

Content Analysis of Argumentation Structures

The Role of Reliability in Argument Mapping

Zur Erlangung des akademischen Grades eines
DOKTORS DER PHILOSOPHIE (Dr. phil.)

von der KIT-Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT) angenommene

DISSERTATION

von

Sebastian Cacean

KIT-Dekan: Prof. Dr. Michael Mäs

1. Gutachter: Prof. Dr. Gregor Betz
2. Gutachter: Prof. Dr. Georg Brun

Tag der mündlichen Prüfung: 15.07.2024



This document is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/deed.en>

INHALTSVERZEICHNIS

1	Zusammenfassung	1
2	Introduction	5
2.1	Argumentation Theory	6
2.2	Political Discourse Analysis	7
2.3	Deliberative Quality Index	9
2.4	Argument Mining	11
2.5	Content analysis of argumentation structures	13
2.5.1	CAAS as Content Analysis	13
2.5.2	The Role of Argumentation Theories	15
2.5.3	The Scope of CAAS	18
2.5.4	CAAS and Normativity	23
2.5.5	CAAS and Reliability	24
2.6	Hermeneutical Underdetermination	25
2.7	Structure of the Thesis	28
2.7.1	Chapter 3: Reliability and Hermeneutical Underdetermination . .	28
2.7.2	Chapter 4: Hermeneutical Underdetermination in Argumentation Analysis	31
2.7.3	Chapter 5: Content Analysis of Argumentation Structures	35
3	Reliability and Hermeneutical Underdetermination	39
3.1	What is Content Analysis?	40
3.1.1	Doing Content Analysis	41
3.1.2	Categorising as Measurement	43
3.2	Reliability and Validity	46
3.2.1	Objectivity as Validity	47
3.2.2	Intersubjectivity as Reliability	48
3.2.3	Reliability-Orientated Content Analysis	54
3.2.4	The Graded Picture of Reliability	57
3.3	Hermeneutical Underdetermination	59
3.3.1	Quantitative vs. Qualitative Content Analysis	59
3.3.2	Reliability and Qualitative Content Analysis	62
3.3.3	Manifest and Latent Content	62
3.3.4	The Challenge of Hermeneutical Underdetermination	67
3.4	Phenomenon Sensitivity	70

3.4.1	Phenomenon Sensitivity and Reliability	71
3.4.2	The Analogy from Error Theory	72
3.4.3	A Recipe for Content Analysis	80
4	Hermeneutical Underdetermination in Argumentation Analysis	87
4.1	Varities of Hermeneutical Underdetermination	89
4.1.1	Node Ambiguity	89
4.1.2	Underdetermination of Granularisation	90
4.1.3	Relation Ambiguity	91
4.2	Varities of Argumentation Analysis	92
4.2.1	Non-Reconstructive Analysis	92
4.2.2	Reconstructive Analysis	93
4.2.3	Applied Formal Logic	96
4.2.4	Informal Logic	98
4.3	Underdetermination in Non-Reconstructive Analysis	100
4.3.1	Node Ambiguity	101
4.3.2	Relation Ambiguity	104
4.3.3	Individuation of Arguments and Reasons	110
4.3.4	Summary	112
4.4	Underdetermination in Reconstructive Analysis	113
4.4.1	Node Ambiguity	117
4.4.2	Individuation of Arguments in Applied Formal Logic	117
4.4.3	Individuation of Arguments in Informal Logic	139
4.4.4	Individuation of Reasons in Informal Logic	146
4.4.5	Relation Ambiguity in Applied Formal Logic	162
4.4.6	Relation Ambiguity in Informal Logic	168
4.4.7	Relation Ambiguity – A Comparative Example	175
4.4.8	Summary	185
4.5	Preliminary Consequences for a Minimal Annotation Scheme of Argumen- tation Structure	188
5	Content Analysis of Argumentation Structures (CAAS)	195
5.1	A Disagreement Measure for CAAS	196
5.1.1	Reliability of Unitising	198
5.1.2	Relational Categories	204
5.1.3	Measuring Disagreement Between Argumentation Structures	208
5.2	Alignment-Based Approaches	211
5.2.1	The Gamma Approach	211
5.2.2	Aligning Argumentative Units	215
5.3	A Probabilistic Conceptualisation of CAAS	221
5.3.1	A Statistical Model for CAAS	224
5.3.2	Significance Testing with CAAS	231
5.3.3	Considerations about Computational Complexity	236
5.4	Summary	237
6	Conclusion	239

A	Appendices	247
A.1	Annotation Guidelines for a Minimal Analysis of Argumentation Structure	247
A.1.1	Visualizing Argumentation Structures with Reason Maps	247
A.1.2	Two Tasks of Analysing Argumentation Structure	250
A.1.3	Used Examples: Germline Editing and Affirmative Action	252
A.1.4	Justificatory Relevance – Some Important Clarifications and Maxims	253
A.1.5	Individuation of Reasons	260
A.1.6	Identifying the Target of Support and Attack Relations	271
A.1.7	Frequently Asked Questions	273
A.2	Mathematical Proofs	279
A.2.1	Pure Significance Testing	279
A.2.2	Constraints on the Parameter Space	280
	Literature	283

1. ZUSAMMENFASSUNG

Die vorliegende methodologische Arbeit entwickelt die Methode *Content Analysis of Argumentation Structure* (CAAS) als reliabilitätstreue und wertneutrale Inhaltsanalyse, die in sozial-empirischen Kontexten zur Analyse der Argumentationsstruktur von Texten eingesetzt werden kann. Das dazu entworfene Kategoriensystem kann direkt angewendet werden oder als Ausgangspunkt für die Entwicklung eines darauf aufbauenden Kategoriensystems dienen.

Zentraler Gegenstand dieser Arbeit ist das Problem hermeneutischer Unterbestimmtheit. Auf der Grundlage theoretischer Überlegungen und verschiedener Beispiele wird gezeigt, dass die Analyse von Argumentationsstrukturen in vielen Fällen zu unterschiedlichen Interpretationen führen kann. Bestehende Interpretationsspielräume können jedoch nicht dadurch verkleinert werden, das Kategoriensystem unter Ausnutzung von Erkenntnissen und Methoden der Argumentationstheorie entsprechend anzupassen. Interpretationsspielräume im Rahmen der Analyse von Argumentationsstrukturen sind in dieser Hinsicht irreduzibel. Damit hängen die Ergebnisse solcher Untersuchungen von subjektiven Interpretationsentscheidungen der Kodierenden ab.

Hermeneutische Unterbestimmtheit ist in inhaltsanalytischen Kontexten problematisch, da sie die Reliabilität der Kodierung gefährden kann. Die Reliabilität der Kodierung kann als Maß darüber verstanden werden, wie stark der die Inhaltsanalytikerin den Ergebnissen der Kodierung vertrauen kann. Reliabilität wird in diesem Zusammenhang als Reproduzierbarkeit (oder intersubjektivität) verstanden und verlangt, dass Ergebnisse der Kodierung nicht in relevanter Hinsicht von den Kodierenden abhängen.

Wenn jedoch die Ergebnisse von Kodierungen von möglicherweise unterschiedlichen Interpretationsentscheidungen der Kodierenden abhängen und daher bei wiederholten Kodierungen desselben Ausgangsmaterials variieren können, ist es nicht ohne weiteres möglich festzustellen, ob beobachtete Unterschiede zwischen Kodierungen unterschiedlicher Texte das Resultat subjektiver Interpretationsunterschiede sind oder ob sie durch Unterschiede in den zu untersuchenden Phänomenen zu erklären sind. Eine Verletzung der Reliabilitätsforderung erschwert es den Forschenden damit, valide Schlüsse aus den generierten Daten für die Beantwortung ihrer Forschungsfragen zu ziehen.

Bisher kennt die Inhaltsanalyse zwei Strategien, um mit hermeneutischer Unterbestimmtheit umzugehen. Die sogenannte *quantitative Inhaltsanalyse* benutzt zufallsbereinigte Reliabilitätsmaße. Diese Maße quantifizieren ob und in welchem Ausmaß Kodierungen bezüglich der Reproduzierbarkeit besser als zufallsgenerierte Kodierungen abschneiden. Wenn die entsprechenden Reliabilitätswerte eines Kodierungsprozesses die empfohlenen

Mindestwerte überschreiten, wird der Datenerstellungsprozess als hinreichend reliabel eingeschätzt. Andernfalls muss das Kategoriensystem so lange verbessert werden, bis hinreichend hohe Reliabilitätswerte erreicht werden. Die Strategie der quantitativen Inhaltsanalyse besteht also darin, den Interpretationsspielraum zu minimieren, indem ein ausreichend präzises Kategoriensystem entwickelt wird.

Diese Strategie führt jedoch nicht immer zum Erfolg. Das Problem besteht darin, dass die jeweilige Forschungsfrage der beschriebenen Strategie bestimmte Randbedingungen auferlegt. Je nach Forschungsfrage ist es unter Umständen nicht möglich, den Interpretationsspielraum hinreichend einzuschränken, ohne die Validität der Inhaltsanalyse zu gefährden. Die sogenannte *qualitative Inhaltsanalyse* wählt einen anderen Weg als die quantitative Inhaltsanalyse. Anstatt Reliabilität zum ausschlaggebenden Kriterium zu erheben, schlägt qualitative Inhaltsanalyse vor, alternative Kriterien für Fälle irreduzibler Interpretationsspielräume anzuwenden. So soll z.B. bezüglich divergierender Interpretationsentscheidungen Transparenz hergestellt werden oder Kodierende sollen zusammenarbeiten, um deliberative Konsensentscheidungen bzgl. der Kategorisierungen zu fällen.

Beide Strategien werden bisher als sich gegenseitig ausschließende und erschöpfende Alternativen verstanden. Entweder sollen Interpretationsspielräume so weit minimiert werden, bis das Kriterium der Reliabilität erfüllt ist oder die Forschenden müssen auf Reliabilität verzichten und alternative Kriterien heranziehen, um die Güte einer Inhaltsanalyse zu qualifizieren. Bislang gibt es in der Inhaltsanalyse also keinen Vorschlag, wie irreduzible Interpretationsspielräume mit der Erfüllung des Reliabilitätskriteriums in Einklang zu bringen sind.

Die vorliegende Arbeit zeigt, dass es eine bisher nicht beachtete Alternative gibt, die es ermöglicht, Argumentationsstrukturen inhaltsanalytisch auszuwerten, ohne dabei auf Reliabilität zu verzichten, selbst wenn die Analyse von Argumentationsstrukturen hermeneutisch unterbestimmt ist.

Die Vorschläge in dieser Arbeit machen sich dabei eine Analogie zu Messungen in den Naturwissenschaften zu Nutze. Während die Ergebnisse von Messungen im Gegensatz zu inhaltsanalytischen Untersuchungen nicht aufgrund von subjektiven Interpretationsentscheidungen auseinanderklaffen können, sind Messungen in den Naturwissenschaften ebenfalls häufig nicht reproduzierbar. Diese Variabilität wird häufig über das Auftreten zufälliger Messabweichungen erklärt und entsprechend im Rahmen eines probabilistischen Ansatzes mathematisch präzisiert. Ähnlich wie bei dem Problem der hermeneutischen Unterbestimmtheit stellt sich im Rahmen der Messtheorie die Frage, unter welchen Bedingungen Unterschiede in Messergebnissen Rückschlüsse zu Unterschieden in den beobachteten Phänomenen zulassen.

Die probabilistische Präzisierung ermöglicht unter Rückgriff auf statistische Methoden die Formulierung präziser Antworten auf diese Frage. Signifikanztests dienen in der vorliegenden Arbeit als Beispiel einer konkreten statistischen Methode. Diese Signifikanztests ermöglichen es den Forschenden zu entscheiden, ob Unterschiede in den Messergebnissen allein durch zufällige Fehler erklärt werden können oder ob sie auf einen Unterschied in den Phänomenen hindeuten.

Obwohl es Unterschiede zwischen Messungen in den Naturwissenschaften und Textko-

dierungen im Rahmen von Inhaltsanalysen gibt, ist es möglich, den Kodierungsprozess probabilistisch aufzufassen. Diese Konzeptionalisierung läuft auf einen probabilistischen Reliabilitätsbegriff hinaus, der im Gegensatz zu zufallsbereinigten Reliabilitätsmaßen in Kontexten irreduzibler Interpretationsspielräume sinnvoll verwendet werden kann. Eine Ausarbeitung dieses Vorschlags erfordert eine probabilistische Formulierung des Annotationsprozesses, die Spezifikation eines statistischen Modells und die Ausarbeitung geeigneter statistischer Methoden.

Dazu wird in der vorliegenden Arbeit der Kodierungsprozess als Zufallsprozess verstanden, insofern Kodierende zufällig aus der Grundgesamtheit aller geschulten Kodierenden gezogen werden. Der Raum aller Interpretationen, verstanden als die (hypothetische) Menge der Kodierungen aller geschulten Kodierenden, wird auf diese Weise durch eine Wahrscheinlichkeitsfunktion ergänzt, die jeder Kodierung einen Wahrscheinlichkeitswert zuweist. Das heißt, der Kodierungsprozess wird durch eine Wahrscheinlichkeitsfunktion über den Ergebnisraum aller Kodierungen modelliert. Die Arbeit zeigt, wie der Wahrscheinlichkeitsraum und die entsprechende Verteilung auf Grundlage einer Stichprobe von Kodierungen geschätzt werden kann. Ferner wird erläutert, wie Signifikanztests eingesetzt werden können, um statistische Rückschlüsse von beobachteten Unterschieden zwischen Kodierungen auf Unterschiede in den Phänomenen zuzulassen.

Insgesamt wird damit gezeigt, dass ein irreduzibler Interpretationsspielraum mit der Erfüllung des Reliabilitätskriteriums vereinbar ist. Die vorgeschlagenen Methoden und das entwickelte Kategoriensystem können in verschiedenen Forschungskontexten zur Analyse von Argumentationsstrukturen eingesetzt werden.

2. INTRODUCTION

Argumentation is ubiquitous. As individuals, we are asked to justify our decisions, preferences and political views; as collaborative groups, we have to decide on a common course of action—often by discussing what is right and just. While decision making is framed by power relations, skewed by epistemic inequalities and thereby seldomly guided by rational reason alone, people are accountable for their actions. So even if decisions are based on intuitions rather than reasons, people are called on to produce post-hoc justifications for their actions (Haidt 2012).

This practice of producing reasons is prevalent in small groups—for instance, among friends, family members and co-workers—and in larger collectives. In democratic societies, political decision making is intricately linked to public argumentation, which is “argumentation that is about a public issue [...] and that is typically expressed by people in their public capacity, e.g. as citizens or politicians [...], in an open forum [...], while addressing a larger audience whose members are people in their public capacity [...]” (Zenker et al. 2020, 3). Public argumentation plays a central role in both the informal and formal public sphere: Policy options are discussed and decided upon in parliamentary debate; the media reproduces these debates and shapes them by providing a forum for public argumentation; citizens are informed by the media and participate by them in public argumentation.

Due to this omnipresence of argumentation and its relevance for our private and public affairs, it is hardly surprising that researchers from diverse academic disciplines, such as political theory, linguistics, discourse analysis, computer science, law, communication science and rhetorics have a keen interest in understanding and analysing argumentation. Admittedly, people as arguers always do that without relying on scientific expertise. The mere participation in argumentation demands a proper understanding of others’ arguments. Nevertheless, in contrast to our ordinary-language practice of exchanging reasons, a scientific perspective on argumentation is usually linked to a specific research question, demands adherence to certain scientific norms and is based on scientific methods.

This methodological work introduces such a set of scientific methods. It is methodological because I will discuss and propose such methods without applying them to answer a research question. I will draw on existing approaches, particularly argumentation theory and content analysis, and mesh them into a methodology, which I will call *content analysis of argumentation structures* (in short, CAAS).

The methodological innovations of CAAS and the challenges it is tailored to meet are best explained by comparing CAAS to existing approaches. In the following, I will

present four paradigms of analysing argumentation in a nutshell—namely, *argumentation theory*, *political discourse analysis*, the *deliberative quality index* and *argument mining*. In particular, I will identify their restrictions and describe the main ideas of overcoming them.

2.1 ARGUMENTATION THEORY

The term *argumentation theory* is, in some sense, a misnomer: It does not refer to one particular theory but is instead an umbrella term for various approaches and theories to study argumentation, which can differ in their methods and may even be incompatible. It includes, among others, Pragma-Dialectics, Informal Logic, Applied Formal Logic, Rhetorics, Formal Dialectic and the Toulmin Model—some of which will be discussed in more detail in Chapter 4. In its broadest sense, argumentation theory “covers the study of argumentation in all its manifestations and varieties, irrespective of the intellectual backgrounds of the theorists, their primary research interests, and their angles of approach” (Eemeren et al. 2014, 7). What argumentation theories share is their general objective, which is “a practical one: to provide adequate instruments for analysing, evaluating, and producing argumentative discourse” (Eemeren et al. 2014, 12).

What are these instruments? They usually comprise three types:¹ First, explications of crucial concepts, such as *argumentation*, *reason*, *argument* and *refutation* clarify these notions, which are, otherwise, too vague and too ambiguous for a rigorous study of argumentation. Second, argumentation and its constituent components must be identified in oral or written texts. Since natural language is often messy, ambiguous and context-dependent, there are no mechanical ways to do so. Rather, an argumentation theory has to provide heuristics, methods, examples and best practices that help to distinguish argumentative components from other elements in our natural language. Often, argumentation theories adopt a reconstructive approach: Instead of merely identifying arguments and their components in natural language texts, the detected text segments are converted into a streamlined argument form. The reconstruction usually involves a reformulation of identified text segments and often the addition of implicit premises and implicit conclusions. Finally, the study of argumentation contains normative dimensions: Often, argumentation should be evaluated to identify fallacious reasoning and improve argumentation. To that end, an argumentation theory has to introduce normative standards for evaluating arguments and methods to apply them.

One obvious aim of argumentation theory is to increase the quality of argumentation. For instance, the tools provided by Informal Logic and Critical Thinking are used in educational contexts to increase students’ argumentation skills. It is similarly possible to aid argumentation in more specific domains. The argumentative turn in policy analysis, as described by Hansson and Hirsch Hadorn (2016b), aspires to provide prospective support for political decision making. According to Hansson and Hirsch Hadorn (2016a), the tools of traditional decision theory, such as cost-benefit analysis or risk analysis, are not applicable in situations under great uncertainty. In these contexts, political decision makers may not be aware of all relevant decision options, may lack knowledge about consequences and their probabilities and may face normative uncertainties. Conventional approaches to

¹This is a very simplified model. See Eemeren and Grootendorst (2004), who introduce a more refined model and five components of argumentation theory.

decision making are ill-equipped to deal with these situations. However, these conditions are part of the politician's daily life and not only an exception.

By relying on insights from philosophy and argumentation theory, the argumentative approach to policy analysis provides decision support in two ways: First, it clarifies and explicates the conditions under which conventional decisions theory is not applicable (see Hansson and Hirsch Hadorn 2016a, 23–29). Second, it provides the decision maker with tools that complement and replace conventional approaches to decision making. Instead of applying one specific framework of decision making, such as cost-benefit analysis, policy makers should scrutinise arguments that are put forward or raised as objections against policy options—independent of the decision-theoretic framework on which these arguments are based. To that end, Betz and Brun (2016) and Betz (2016) introduce methods from argumentation theory to analyse practical argumentation. They describe argument schemes of practical reasoning under great uncertainty, which can be used as templates to formulate valid arguments for or against certain policy options. Additionally, Betz and Brun (2016) introduce the method of argument mapping as a tool to understand and evaluate the complex interplay between different arguments and objections. Finally, Hansson (2016) discusses several typical fallacies people are prone to when faced with uncertainties and suggests ways to improve decision making under uncertainty. The case studies in the last Chapter of Hansson and Hirsch Hadorn (2016b) illustrate the relevance of these tools for political decision making and demonstrate that insights from argumentation theory can be fruitfully applied in very specific argumentative contexts.

2.2 POLITICAL DISCOURSE ANALYSIS

I. Fairclough and Fairclough (2012) proposed a form of political discourse analysis (in short, *PDA_{F&F}*) that is of particular interest for understanding CAAS since they explicitly endorse methods that are based on argumentation theory to analyse and evaluate political discourse.

Discourse analysis is a branch of linguistics that studies the broader aspects of language and its use in social contexts.² There are numerous definitions of the target concept *discourse*, which all fall into three categories: “(1) anything beyond the sentence, (2) language use, and (3) a broader range of social practice that includes non-linguistic and non-specific instances of language” (Tannen, Hamilton, and Schiffrin 2015, 1).³

Discourse analysis is based on two fundamental assumptions: First, interpreting what someone says or writes is context-dependent (Paltridge 2012, 2). The same utterance can be understood differently in different situations. Accordingly, discourse analysts include their knowledge of all relevant extra-linguistic factors in their study of language use. Second, discourse analysis draws on Austin's (1962) seminal insights that language use is not only a transmission of meaning. Rather, we can perform actions by uttering words

²The term *discourse analysis* does not represent a concise and unique set of methods (Titscher et al. 2000, 144). Rather, different authors present partially diverging accounts of discourse analysis. Here, I try to describe general features that appear to be common to all of them. The same holds for critical discourse analysis.

³For a survey of different definitions, see Jaworski and Coupland (2006).

to reach specific goals (Jaworski and Coupland 2006, 13). Pronouncing two people as husband and wife is an often-mentioned paradigmatic example in this connection.

In discourse analysis, this performative character of language is taken as a motivation for a social-constructivist underpinning of discourse. On this view, a mutually constitutive relationship exists between language use and social reality: “We shape, produce, and reproduce the world through language in use. In turn the world we shape and help to create works in certain ways to shape us as humans” (Gee and Handford 2012, 5). This constructivist perspective on discourse has far-reaching methodological consequences for discourse analysis: “Where other qualitative methodologies work to understand or interpret social reality as it exists, discourse analysis tries to uncover the way that reality is produced” (Hardy, Harley, and Phillips 2004, 19). Accordingly, discourse analysis “examines how the use of language is influenced by relationships between participants as well as the effects the use of language has upon social identities and relations” (Paltridge 2012, 2).

Critical discourse analysis is a young branch of discourse analysis based on critical linguistics ideas (Titscher et al. 2000, 144) and critical social analysis (N. Fairclough 2012). Its focus is on the interdependence of language, power relations and ideology within and between different societal groups. Critical discourse analysts maintain that inequalities and social wrongs are reproduced, caused and legitimised by language. It is not confined to describing and explaining the relationship between discourse and social reality but “also evaluates them, assesses the extent to which they match up to various values, which are taken (more or less contentiously) to be fundamental for just or decent societies” (N. Fairclough 2012, 9). This evaluative perspective is normative since it is grounded in views of human well-being and morality. In addition to this normative critique, critical discourse analysis seeks to provide explanatory critique by analysing the structures that cause social injustices. The resulting explanations are then used to formulate suggestions for changing unjust social realities for the better. Critical discourse analysis is, therefore, “politically involved research with an emancipatory requirement: it seeks to have an effect on social practice and social relationships” (Titscher et al. 2000, 147).

Political discourse analysis is—as the name suggests—about political discourse, which can be “defined as talk and text produced in regard to concrete political issues (language in politics) or through the actual language use of institutional political actors, even in discussions of nonpolitical issues (language of politicians)” (Kampf 2015, 3). In other words, political discourse encompasses public argumentation as defined by Zenker et al. (2020) (see above) and includes, for instance, parliamentary debate, media interviews about political issues, political talk shows and political speeches.⁴ Political discourse analysis is a variant of critical discourse analysis and “deals especially with the reproduction of political power, power abuse or domination through political discourse, including the various forms of resistance or counter-power against such forms of discursive dominance” (Dijk 1997, 1).

While the central role of argumentation in political discourse is acknowledged by most political discourse analysts (Dijk 1997, 29–30), I. Fairclough and Fairclough (2012) are the first authors who integrate “critical discourse-analytical concepts with the analytical framework of argumentation theory, on the basis of viewing political discourse as primarily

⁴For a more thorough discussion of defining *political discourse*, see Dijk (1997).

argumentative discourse” (I. Fairclough and Fairclough 2012, 17). They seek to “increase the capacity of CDA to pursue its aim of extending critique to discourse” (I. Fairclough and Fairclough 2012, 78) by analysing and evaluating argumentation.

According to I. Fairclough and Fairclough (2012), political discourse is primarily argumentative and involves, first and foremost, practical argumentation—that is, arguments in favour of or against suggestions to act in certain ways. Other non-argumentative pre-genres of discourse, such as narration, description and explanation, are viewed as subordinated to argumentation. These other pre-genres are embedded within political argumentation to the extent that they can be interpreted as premises in practical arguments.

The analytical framework endorsed by I. Fairclough and Fairclough (2012) is grounded in Informal Logic and Pragma-Dialectics and labours under the assumption that political arguments are best understood as means-ends reasoning. The critical evaluation of public discourse will then proceed along the following lines: After identifying pro and con arguments in public discourse, they are reconstructed by using one of two argument schemes of practical reasoning, which are based on Walton’s (2006, 2007) argument schemes for practical reasoning. Transforming arguments into an explicit premise-conclusion structure often involves identifying implicit premises, thereby making them accessible to a critical evaluation. The reconstructed argumentation can then be evaluated by answering critical questions, which question the acceptability of premises or the conclusion or the inferential link between the premises and the conclusion (I. Fairclough and Fairclough 2012, 62–67).

2.3 DELIBERATIVE QUALITY INDEX

Deliberative democracy is a form of democracy that puts deliberation at the center stage of analysing political will formation. While deliberative theory lacks a shared conceptualisation of *deliberation* (Carpini, Cook, and Jacobs 2004; Chambers 2003), many clarifications involve the idea that deliberation includes direct and mediated forms of “debate and discussion aimed at producing reasonable, well-informed opinions in which participants are willing to revise preferences in light of discussion, new information, and claims made by fellow participants” (Chambers 2003, 309). According to this understanding, collective decision making is not viewed as voting-centric but talk-centric. Instead of understanding political will formation as a mere preference aggregation by voting, deliberative theory focuses on deliberation to form preferences prior to voting.

The general idea is that deliberation is good for democracy if it lives up to certain normative standards. Deliberative theories clarify this basic thought by formulating normative criteria and elucidating in what way deliberation is desirable and what goals are being pursued by it. For instance, deliberation should be rational in the sense that outcomes of the collective decision-making process are more influenced by the quality of arguments and not by power relations, psychological biases or personal interests.

However, whether deliberation measures up to these aspirations is an empirical question. Accordingly, a deliberative theory must be sufficiently precise to allow empirical scrutiny. In other words, it must be possible to measure the quality of deliberation to test the various hypotheses deliberative theorists devise to describe the role of deliberation in democratic societies.

The deliberative quality index (Steenbergen et al. 2003; Steiner et al. 2004), in short DQI, is one such “measurement instrument that could help bridge the gap between political theory and empirical research” (Steenbergen et al. 2003, 43).⁵ It can be used to measure the quality of political deliberation either as a dependent or as an independent variable (Steiner et al. 2004, 5). The first case comprises research questions that examine the circumstances under which the quality of deliberation suffers or improves; the second case addresses the influence of different quality levels of deliberation on policy outcomes or other quantities of interest.

Habermas’ deliberative model is used as the underlying normative model to conceptualise the DQI’s notion of deliberative quality. It comprises six primary normative constraints that formulate the regulative ideal that real deliberation should strive to mimic (Steenbergen et al. 2003, 25–26). First, everyone affected by a decision should be able to participate in discussing the decision without being pressured or coerced. Second, the deliberation should be rational and, in particular, based on rational argumentation. Interlocutors should share relevant information and exchange their reasons inasmuch as they are justificatory relevant. The third component puts an additional constraint on reasons: Not any justificatory relevant reason will increase the deliberative quality. Rather, reasons should allude to the common good by, for instance, considering the well-being of others or the community at large. Fourth, participants should treat one another with respect. Fifth, even if consensus is not attainable, “participants in a discourse should at least attempt to reach mutually acceptable compromise solutions” (Steenbergen et al. 2003, 26). Finally, participants are required to be authentic. They are not allowed to obscure their intentions and preferences but have to express them sincerely and faithfully.

The basic idea of devising an aggregated quality index out of these normative components is this: If possible, each component is operationalised as a category or a set of sub-categories with measurable values.⁶ The different category values for each category are mapped to codes, which are numerical values that measure to which extent the discourse unit approaches the underlying normative ideal. Measurable means that a human annotator can apply the category system to individual discourse elements that constitute the discourse. In other words, the annotator must be able to categorise each discourse element according to all defined categories. In the context of deliberation, a discourse element is a speech act that expresses a demand—“that is, a proposal on what decision should or should not be made” (Steenbergen et al. 2003, 27)—and speech acts that are connected to such a demand by formulating reasons in favour or against it.

Here, we are only interested in those categories that are connected to argumentation.⁷ These include three categories: level of justification, content of justification and respect for counterarguments.

⁵For an overview of other approaches to measure deliberation, see (Neblo 2011, 545–52).

⁶The DQI operationalises all but one of the six key components. The requirement of authenticity would demand to consider the intentions of the speaker, which are not directly observable. Accordingly, authenticity is not included in the DQI since “the speculative nature of such a judgment [about the speaker’s intentions] is bound to introduce large amounts of (possibly systematic) measurement error” (Steenbergen et al. 2003, 26).

⁷Accordingly, we can refrain from understanding how the different category values can be aggregated to one number that represents the overall deliberative quality of the discourse element. For details, see Steenbergen et al. (2003, 39–41).

The *level of justification* explicates the second normative component, which requires deliberation to contain rational argumentation. It represents an aspect of the epistemic quality of deliberation and gauges whether “a speech gives complete justifications for demands” (Steenbergen et al. 2003, 28). After identifying a discourse unit—a speech act formulating a demand—the annotator has to decide whether the interlocutor offered reasons for the demand and then judge the completeness of the justifications. A justification is considered complete if there is a “linkage between conclusion [the demand] and reason” (Steiner et al. 2004, 171). Completeness is not supposed to qualify the quality of arguments in terms of the plausibility or acceptability of the given reasons but whether there is a justificatory relation between the reasons and the demands.

There are four different levels of justification (described in ascending order) (Steenbergen et al. 2003, 28). The first one (*no justification*, code 0) is used if the proponent of a demand offers no justification at all. The second category value (*inferior justification*, code 1) is to be assigned to a discourse unit if an incomplete justification is provided—that is, reasons that either lack a linkage to the demand or are mere illustrations. If the linkage is complete, the level of justification represents a *qualified justification* (code 1). Finally, if the interlocutor offered at least two complete justifications for their demand, the justification is considered *sophisticated* (code 2).

The category *content of justification* is an operationalisation of the third normative component, which requires that reasons should relate to the common good instead of being based on group interests alone (Steenbergen et al. 2003, 28–29). While the level of justification concerns the formal structure of arguments, this category relates to the content of arguments. It comprises four category values. The category value *explicit statement concerning group interests* corresponds to the lowest code (0) and is assigned if “one or more groups or constituencies are mentioned in a speech” (Steenbergen et al. 2003). If the reasons do neither allude to constituency/group interests nor to the common good, the category value *neutral* (code 1) is to be used. If there is an explicit reference to the common good, the codes 2a or 2b can be assigned. The former is used if the common good is conceptualised in utilitarian terms, and the latter if it is based on Rawls’ difference principle.

The normative component of respect is conceptualised via three sub-categories (Steenbergen et al. 2003, 29). One of them, the sub-category *respect towards counterarguments*, operationalises how the speaker relates to mentioned or anticipated counterarguments. It comprises four category values. If the speaker ignores counterarguments, the lowest code (0) is assigned; if they acknowledge counterarguments and degrade them by using negative statements, the second lowest code (1) is assigned. If, on the other hand, such an acknowledgement is neither accompanied by a negative nor a positive statement the annotator should assign the code 2. Finally, if counterarguments are positively valued, the corresponding discourse unit receives the highest code (3).

2.4 ARGUMENT MINING

Argument mining, sometimes also referred to as argumentation mining, uses sophisticated algorithms to automatically identify and extract arguments, reasons, objections, and other argumentative components along with their relationships to each other as they are presented

in natural language (Lawrence and Reed 2019). The resulting structured data can then be analysed using computational models of argumentation (Lippi and Torroni 2016).

Once it matches the quality of human annotation, argument mining could enable the extraction of argumentation structure from texts on a massive scale (Lippi and Torroni 2016; Lawrence et al. 2014; Stede and Schneider 2018). It has, therefore, the potential to solve the scalability problem of manual analysis, which is limited by the capabilities of human annotators. The problem is that the extraction of argumentation is a time-consuming process that requires intensive training. Accordingly, human annotation of argumentation structure does not scale well unless one is willing to invest tremendous resources into the effort.

Argument mining is in its usability and potential scope of application as generic as argumentation theory. In contrast to DQI, which has a more specific goal, namely determining the deliberative quality in debates, argument mining can be used for different purposes. Argument mining could, for instance, support strategic choices in policy or business contexts by feeding decision and reasoning models with its results (Lippi and Torroni 2016). Another use case is the aggregation and visualisation of complex debates. Argument mining would extract all relevant arguments of a large debate, which might be scattered over heterogeneous sources such as blog posts, articles, books and discussion forums. Based on these results, users could then receive an aggregated overview of the debate (Habernal 2014). In educational contexts, argument mining could help students improve their writing skills by giving them feedback on their argumentation (Stede and Schneider 2018). The same systems could support teachers in scoring essays by pointing to argumentative weaknesses.⁸

Human annotation of argumentation structure is central in furthering automatic argument extraction. On the conceptual level, “manual analysis can offer unique insight into how this task can be automated” (Lawrence and Reed 2019, 10); on the practical level, human annotations are used as training data for the algorithms and as gold standards to evaluate the performance of an automatic extraction (Lippi and Torroni 2016).

The role of human annotation in argument mining is of particular interest for this work since the argument mining community is concerned with questions that are similarly relevant for CAAS: Which argument models are helpful in extracting argumentation structure from natural-language text? What kind of training enables humans to perform a manual extraction? Are there relevant differences between different domains? Can annotation efforts be reproduced? The argument-mining community approaches these questions with conceptual clarity and empirical means. The empirical results provide invaluable insights into human annotation performance.

The mutual interest in these questions and the fact that human annotation in argument mining is, from a methodological perspective, the closest sibling to CAAS (see below) make argument mining a vital foil for conceptualising CAAS.

⁸For other potential use cases, see Stede and Schneider (2018), Section 10.3.

2.5 CONTENT ANALYSIS OF ARGUMENTATION STRUCTURES

We can now use the described approaches to sketch the basic ideas and distinguishing features of CAAS. In short, CAAS is a reliability-orientated and value-neutral content analysis of argumentation structures that is based on argumentation theory. Let me unwrap each of the defining properties in a nutshell.

2.5.1 CAAS AS CONTENT ANALYSIS

First, CAAS is conceptualised as a subdiscipline of content analysis, which is a set of research methods to analyse texts—or meaningful matter in general—to answer specific research questions.

From a practical viewpoint, the main distinguishing feature of content analysis as compared to other social research methods is its use of fixated category systems, or annotation schemes, to categorise text. Ideally, the different category values of each category are precisely defined in a coding manual that is used to instruct annotators. The actual categorisation is performed as a two-step process. Instead of categorising the target material as a whole, the text is divided into different text segments and only a subset of these segments is supposed to be categorised. Relevance criteria, which must be defined during the design phase of the research effort, determine which text segments are supposed to be categorised—so-called coding units. The idea is to assign each coding unit one category value for each category. In other words, applying the category system can be interpreted as an annotation of the text. The generated data will then be used to draw inferences to answer a research question.

The DQI approach is clearly a form of content analysis. The paradigmatic target object is parliamentary debate. Coding units are defined as speech acts that formulate a demand (Steenbergen et al. 2003, 27). Annotators are then supposed to classify these coding units according to the described category system, which operationalises dimensions to assess the quality of deliberation (Steenbergen et al. 2003, 27–30).

Similarly, most of the manual annotation studies performed in the context of argument mining follow the content-analysis paradigm. Coding units correspond to text segments that express or formulate argumentative components and will be called argumentative units.⁹ The used argumentation model is translated into an annotation scheme, which is applied to categorise argumentative units according to their justificatory role (e.g., premises or claims) and to categorise justificatory relations between argumentative units (e.g., that

⁹Note that I distinguish between argumentative components and argumentative units. The latter refers to text segments annotators identify as argumentatively relevant—in other words, their identified coding units. The term *argumentative component*, on the other hand, refers to the argumentative phenomenon that is formulated, expressed or referred to, such as a reason or an objection. They differ in their identity criteria. Different argumentative units of different annotators might represent the same argumentative component. For instance, if one annotator decides to annotate a text segment as an argumentative unit together with the linguistic cue, but another annotator omits the linguistic cue, they identified two different text segments as an argumentative unit since the text segments will have different start points or different end points. However, both units represent the same argumentative phenomenon. In the literature, argumentative units are also called *argumentative discourse units* (Peldszus and Stede 2013; Stede and Schneider 2018), *argument components* (C. Stab and Gurevych 2014) or *argument units* (Eckle-Kohler, Kluge, and Gurevych 2015).

one unit is presented as an objection against another unit).¹⁰ A coding manual, which includes the category definitions, clarifications and further instructions, is used to train annotators and is often published for transparency and repeatability.

The version of political discourse analysis proposed by I. Fairclough and Fairclough (2012) does not fit into the picture of content analysis. Content analysis is characterised by employing criteria that define coding units and by applying an annotation scheme with fixed definitions of category values.

In some sense, $PDA_{F\&F}$ uses criteria that determine what text segments are relevant for “categorisation” (even though discourse analysts do not refer to them as *coding units*): Text segments are relevant if they explicitly express arguments or if they can be reinterpreted as premises of arguments (I. Fairclough and Fairclough 2012, 13).

However, in contrast to the DQI approach and the human annotation efforts of argument mining, $PDA_{F\&F}$ does not employ a fixated category system. The discourse analyst is not bound to category values that are defined prior to their analysis. Instead, discourse analysts can (more or less) decide how to analyse their “coding units” from case to case.

Let me illustrate this point by comparing the DQI and $PDA_{F\&F}$ in more detail. Both approaches aim to evaluate argumentation. The DQI uses three categories to evaluate the epistemic qualities of argumentation (see above). One of them, the level of justification, is used to evaluate the inferential link between premises and conclusion (see above). $PDA_{F\&F}$ demands to reconstruct arguments in their premise-conclusion structure by employing argument schemata. The argument is then to be evaluated by assessing how it performs with respect to critical questions. Each argument schema is equipped with a set of critical questions. Some will target the premises or the conclusion by questioning their plausibility or acceptability, and some will target the inferential link between the premises and the conclusion (I. Fairclough and Fairclough 2012, 62–66). Answering the latter will, therefore, play a similar role in evaluating the argument as determining the level of justification in the DQI approach.

However, in contrast to the introduced category values of the level of justification, there are no fixated category values in $PDA_{F\&F}$. The discourse analyst would have to fix the different possibilities of how an argument can perform with respect to critical questions, which they does not. Instead, the discourse analyst uses critical questions as an inspirational tool that guides their evaluation. They neither have to use every critical question in their assessment nor are they confined to using critical questions. Additionally, the specific way to answer these questions is not fixed prior to the analysis. In this way, the individual discourse analyst is much more unrestricted in their analysis. Accordingly, the results will differ immensely: While the content analyst will end up with a set of assigned labels, the discourse analyst will formulate their results in a qualitative manner.¹¹

Content analysis, as hitherto described, is a very suitable methodological background for the analysis of argumentation structure. The different argumentative components are usually formulated or expressed by specific text segments within a larger speech act or

¹⁰See Stede and Schneider (2018), Chapter 4 for an overview.

¹¹This is not to say that the content analysts will stop there. Instead, the generated data will be used to draw further inferences, which often includes applying quantitative methods.

text containing text segments with other functions. In other words, the identification of argumentative components demands to decide whether text segments have a justificatory function or some other function. Consequently, text segments are coding units to the extent that they express argumentative components. Monadic category values will correspond to different types of argumentative components, and relational category values can be used to represent the different justificatory relations between these argumentative components.

2.5.2 THE ROLE OF ARGUMENTATION THEORIES

CAAS is grounded in insights from argumentation theory but is also agnostic towards the specific details of different argumentation theories. Instead, I try to ground CAAS on a common core of different argumentation models. Let me motivate this commitment by describing the extent to which the other accounts are based on argumentation theories.

The DQI approach includes categories that presuppose a basic analysis of argumentation structure. The three categories *respect for counterarguments*, *level* and *content of justification* demand to identify argumentative units and their justificatory relations. However, the DQI is not based on argumentation theory. To my mind, this neglect is unfortunate since argumentation theories offer ample help to facilitate the analysis of argumentation structure. In particular, they can guide the formulation of coding instructions since they come with various key concepts and their clarifications.

I do not suggest that content analysts should reuse these concepts without any adjustments. However, the formulation of coding instructions should be informed by insights from argumentation theory since such a grounding can increase the clarity of coding instructions, which will benefit the reliability of the coding process.

The DQI category *level of justification* can be used to illustrate this point. The DQI is in line with the basic structural understanding of arguments. Arguments have to be analysed with respect to three constituent components: A part that is supposed to be justified (the conclusion, or *demand* in the DQI terminology), elements that are used as justification (the premises, or reasons) and the justificatory link or connection between premises and conclusion. The level of justification is used to judge whether speakers provide reasons for their claims and to evaluate the quality of the justificatory link between premises and conclusion. Let us concentrate on its category values *inferior* and *qualified justification*.

A justification is qualified if, first, the speaker provides a reason for their conclusion and, second, the linkage between reason and conclusion is complete. In contrast, a justification is only inferior if this linkage is incomplete—either because “no linkage is made” or if the “conclusion is merely supported with illustrations” (Steenbergen et al. 2003, 28).

The key notion of completeness is to be understood as a binary attribute and doesn’t come in grades: Either the connection between reasons and conclusion is complete, or it isn’t. Additionally, the notion of completeness is used exclusively to evaluate the inferential link between premises and conclusion in the following ways: First, whether an inference is complete does not depend on what the speaker *thinks about* the quality of the inferential link. Both category values apply to cases where speakers provide reasons or justifications. Naturally, a speaker will usually maintain that the linkage between their reasons and what they try to justify is complete. Consequently, categorising what the author thinks

about the inferential link would be trivial. Instead, it is about the inferential link per se. Second, completeness does not depend on the epistemic qualities of the premises. In other words, the inferential connection should be assessed without considering the plausibility or acceptability of the premises.¹²

The problem is that the authors do not offer a sufficient characterisation of their completeness notion. They provide illustrating examples which lead, however, to further questions. According to Steenbergen et al. (2003), the following example represents a complete justification

Does the hon. Lady [Caroline Spelman, Conservatives, Meriden] agree that there is a further point on the separate taxation of men and women? Women who are abused in the household sometimes find it difficult to get away from the home. Separate taxation helps women to have the courage to move out on an abusive household (Steenbergen et al. 2003, 32).

and the subsequent one an incomplete justification:

Does my hon. Friend [Eleanor Laing, Conservatives, Epping Forest] agree that, if the rumours are true that people will not need receipts to claim the child care allowance, they could indeed spend the money on washing machines (Steenbergen et al. 2003, 32)?

In the first example, the speaker argues in favour of separate taxation for men and women (the demand) by saying that such a separation could help women to extract themselves from abusive households (the justification); in the second example, the speaker argues that parents should only receive childcare allowance if they are required to file in receipts as to how they spend that money (the demand). Otherwise, they might spend it on things that would not profit the child (the justification).

In both cases, the speaker provides something as justification. But what distinguishes the inferential link in the first example from the one in the second example? The justification in the second example is supposedly incomplete because the expressed reason “is not backed up by an argument or evidence” (Steenbergen et al. 2003, 32). This explication is, however, unsatisfactory. First, it seems to suggest that the quality of the inferential link depends on the existence of second-order reasons. It is not the inferential link in isolation that matters but whether the speaker provides further justifications for their reasons. This interpretation is at least at odds with explications of inferential adequacy that argumentation theorists provide. Second, there seems to be no difference in terms of second-order reasons between both examples. Similarly to the second example, the first does not include any evidence or other back-up for the provided justification.

Perhaps the authors mean that such a back-up is only needed under certain conditions. In the second example, it might seem implausible, perhaps even preposterous, that parents act contrary to their parental nature. Accordingly, a complete justification might only

¹²The authors write that they do not want to “judge how good an argument for a demand is or whether [they] agree with it” (Steiner et al. 2004, 171). One could object that evaluating the inferential link is already part of an overall assessment of an argument’s quality. While this is the usual view in argumentation theory, the quoted statement should not be interpreted as revealing an inconsistency. Instead, they mean that they refrain from evaluating the premises’ truth, plausibility or acceptability.

need further back-up if the first-order reasons are not acceptable or implausible on their own. One problem is that the authors do not provide any such explication. Additionally, this interpretation presupposes evaluating the epistemic qualities of premises and reasons, which the authors want to avoid.¹³

These considerations illustrate that the discussed category values lack, to some extent, clarity. To my mind, the clarity of such categories can be increased by considering insights from argumentation theory—a claim I cannot substantiate here.¹⁴

In contrast to the DQI, $PDA_{F\&F}$ and argument mining are based on argumentation theories. I. Fairclough and Fairclough (2012) follow a reconstructive approach. They adapt Walton's argumentation schemes, use them to reconstruct natural language arguments in their premise-conclusion form and evaluate arguments as argumentation theory suggests.

The annotation studies in the context of argument mining are based on different argumentation models.¹⁵ For instance, Habernal and Gurevych (2016) use a generic argument scheme based on Toulmin's (1958) argument model to annotate argumentation structure in user-generated Web discourse. Peldszus and Stede (2015) ground their argumentation model on Freeman (1991, 2011) to analyse argumentative microtexts. And Visser et al. (2021) annotate argument schemes in transcripts of televised US presidential election debates in 2016 by using Walton's argument schemes and compare the results with an alternative annotation based on Wagemans' Periodic Table of Arguments.

This heterogeneity is, however, also problematic to some extent. The use of a specific argumentation model is often motivated by the specifics of the target object. This observation might suggest that the choice of the argumentation model depends on the domain at hand.¹⁶ But then, any argument mining technique based on data produced in a specific domain with a specific argumentation model is limited to that domain (Lawrence and Reed 2019, 42). As a consequence, we would have different argument mining approaches with their own training data, specialised to specific domains, without being able to adapt them to new domains easily.

CAAS is similarly based on insights from argumentation theory. In Chapter 4, I will discuss different argumentation theoretic accounts and consider their various implications for analysing argumentation structure. While CAAS is not conceptualised as helpful for the automatic extraction of argumentation structure, it would be similarly problematic to ground CAAS on the specifics of one particular argumentation theory. Instead, I will ground CAAS on a shared core of different argumentation theories compatible with most of the more specific accounts. This approach has at least three advantages. On a methodological level, CAAS can be supplemented with other elements from more specific theories to account for domain-specific and other contextual needs. On a more practical level, the results of a CAAS study will always be partially comparable with other studies based on more specific argumentation theories. Finally, CAAS will not be confined to a specific

¹³Cf. Footnote 12.

¹⁴In discussing the DQI, Friberg-Fernros and Schaffer (2017) suggest evaluating the inferential link by using the notion of deductive validity, a concept which stems from argumentation theory and which can be considered an explication of what *complete justification* means.

¹⁵For an overview, see Habernal (2014)

¹⁶See, e.g., Habernal, Eckle-Kohler, and Gurevych (2014).

argumentation domain. To that end, I will now clarify the exact scope of CAAS more closely.

2.5.3 THE SCOPE OF CAAS

So, what is the target object of CAAS? In other words, what is meant by *argumentation* and *argumentation structure*?

Pragma-Dialectics defines argumentation as

a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint (Eemeren and Grootendorst 2004, 1).

This dialectical definition of *argumentation* focuses on an argumentation's social and functional character and preserves the process-product ambiguity of the term (Eemeren 2001, 11). It is, however, for at least two reasons a too narrow definition for a scope specification of CAAS.

First, the singular form of 'standpoint' might be interpreted as suggesting that argumentation is always about one main claim or one standpoint or that argumentation is always argumentation from the perspective of a specific proponent and that dissenting proponents always react with their own argumentation. Here, I will stick to a more inclusive definition of argumentation similar to the term *debate*. Instead of understanding the word *argumentation* as *argumentation for* a standpoint, I prefer to understand it as *argumentation about* some topic, which can encompass very different, even contradicting viewpoints and arguments by possibly different proponents or interlocutors. In particular, *argumentation* is, in this sense, not restricted to the discussion of one standpoint or one claim. Instead, in a complex debate, we will often encounter different central claims, which are related to each other in some way.

While this first point might be a mere terminological issue—after all, what I call argumentation (singular) is, from the perspective of the above cited dialectical definition, simply a set of argumentations (plural) that are connected in some way—the second point is more important. According to the dialectical definition, an argumentation is definitionally dependent on the intentions of arguers. If reasons or arguments are not formulated to convince or persuade others or are not aimed at resolving a difference of opinion, the activity is not considered an argumentation (Eemeren et al. 2014, 2). While I agree that *usually* arguments are put forward to persuade and to dissolve conflicts, there are contexts in which arguments and reasons are formulated with other aims in mind. One obvious example is a text that merely intends to provide an overview of arguments and reasons without taking any side. Authors do not want to convince others in such cases but aim to explain and clarify arguments. Obviously, such texts have an argumentation structure, and CAAS aims to include such cases as potentially interesting target objects of research.

This observation is not an objection against the dialectical definition but rather motivates another, more including characterisation of *argumentation*: Since I do not want to confine the target object of CAAS to argumentation in the narrow sense of the dialectical definition,

I refrain from fixing the specific aims of performing argumentation.

From the perspective of a structural analysis, it is more adequate to focus on the constituent components of argumentation in the definition of *argumentation*. In other words, if we are interested in analysing the structure of argumentation, we should have an understanding of the elements that make up an argumentation. The dialectical definition only mentions “constellation of propositions”, which is not false but also not very specific to argumentation. It is, fortunately, not difficult to name some distinguishing constituent components of argumentation.

I will use the terms *argumentation about x* and *debate about x* interchangeably. A debate or argumentation about some specific topic *x* is a (possibly complex) nexus of claims about questions concerning *x* and different utterances formulated in favour of these claims or as objections, which include arguments, reasons, objections, counterarguments, assumptions, premises and other elements that are presented as a justification.

At this point, it is not important to explicate the different mentioned element types in more detail—for instance, by distinguishing between reasons and arguments. I will use the technical term *argumentative component* as an umbrella term for these elements. What is important is that, first, argumentative components must often be understood as standing in justificatory relations to other argumentative components and, second, that these components can form complex hierarchies.

Regarding the first point: A reason is always to be understood as a reason for something else, for instance, the truth of some statement. Similarly, an objection is always an objection against something. Almost all argumentative components are relational by definition. Consequently, the constituent components of argumentation stand in justificatory relations to other components. I characterise these different relations as *justificatory* relations because one of their relata is presented as (partially) justifying something that stands in a specific semantic relation to the other. For instance, an objection against a statement can be understood as a presented justification for the falsehood or implausibility of the statement.

This first point suggests that the resulting structure of argumentative components and their justificatory relations can form more complex hierarchies. For instance, the refutation of an objection is already a second-order justification in the sense that it is related to something (the objection) that is itself related to something else (the target of the objection). And it does not have to stop there since the refutation can itself be backed up by further justifications or be the target of objections.

With this in mind, I will define the *argumentation structure* of an argumentation as the set of its argumentative components together with their interlinking justificatory relations. At this point, I do not want to take any particular stance towards what types of argumentative components and relations there are. The corresponding decisions depend on one’s analytical aims and vary between argumentation theories. But it is helpful to think of argumentation structure as something that can be represented and visualised by a graph. A graph consists of nodes, which can be visualised by boxes, and relations between these nodes, which can be visualised by edges between these nodes. In a graph that represents the structure of an argumentation—an argumentation graph, or argumentation map—the nodes will correspond to argumentative components, and the edges will represent their justificatory

relations to each other.

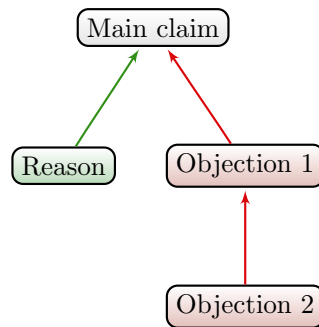


ABBILDUNG 2.1
Simple example of an argumentation map.

Figure 2.1 depicts a simple illustration. There is one node representing a reason, two representing objections, and one node representing the central claim of the argumentation. One edge type represents a relation of justification, visualised as a green arrow, and another represents a relation of objection, visualised as a red arrow. As the figure illustrates, the second objection is already a second-order objection because it does not object to the central claim but objects to the first objection.

In contrast to the dialectical definition of *argumentation*, the suggested definition I introduced to determine the scope of CAAS does not refer to any aspects relating to the social activity of performing argumentation. It abstracts away from who is putting forward which claim or argument, from the actual process of putting forward claims and arguments and from the specific aims of expressing arguments. I take this analytical perspective to be harmless. It merely presupposes that we can speak of argumentative components (e.g., arguments, reasons and claims) and how they relate to each other without considering their genesis and dialogical context.

This is not to say that a structural analysis cannot consider such procedural and dialogical aspects. Quite the opposite: The structural perspective is compatible with a dialectical perspective on argumentation and can be complemented with dialectical aspects. Often, these aspects must be considered to identify argumentative components and their relationships. Additionally, the results of a structural analysis can be used as a starting point to analyse dialectical aspects. For instance, based on a structural analysis, the standpoints of proponents and opponents can be analysed and visualised in an argument map.¹⁷

What is now the scope of the other discussed accounts? Both the political discourse analysis of I. Fairclough and Fairclough (2012) and the deliberative quality index analyse argumentation structure. However, they do so only to a limited extent and are not well equipped to analyse complex argumentation structures—that is, structures that exhibit a multi-level hierarchy of argumentative components.

In order to describe these limits more nuanced, it is helpful to distinguish between micro- and macrostructure of argumentation. The microstructure of argumentation relates to the internal structure of arguments. If we follow the suggestion to understand an argument as

¹⁷For an example, see Betz and Cacean (2012).

a set of premises presented as a justification for a conclusion, the internal structure of an argument is this set of statements. The microstructure can then be analysed and categorised by introducing further concepts, such as argument schemes. The macrostructure, on the other hand, concerns the different justificatory relations between different arguments and counterarguments.

As long as there are no criteria for the individuation of arguments—that is, criteria or definitions that determine which argumentative components belong to one and the same argument—this distinction is somewhat vague. Without going into the details of such criteria at this point,¹⁸ the provided argument explication has already some consequences for the individuation of arguments: An argument contains a part that is presented as a justification (the premises) for the other part (the conclusion). As a consequence, argumentative components that are presented as counterconsiderations, either against the conclusion or against the premise set, will not be considered as part of the argument because they do not have the function of justifying the conclusion.

$PDA_{F\&F}$ focuses on the microstructure of argumentation and offers no methodological means to analyse macrostructure. As elaborated above, $PDA_{F\&F}$ is based on argumentation schemes used to reconstruct the arguments' internal premise-conclusion structure. While it is undoubtedly possible to analyse several arguments with these methodological means, the authors do not tell us how to analyse the interplay between different arguments and counterarguments. This is not to say that they are ignorant about deliberation involving dissent and objections—quite the contrary. But they suggest including such an analysis into the internal structure of a concluding argument that already “[weighs] all these considerations together in order to arrive at a judgement on balance” (I. Fairclough and Fairclough 2012, 12). While I do not want to argue against such a suggestion, I. Fairclough and Fairclough (2012) do not offer any suggestions of how to base such an analysis on the macrostructure of argumentation.

As the analysed examples in I. Fairclough and Fairclough (2012) suggest, the reasons might be that the authors focus on instances of argumentation in the dialectical sense: An utterance or more complex speech act of *one* speaker who argues for their standpoint to convince an audience. It is plausible that even if the speaker anticipates or refers to objections against their standpoint, they will at least implicitly convey to the audience that, in balance, their standpoint is justified. What they do not consider are more complex cases where such a balancing is missing: For instance, in a parliamentary debate, the speech acts of different interlocutors, each who argues for their standpoint, might be related in complicated ways such that an analysis of the macrostructure reveals essential information.

The DQI was explicitly devised to analyse parliamentary debate but performs only slightly better in this respect. Just as $PDA_{F\&F}$, the DQI approach focuses on speech acts of individual speakers who argue for a standpoint—a demand in the DQI terminology—and only marginally considers the justificatory references between speech acts of different interlocutors.

The *content* and *level of justification* concern the microstructure of argumentation. The former demands categorising individual arguments according to their content, and the latter

¹⁸In Chapter 4, I will say more about the individuation of arguments and reasons.

qualifies the justificatory link between premises and conclusion (i.e., the demand). Only the subcategory *respect for counterarguments* has something to do with the macrostructure of argumentation as it applies to reactions to counterarguments. Accordingly, assigning codes for that category presupposes that the annotator can decide whether a speech anticipates or refers to objections against a demand and requires them to classify the speaker's reaction to such objections, which might even include a refutation of the objection. However, the DQI is not devised with the intent to analyse more complicated structures—in other words, cases where the reason hierarchy is complex.

Argument mining is not confined to any particular argumentation model and can, therefore, analyse the macrostructure of argumentation. However, argument mining is (currently) often based on argumentation models confined to the microstructure of argumentation or a restricted version of the macrostructure.

For instance, Habernal and Gurevych (2016) allow the annotation of multiple arguments (143). However, each analysis is confined to the argumentation of a single author and does not consider “the relations that can exist among arguments and their authors in time” (130) and refrain from considering complex argument hierarchies (143). Similarly, C. Stab and Gurevych (2014), who analyse essays, are confined to monological argumentations in favour of one claim by one author. Although their argumentation model includes an attack relation and is thereby not confined to microstructure, their argumentation structures are always trees.

There is nothing wrong with these restrictions since they are motivated by using an argumentation model that is as simple as possible while still expressive enough to model all argumentation structures that can be expected in their chosen target domain. If, however, the target domain is not restricted to a specific type or genre of argumentation, as in the case of CAAS, the argumentation model should not be confined to tree-like structures but should allow arbitrary graph structures.

The approach to argument individuation in some of the used argumentation models is another problematic issue—at least from the perspective of CAAS. They seem to equate *argument* with *argumentation* in the dialectical sense. For instance, Habernal (2014) “examine the relationships between the different components of a given argument, not a relationship that can exist between arguments” (2). Since they include in their premise-conclusion scheme attacking components (pre-claim attack and post-claim attack), they consider counterconsiderations as part of the argument. According to this view, one, or rather *the* argument of an author includes everything they have to say in favour of their standpoint, including anticipated counterconsiderations and refutations.

While this might be considered an irrelevant terminological quirk, the analysis of argumentation structure requires a more fine-grained individuation of arguments. A paradigmatic and fundamental interest in analysing argumentation structure is counting arguments—for instance, as a simplistic measure of argumentative complexity. If, however, the researcher wants to count arguments and reasons within a text, they usually presuppose an understanding of the argument concept that allows an author to formulate more than one argument or reason for their claim. Similarly, the researcher will typically want to count the amount of objections the author anticipated and whether they could invalidate them by further arguments. In other words, the researcher is, in these contexts, not interested in

counting argumentations in the dialectical sense, which will be trivial since the text of *one* author or the complex speech act of *one* speaker in favour of one standpoint represent *one* argumentation (in the dialectical sense). Consequently, CAAS demands a conceptualisation of argument individuation that aligns more with the intuitions I just described. For this and other reasons, I will devote much attention to argument individuation in Chapter 4.

2.5.4 CAAS AND NORMATIVITY

As described above, the DQI and $PDA_{F\&F}$ are normative accounts. They analyse political argumentation and deliberation with the intention to evaluate its quality. The DQI includes evaluative vocabulary in its formulation of category values and is used to investigate driving factors of deliberative quality. $PDA_{F\&F}$ explicitly adopts the critical perspective of critical discourse analysis, which “focuses on what is wrong with a society [...], and how ‘wrongs’ might be ‘righted’ or mitigated, from a particular normative standpoint” (N. Fairclough 2013, 7).

Similar to the DQI and $PDA_{F\&F}$, argument mining is often described as aspiring to evaluate argumentation: “Assessing the degree of strength of an argument, or comparing the quality of arguments to one another, is a goal that many practical future applications will envisage” (Stede and Schneider 2018, 113). But so far, most current and past efforts restrict themselves to identifying argumentation structure in texts without evaluating them. Automatically identifying argumentative components and their relations is seen as a preparatory step for the evaluation of argumentation.¹⁹

CAAS is value-neutral in the following sense: CAAS considers value neutrality as a regulative ideal for analysing argumentation structure. Argumentative elements and their justificatory relations should be identified without evaluating the strengths and weaknesses of arguments.²⁰ On this view, applying the methods of CAAS is intended as a descriptive effort independent of judging how good an argumentation is. To that end, the used categories of the CAAS annotation scheme are not based on normative concepts. In contrast, the DQI already evaluates justificatory relations by introducing category values that judge them according to their completeness (see above). Increasing the amount of complete relations will increase the overall deliberative quality. In CAAS, however, the justificatory relations used to characterise an argumentation structure are purely descriptive and not evaluative about the actual inferential quality of the link between premises and conclusions.

However, the requirement of being value-neutral does not mean that CAAS cannot be used in normative contexts. Quite the contrary, CAAS can be complemented with evaluative concepts. It is possible to supplement the category system of CAAS with additional categories that evaluate single arguments or whole argumentations.

¹⁹There are, however, some first preliminary attempts to approach the goal of automatically evaluating argumentation. For an overview, see Stede and Schneider (2018), Chapter 8.

²⁰Sometimes, an evaluative perspective towards arguments is necessary to resolve ambiguities in the identification of argumentation structure. The desideratum of value freedom is regulative to the extent that the evaluation of arguments should not play any role in CAAS unless such an evaluation is necessary to identify argumentation structure. In Chapter 4 (see also Maxim 10 on page 268), I will specify under which conditions such an evaluation is necessary.

2.5.5 CAAS AND RELIABILITY

The final defining feature of CAAS is its faithfulness to reliability. In its most general form, reliability can be understood as a form of trustworthiness towards the data-making process (the categorisation of coding units) and its results (the assigned category values). On this trustworthiness view, “reliability is the ability to rely on data that are generated to analyze phenomena a researcher intends to study, theorize, or use in pursuit of practical decisions” (Krippendorff 2016, 139). In other words, a reliable data-making process produces data that allows the content analyst to infer something about the target domain.

Reliability criteria formulate conditions under which the content analyst can trust such inferences based on the generated data. There are different ways to elaborate this line of reasoning and, accordingly, different possibilities to formulate reliability criteria.

The most prominent suggestion in content analysis is to conceptualise reliability as reproducibility. The basic idea is this: If the data-making process is affected by biases, misunderstandings of coding instructions or other idiosyncracies of annotators, the generated data can hardly be trusted. In such cases, it is difficult to judge whether observed differences in annotations correspond to differences in the observed phenomena or result from differences in subjective choices of annotators. Hence, if coding results are not intersubjective—that is, if they depend on subjective decisions that can differ between annotators—it is difficult to infer something about the target domain from the generated data.

Reproducibility tests are a simple way to reveal violations of intersubjectivity: If the repeated categorisation of one and the same coding unit by different annotators yields different results, these differences can only be explained by differences in subjective choices of annotators. One suggestion is, therefore, to demand that repeated annotations of the same coding units should result in the same categorisations.

However, this suggestion requires *perfect* agreement between different annotators on the same coding units. Suppose the content analyst performs a reliability test on, let’s say, one hundred coding units with two annotators. Even if both annotators disagree on only one coding unit, the data-making process is considered unreliable. Since such perfection is rare, the version of reliability-orientated content analysis presented in standard textbooks is based on a weaker notion of reliability (Krippendorff 2011, 95). This version, which I will call the *received view of reliability-orientated content analysis* (or, in short, the *received view*), conceives reliability as a graded concept with perfect agreement on the upper end and agreement that is not better than throwing dice on the lower end. On this view, the reliability of the data-making process is measured by a numerical value (between zero and one) and reliability requirements are formulated as threshold criteria: Only if the reliability coefficient exceeds a certain threshold, the data-making process is considered sufficiently reliable.²¹

The DQI and most human annotation efforts from argument mining are committed to the received view. They regard reliability as an essential desideratum and assess reliability by

²¹There are different suggestions to give this idea a precise mathematical form and different threshold criteria, which fit different contexts and different methodological views. The basic idea is elaborated in Section 3.2.4.

replicating annotations and calculating reliability coefficients.

$PDA_{F\&F}$, on the other hand, is not committed to reliability constraints, which can be explained in the following way. As elaborated above, political discourse analysis is not based on a fixed category system. Instead, analysts can, to some extent, decide from case to case which analytical or evaluative perspective they deem important. In particular, they are not confined to using a fixed set of category values to characterise their target material. Accordingly, the results of replications are not expected to coincide. Rather, subjective decisions are allowed to result in differences in outcomes. Instead of demanding reliability in the form of reproducibility, discourse analysts emphasise transparency as an important requirement. Analysts should explain their subjective choices and make them explicit.

CAAS is reliability-orientated content analysis. However, it is not based on the received view but employs an alternative conceptualisation of reliability. The main contribution of this work is an elaboration of this alternative. The need to develop an alternative stems from a challenge, which will be described in the next section.

2.6 HERMENEUTICAL UNDERDETERMINATION

The analysis of natural-language argumentation involves a systematic study of texts, which includes at least the identification of arguments and reasons as well as their connecting justificatory relations. Usually, this task and all subsequent steps require the consideration of semantical and pragmatic aspects that cannot be read mechanically from the text but demand considering co-text, background knowledge of analysts qua them being competent speakers and additional extra-linguistic context information. Consequently and as emphasised by argumentation theorists, the analysis of natural-language argumentation is a hermeneutical process, and the results are an interpretation: The identification of argumentative units demands the interpretation of context-dependent linguistic cues (Fisher 2004, 16). Similarly, classifying the internal structure of argumentation is often difficult without relying on contextual and other pragmatic factors (Eemeren and Grootendorst 2004, 4). Finally, the reconstruction of arguments is, in many cases, underdetermined—that is, allows for different interpretations—independent of whether analysts use known argumentation schemata (Eemeren and Grootendorst 2004, 4; Walton 2012, 33) or if they prefer more scheme-independent reconstructions methods (Betz 2010, 180).

This more or less obvious observation poses significant questions to the extent that all three accounts portray themselves as scientific methods, which are, consequently, bound to certain scientific standards. If the process of analysing argumentation is an interpretation, are the obtained results without alternatives? Are the results subjective and vary between different analysts? Is there an awareness of interpretational leeway? How do the different accounts deal with this kind of hermeneutical underdetermination? Does hermeneutical underdetermination lead to a violation of scientific criteria such as reliability?

Discourse analysts emphasise the qualitative and interpretative nature of analysing discourse (Greckhamer and Cilesiz 2014, 13; Hardy, Harley, and Phillips 2004, 1; Titscher et al. 2000, 146). In particular, interpretations are open-ended: New contexts and information can change the results of a discourse analysis.

Gee and Handford (2012) discuss this so-called *frame problem* and suggest the following solution: Analysts should widen the considered context “until the widening appears to make no difference to our interpretation” (Gee and Handford 2012, 5). The underlying assumption is that the context can be expanded until a point of saturation, where any additional information has no further bearing on the admissibility of interpretations. This strategy has, however, some problems. First, it is unclear whether it will always yield a unique interpretation: Expanding the context may introduce defeating information, for instance, information that is inconsistent with previous information. In such cases, the interpretational leeway might even increase. Moreover, even if widening the context leads to a point where additional information has no further relevance to the admissibility of interpretations, there might still be more than one admissible interpretation. Second, researchers usually work under limited resources and often deal with other practical hindrances, limiting their possibilities to include arbitrary broad context information.

In consequence, discourse analysts will, at least in some cases, be confronted with some indeterminacy in the interpretation of their data. But is this a problem? Gee (2011) maintains that agreements among analyst matters and that interpretational openness might raise problems for the validity of research efforts. However, he also points out that disagreement is not conclusive for invalidity. Transparency is not only a complementary requirement but is often endorsed as partially compensating for a lack of reproducibility. It demands that it should be intelligible and recognisable how discourse analysts arrive at their conclusions (Titscher et al. 2000, 164; Greckhamer and Cilesiz 2014).

But is transparency always enough in the face of interpretational openness in discourse analysis? This question can hardly be answered in general since it will depend on the specific research questions, which might differ largely between different discourse analytical efforts. Since discourse is pervasive in all social and cultural contexts, discourse analysis can be applied in diverse disciplines, such as sociology, history, anthropology, archaeology, economics, human geography, linguistics, management science and communication science (Gee and Handford 2012). Political discourse analysis, in particular, seeks to make genuine contributions to political science (Dijk 1997). It stands to reason that in some of these research contexts, interpretational indeterminacy is problematic.²²

Discourse analysis in itself provides, however, no specific methodological means to deal with hermeneutical underdetermination—besides the already discussed strategy to widen the extra-linguistic context.²³ Not only does discourse analysis lack the means to decrease interpretational indeterminacy sufficiently, but it also doesn’t provide any tools to assess the severity of interpretational indeterminacy.²⁴

How does the DQI deal with hermeneutical underdetermination? Similarly to discourse analysis, there is an awareness of the interpretational character of analysing argumentation:

²²In Chapter 4, I will elaborate generally why and under which conditions interpretational indeterminacy is problematic.

²³A variant of this strategy, which has the same problems, is suggested by Jaipal-Jamani (2014) who proposes a form of transdisciplinary convergence. Interpretational indeterminacy should be minimised by approaching the research problem from different disciplinary perspectives.

²⁴There are some notable exceptions. For instance, Hux et al. (1997) discusses using reliability coefficients in discourse analysis to measure the agreement among analysts.

Assessing the quality of discourse requires interpretation. One needs to know the culture of the political institution, the context of the debate, and the nature of the issue under debate, to get a true understanding of how actors in the institution use and interpret language. [...] Our project requires careful human judgment. [...] Both the qualitative and our approach require extensive contextualized interpretation. [...] the DQI incorporates coding guidelines that should help coders arrive at similar interpretations of a discourse. This allows for a certain degree of intersubjectivity, although it does not impose it because different coders can have legitimate differences in interpretation. (Steiner et al. 2004, 60).

The DQI suggests a different solution to the problem of hermeneutical underdetermination than the discourse analyst. Instead of widening the (informational) context used to analyse coding units, the DQI follows the received view of reliability-orientated content analysis: The annotation scheme should be based on precise category definitions such that annotators arrive at the same or similar interpretations. The idea is, therefore, to minimise interpretational differences among annotators by improving the annotation scheme and their training.

In contrast to discourse analysis, the received view does not only suggest a means to minimise interpretational leeway but employs a tool for monitoring how successful this strategy is. As described above, reliabilities can be calculated and measure disagreements among annotators. While there might be different causes for why annotators disagree—and thereby different causes for low reliability values—interpretational differences are one contributing factor. In this way, reliability values can be taken as indicating whether annotation instructions are precise enough to minimise the interpretational leeway sufficiently. Consequently, if reliability values exceed the recommended thresholds, interpretational differences can be considered sufficiently small.

The human annotation efforts in the context of argument mining are similarly based on the received view and follow the already described stance towards hermeneutical underdetermination: The interpretative nature of identifying argumentations structure is acknowledged;²⁵ at the same time, precise and detailed annotation guidelines should make sure that annotators reach a sufficiently high level of intersubjectivity—that is, a sufficiently high level of interannotator agreement.

The argument mining community seems to be optimistic that this strategy is successful. At least, hermeneutical underdetermination is seldom mentioned as a problematic issue.²⁶ There are, however, some notable exceptions: According to C. Stab et al. (2014), hermeneutical underdetermination, or in their words, the ambiguity of argumentation structures, “represents a major challenge for argument annotation studies and consequently the creation of reliable gold standards for argumentation mining” (7–8). Similarly, Kirschner, Eckle-Kohler, and Gurevych (2015) note that ambiguity is “one of the main challenges in identifying argumentative relations on a fine-grained level in scientific publications” (9).

²⁵See, e.g., Mochales and Ieven (2009).

²⁶For instance, the review Lawrence and Reed (2019) lists major challenges for argument mining in their conclusion. Hermeneutical underdetermination is not one of them. Similarly, the textbook Stede and Schneider (2018) does not discuss the issue of interpretational indeterminacy.

What is more, these authors seem to be less optimistic about the possibilities to narrow down the degree of interpretation and even suggest that there is a methodological problem with the received view of reliability-orientated content analysis when applied to analyse argumentation: The received view labours under the assumption that there is only one correct interpretation, although the identification of argumentation structure does often allow for different interpretations.²⁷ Kirschner, Eckle-Kohler, and Gurevych (2015) write that “all the measures used to calculate IAA [interannotator agreement] assume that there is one single correct solution to the annotation problem. Actually, we believe that in many cases several correct solutions exist depending on how the annotators interpret the text” (9).

The received view is thereby ill-equipped to deal with cases where interpretational indeterminacy cannot be dissolved. At least, it condemns hermeneutical underdetermination as inconsistent with a reliable annotation process. In the first instance, the challenge of hermeneutical underdetermination can now be formulated as follows: How can the content analyst achieve a reliable data-making process if the degree of interpretation cannot be further reduced by improving the annotation scheme?

2.7 STRUCTURE OF THE THESIS

The main contribution of this work is the development of a methodological framework that meets the challenge of hermeneutical underdetermination in the context of CAAS. The central idea will be laid out along the following steps: I will, first, elaborate on the role of reliability in content analysis, show that the received view is not able to deal with an irreducible degree of interpretation and sketch an alternative conceptualisation of reliability based on an analogy from dealing with measurement errors in natural sciences (Chapter 3). I will, second, motivate that the identification of an argumentation’s macrostructure might allow for different interpretations even if we use everything argumentation theory has to offer to design an annotation scheme with as precise category definitions as possible (Chapter 4). Finally, I will use the general considerations laid out in Chapter 3 to devise a probabilistic conceptualisation of reliability for the case of CAAS and discuss some technical difficulties (Chapter 5).

In the subsequent chapters, I will unfold these steps into the following line of reasoning.

2.7.1 CHAPTER 3: RELIABILITY AND HERMENEUTICAL UNDERDETERMINATION

My central aim in Chapter 3 is to sketch a generic solution to the problem of hermeneutical underdetermination in content analysis. I characterise content analysis as a research technique that is based on applying a coding scheme to annotate meaning-bearing matter—most often texts—to generate data, which is used to infer something relevant to answer a research question (3.1).

The categorisation of coding units—that is, the application of the annotation scheme—by annotators can be assessed by two different desiderata: The validity of the annotation

²⁷Whether the received view is indeed committed to this uniqueness assumption, is an intricate issue, which I will analyse in Section 3.3.4.

process measures the extent to which annotators categorise coding units correctly as determined by the category definitions of the annotation scheme (3.2.1). Reliability can be interpreted as an indicator of validity and measures the extent to which the content analyst can trust the results of the annotation process. This trustworthiness view of reliability is usually conceptualised as a form of reproducibility (or intersubjectivity): It demands that annotations should not depend on the particular annotator (3.2.2).

The desideratum of reliability implies important specific requirements for the annotation process (3.2.3) and is conceptualised as a graded notion by the received view of reliability-orientated content analysis (3.2.4). On this view, the reliability of the data-making process can be assessed in the following way: The coding effort is, first, reproduced by different annotators. The resulting annotations are then used to calculate reliability values based on chance-corrected reliability measures, which assess to what extent coders perform better than chance. If, finally, these values exceed recommended threshold values, the data-making process is considered (sufficiently) reliable.

The question is whether content analysis is able to deal with hermeneutical underdetermination (3.3).

There are two different types of content analysis (3.3.1), which differ in their stance towards the role of reliability. What is often referred to as *quantitative content analysis* is the received view of content analysis and acknowledges reliability as a central requirement. *Qualitative content analysis*, on the other hand, is not reliability orientated and emphasises other constraints such as transparency of the annotation process (3.3.2).

The difference between manifest and latent content, which is often presented as another distinguishing feature between both types of content analysis, helps to explain their diverging views about reliability: If the reliability of the data-making process can only be increased by decreasing the evidential relevance of the annotation scheme, the qualitative content analyst is more willing to make concessions to reliability than the quantitative content analyst (3.3.3).

The differences between both types of content analysis have significant consequences for how they deal with hermeneutical underdetermination (3.3.4): Quantitative content analysis suggests minimising interpretational leeway by providing as precise category definitions as possible. According to one interpretation, quantitative content analysis even presupposes that there is always only one correct categorisation for each coding unit. Consequently, quantitative content analysis cannot meet the challenge of hermeneutical underdetermination. Qualitative content analysis, on the other hand, acknowledges the existence of an irreducible degree of interpretation but gives up on reliability and is, thereby, unable to meet the challenge of hermeneutical underdetermination.

The central tenet of Chapter 3 is that both discussed paradigms of content analysis do not exhaust the set of available options. Instead, I propose a probabilistic conceptualisation of reliability that meets the challenge of hermeneutical underdetermination (3.4).

To approach the basic idea, it is helpful to employ another perspective on reliability (3.4.1): Reproducibility (or intersubjectivity) is important because the content analyst must be sure “that the variance of the generated data is explainable by the differences that coders detected among the phenomena they recorded” (Krippendorff 2016, 139). The content

analyst must trust that a difference between coding results corresponds to a difference in the phenomena and is not the result of divergent interpretations. If this requirement, which I will call *phenomenon sensitivity*, is violated, it is difficult to infer something about the target phenomenon based on the generated data.

Phenomenon sensitivity comes in two variants: *Strong phenomenon sensitivity* demands that *every* difference in the generated data is only explainable by a difference in the phenomena and henceforth not by a difference in interpretation. Strong phenomenon sensitivity is, therefore, only satisfied if only one correct interpretation exists. The question is now whether there is a weakening of strong phenomenon sensitivity that allows the content analyst to reliably infer something about the target phenomenon even under conditions of hermeneutical underdetermination.

It is helpful to understand that the requirement of strong phenomenon sensitivity is not even satisfied in the natural sciences (3.4.2). To some extent, the replication of measurements will always produce a variance in measurement results. The scientist can, therefore, not fully trust single measurements. The *theory of errors* provides a probabilistic conceptualisation of this uncertainty: The variance in measurements of one quantity value is interpreted as the result of random errors. These random errors are why the measurements of one quantity value are distributed around the value. Accordingly, the differences between the measurements of two quantity values will be caused by the differences between these values and by random errors. Strong phenomenon sensitivity is, therefore, violated by measurements in natural sciences: The researcher cannot always be sure whether observed differences result from random errors alone or are also caused by differences in the phenomena.

That does, however, not preclude the researcher from inferring something about their target phenomena based on measurements. In other words, measurements in natural sciences satisfy some weakening of strong phenomenon sensitivity: The probabilistic conceptualisation of random errors allows the formulation of conditions under which inferences from the data to target phenomena are (statistically) valid. Random errors are probabilistically distributed. The researcher estimates this distribution based on repeated measurements. This probabilistic quantification of random errors enables them to decide whether differences in measurement results cannot be explained by random errors alone but can be taken to indicate a difference in the phenomena.

The theory of errors can be used as an analogy to conceptualise a similar probabilistic approach of “measurements” in content analysis. The generic recipe to meet the challenge of hermeneutical underdetermination by specifying such a probabilistic approach is this (3.4.3): We, first, need a probabilistic description of the annotation process, which is tied to a specific interpretation of these probabilities. I argue that these probabilities are best understood as physical probabilities tied to a chance set-up. On this view, one specific annotation is the result of randomly picking one annotator from the population of properly trained annotators and letting them annotate the target material. Second, this general probabilistic conceptualisation should be modelled by a specific statistical model that includes all relevant probabilistic assumptions. Finally, based on the statistical model, the analyst can specify conditions under which inferences from differences in annotations to differences in the phenomena are (statistically) valid.

2.7.2 CHAPTER 4: HERMENEUTICAL UNDERDETERMINATION IN ARGUMENTATION ANALYSIS

What is the extent of hermeneutical underdetermination in CAAS? In Chapter 4, I address this pivotal question by comparing three accounts of analysing argumentation structure—one non-reconstructive and two reconstructive accounts (applied formal logic and informal logic). The main argumentation is this: All three accounts cannot guarantee a unique identification of argumentation structure. Although reconstructive approaches can provide precise concepts and profound criteria to analyse argumentation, they do not necessarily help to narrow down the degree of interpretation. In particular, the degree of interpretational distance between annotators depends on whether they have a sufficiently large overlap of their background knowledge that they use to reconstruct arguments. In terms of narrowing down the degree of argumentation, the two reconstructive accounts have, therefore, not necessarily an advantage over the non-reconstructive account. These claims will be justified by analysing examples and on theoretical grounds. If there are no other motivations than narrowing down the degree of interpretation, the content analyst should, therefore, ground their annotation scheme on a non-reconstructive approach since it is less demanding in terms of annotator training.

This argument will be elaborated along the following lines:

I begin by describing the nature of hermeneutical underdetermination that is relevant in the context of CAAS (4.1). *Node ambiguity* is ambiguity in connection to the identification of justificatory relevant text segments and their justificatory role (4.1.1). *Underdetermination of granularisation* is ambiguity with respect to deciding how many arguments and reasons are expressed in a longer text segment (4.1.2). Finally, *relation ambiguity* is ambiguity with respect to the identification of justificatory relations between argumentative components (4.1.3).

I broadly distinguish *reconstructive* and *non-reconstructive accounts* of argumentation analysis (4.2). A non-reconstructive analysis is based on our ordinary-language competence to comprehend argumentation without reflecting on the underlying concepts and without relying on argumentation-theoretic accounts (4.2.1). Reconstructive approaches, on the other hand, analyse argumentation by reconstructing arguments in their premise-conclusion form (4.2.2). They comprise Applied Formal Logic and Informal Logic.

Applied Formal Logic is one central paradigm in argumentation theory (4.2.3). It requires reconstructing arguments as deductively valid (*reconstructive deductivism*) and draws heavily on concepts from formal logic. An argument is called deductively valid if it is truth-preserving. This notion is explicated in formal logic by introducing formal languages, which requires the applied logician to formalise natural language arguments.

According to *Informal Logic*, most natural language arguments are not deductive but defeasible (4.2.4). Consequently, they have to be judged by criteria other than deductive validity. Arguments can be reconstructed with the help of argument schemes; they can be evaluated by invoking critical questions and assessing the acceptability, relevance and sufficiency of premises.

How do these different accounts perform with respect to minimising the degree of interpre-

tation in CAAS (4.3–4.4.7)?

Although arguers may make their argumentation structure maximally explicit by extensive signposting, they will often not do so. As a consequence, a non-reconstructive analysis of argumentation structure will—depending on the text at hand—allow for different interpretations (4.3).

First, annotators might face node ambiguity (4.3.1). There are different challenges to identifying text segments as justificatory relevant and to discern their justificatory role: It is often challenging to distinguish justificatory roles from other roles, such as providing causal explanations or elaborative illustrations. Additionally, linguistic cues can be ambiguous themselves. What they signpost will often depend on the context.

Identifying and categorising justificatory relations can also allow for different interpretations (4.3.2). While it is usually possible to uniquely classify justificatory relations as either an attack or support relation, the target or source of a relation may be underdetermined.

Finally, the individuation of arguments and reasons faces two paradigmatic problems (4.3.3). First, the arguer might repeat arguments and reasons by rephrasing them. Accordingly, it can be difficult to realise whether different text segments are intended as additional arguments or repetitions. Second, the analyst might be confronted with *horizontal underdetermination*: If the arguer presents different text segments as justifying one claim, the analyst has to understand whether these text segments make up one or several arguments or reasons.

Reconstructive and non-reconstructive accounts of argumentation analysis are complementary. Usually, the analyst will ground their reconstructive analysis on a non-reconstructive preparatory work. The question is, therefore, whether the reconstruction of arguments can help to meet some of the hitherto mentioned challenges. I will, first, discuss node ambiguity and granularisation of argumentation (4.4.1–4.4.4); then, I will move on to consider more specific accounts that can deal with relation ambiguity, which relates to the macrostructure of argumentation (4.4.5–4.4.7).

The reconstruction of arguments as envisaged by applied formal logicians cannot help substantially to narrow down the degree of interpretation with respect to the individuation of arguments (4.4.2). The problem is that there are often different possibilities to reconstruct arguments. According to the principle of charity, the analyst should pick the most charitable reconstruction in terms of argument strength and, in particular, premise plausibility. However, judging the plausibility of premise candidates depends on the analysts' background knowledge. To the extent that this background knowledge varies between analysts, they will arrive at different interpretations as long as annotators analyse texts independently.

In informal logic, the reconstruction and evaluation of arguments are grounded on sufficiency and premise relevance. Each of these notions provides the informal logician with a different way to determine the granularisation of argumentation. The analyst can either use the concept of sufficiency to individuate arguments or the concept of relevance to individuate reasons.

The individuation of arguments is based on argument schemes and inference rules (4.4.3).

Argument schemes describe recurring forms of argumentation. By instantiating the placeholders of such a scheme, the analyst arrives at an argument reconstruction that represents an individuated argument. I argue that argument individuation based on argument schemes might diverge between different analysts for the following reasons: First, the analyst will encounter arguments that do not fit any of the given schemes since the prominent accounts are incomplete. In this case, they have to resort to other reconstruction strategies. Second, an argumentation can sometimes be reconstructed with different schemes, resulting in different possibilities to individuate arguments. Finally, argument schemes are not based on a systematic theory of argument reconstruction. Consequently, there might be good reasons to reconstruct a specific argument in a way that differs from a reconstruction guided by argument schemes. The alternative of using inference rules will encounter similar difficulties as the individuation of arguments in the paradigm of applied formal logic.

The informal logician uses the prominent convergent-linked distinction to distinguish different types of arguments. As it turns out, the distinction can be used as a criterion to individuate reasons (4.4.4). I will discuss two different criteria, which I borrow from Freeman (2011). An *intuition-based criterion* that grounds the individuation of reasons in our judgements of independent probative relevance and a *syntactical criterion* that grounds the individuation of reasons in the arguer's inference rules. The intuition-based criterion is helpful and easy to apply—in the sense of neither depending on sophisticated theoretical baggage nor the reconstruction of arguments. However, it does not exclude cases of horizontal and vertical hermeneutical underdetermination in reason individuation. The syntactical criterion demands reconstructing arguments. The identification of reasons hinges crucially on the inference rules the analyst attributes to the arguer. Since there is no canonical way to determine these inference rules uniquely, the syntactical criterion cannot exclude cases of horizontal and vertical hermeneutical underdetermination in reason individuation.

How do reconstructive accounts perform with respect to relation ambiguity? Are they able to narrow down the corresponding degree of interpretation?

I will discuss two different accounts of macrostructure: the theory of dialectical structures (Betz 2010) as a representative of the applied formal logic paradigm (4.4.5) and Carneades (T. F. Gordon, Prakken, and Walton 2007) as a representative of the informal formal logic paradigm (4.4.6).

The theory of dialectical structures (Betz 2010) introduces justificatory relations between arguments that are based on deductive argument reconstructions. It is parsimonious—in that it defines only two relations between arguments that are reduced to basic semantic relations between statements—but still expressive enough to analyse the macrostructure of complex debates. The theory can even help to disambiguate apparent cases of relation ambiguity. However, choosing between different interpretations of the macrostructure will, again, involve the principle of charity. Hence, interpretations between analysts might vary inasmuch as their background knowledge diverges. Therefore, this theory cannot guarantee to eliminate the variability between analysts in the identification of debate macrostructures.

The analysis of argumentation structure in informal logic is intricately linked to argument diagramming—the visual representation of argumentation structures. By now, argument diagramming techniques are used in diverse scientific contexts such as informal logic, legal

reasoning and computer science. Most of the techniques used in informal logic focus on analysing the internal structure of arguments. But many accounts support, at least partially, the analysis of how different arguments are related to each other as, for instance, Toulmin's -Toulmin (1958) very influential model.

Carneades model of argument (T. F. Gordon, Prakken, and Walton 2007), which is primarily designed for the context of legal reasoning, can be considered a paradigmatic account of analysing macrostructure from the viewpoint of informal-logic since it is based on Walton's arguments schemes and their associated critical questions. I will argue that while Carneades can circumvent some of the problems to distinguish between "normal" premises and critical questions, on the one hand, and between undercutting and undermining defeaters, on the other hand, its reliance on proof standards to distinguish between different types of justificatory relations is a drawback when it comes to relation ambiguity. Without a systematic and complex theory of proof standards that can account for the heterogeneity and extensiveness of ordinary-language argumentation, different analysts might come to different interpretations in analysing macrostructure.

The theoretical considerations laid out in Sections 4.4.5 and 4.4.6 will be complemented in Section 4.4.7 with a small case study. I will use the example of affirmative action to illustrate how a reconstructive analysis—be it based on applied formal logic or informal logic—can dissolve relation ambiguities that prevail during a non-reconstructive analysis of macrostructure. The basic strategy is to reconstruct different versions of one argument that correspond to the different possibilities of interpreting its justificatory relations to other arguments and, then, to compare the different versions according to hermeneutical principles such as the principle of charity. Reconstructions that perform best with regard to these criteria determine the most adequate interpretation of their justificatory relations.

It turns out that the differences between both paradigms do not matter much for the disambiguation of relation ambiguities. If anything, Carneades model might be faced with additional degrees of interpretational freedom since it introduces subcategories of justificatory relations that are irrelevant to the theory of dialectical structures (the distinction between undercutting and undermining defeaters on the one hand and the distinction between premises of an argument scheme and its critical question on the other hand).

While the described strategy is, in principle, able to narrow down the degree of interpretation, it inherits all of the difficulties already discussed in previous sections. The comparison of different reconstructions according to the hermeneutical principles might involve trade-offs and depends on the background knowledge of the analyst. Consequently, different analysts might come to different conclusions in their interpretation of justificatory relations between arguments.

Based on the findings arrived at in the previous sections, I will in Section 4.5 describe the most important implications for devising a minimal category system for CAAS, which can be found in Appendix A.1. This annotation scheme has three main features. First, it focuses on the macrostructure of argumentation and can be used to analyse any argumentative text. The argumentation structure will be modelled by a graph and is not confined to tree structures. Second, the category system is confined to relational categories between text segments. There are only two justificatory relations—a support and an attack relation. Finally, the annotation scheme is based on a non-reconstructive approach.

2.7.3 CHAPTER 5: CONTENT ANALYSIS OF ARGUMENTATION STRUCTURES

Since the analysis of argumentation structure may often face an irreducible degree of interpretation (Chapter 4), CAAS has to meet the challenge of hermeneutical underdetermination. The main aim of Chapter 5 is to apply the described generic probabilistic conceptualisation of reliability (3.4) to the case of CAAS. In other words, I will explain how the content analyst can satisfy the reliability requirement in the CAAS context (5.3), even though the analysis of argumentation structure is connected to an irreducible degree of interpretation. This probabilistic conceptualisation of reliability requires two preparatory steps: First, I will devise a disagreement measure that can be used to estimate the extent of hermeneutical underdetermination of a text's argumentation structure (5.1). Second, I will solve the practical problem of aligning coding units of different annotators (5.2), which is needed for applying the disagreement measure and for conceptualising the statistical model I suggest for CAAS (5.3.1).

Naturally, an estimation of the extent of hermeneutical underdetermination should be based on a disagreement measure between annotation results (5.1).²⁸ In content analysis, the most simple idea to conceptualise the disagreement between two annotations is to take the relative share of coding units on which annotators disagree. This idea can, however, not be directly applied to CAAS. There are two problems: First, coding units are not predefined. Rather, annotators unitise their target material—that is, they decide for themselves where coding units (i.e., argumentative units) start and end. Accordingly, there can be disagreements in unitising, and it is, therefore, unclear what it means to compare categorisations of *one and the same* coding unit between *different* annotations. Second, instead of categorising coding units (by monadic categories), CAAS uses relations between coding units as categories. Consequently, the disagreement measure cannot be based on comparing categorisations of coding units but has to consider relations between coding units.

Fortunately, these hurdles are not specific to CAAS, and I can, therefore, draw on existing suggestions to deal with these problems. In Section 5.1.1, I will discuss several approaches to assess the reliability of unitising and, in particular, how they suggest measuring the observed difference between annotation results. The overlap-based family of α coefficients and the reformulation of unitising as a token-labelling problem cannot be used in CAAS since they cannot account for differences in the granularisation of argumentation, which is highly relevant in CAAS. Instead, I opt for an alignment-based approach. Such an approach is grounded on criteria that determine whether two coding units of different annotators are considered to represent the same semantic unit. Based on such an inter-annotator identification of coding units, the content analyst can compare whether annotators disagree in their categorisation of one semantic unit.

An alignment-based approach can be responsive to differences in granularisations. Additionally, such an approach is able to separate two types of disagreements: Positional disagreements—that is, disagreements in unitising (Δ_u)—and (categorical) disagreements in argumentation structure (Δ_r).

²⁸Based on such a measure and a probabilistic conceptualisation of CAAS (5.3), we could define the degree of interpretation as the expected spread or variance of all correct interpretations. The expected spread, in turn, could be estimated based on the spread within a sample of replicated annotations.

After discussing several suggestions to measure disagreements in annotating relations between coding units (5.1.2), I will suggest measuring categorical disagreements Δ_τ by comparing the graphs representing the argumentation structure based on a graph-edit distance (5.1.3).

The graph-edit distance Δ_τ presupposes an alignment of argumentative units among annotators. We have to determine whether an argumentative unit identified by one annotator represents the same argumentative component as the argumentative unit of another annotator.²⁹ This inter-annotator identification of argumentative units is a practical problem, which I will overcome by adapting the γ approach developed by Mathet, Widlöcher, and Métivier (2015) (5.2). The basic idea of this alignment procedure is to align coding units so that the resulting positional and/or categorical disagreements are minimised (5.2.1). A modification of this approach is necessary since it produces some incorrect alignments in the context of CAAS. First, in some cases, the γ approach will align non-overlapping argumentative units. Second, the γ approach is not able to properly (not) align argumentative units in the case of a diverging choice as to the granularisation of argumentation.

I will solve these problems by departing from the γ approach in two points (5.2.2). First, I will adapt their used pairwise distance function. Second, I will formulate some additional side constraints the alignment procedure has to satisfy.

In Section 3.4.3, I will provide a generic probabilistic conceptualisation of reliability. This enables the content analyst to detect differences in the phenomena of interest even if annotators diverge in their coding results due to interpretational differences. In Section 5.3, I will elaborate on how these preliminary ideas can be applied to analyse argumentation structure.

The general idea is this: The result of annotating the argumentation structure is an argumentation graph—that is, a directed graph consisting of argumentative components as well as attack and support relations between them. Since such an annotation might be hermeneutically underdetermined, different annotators might end up with different argumentation graphs without doing anything wrong. The space of all interpretations is then supplemented with a probability function, which assigns a probability value to each argument map. In other words, the data-making process for a specific text is modelled by a probability function over the outcome space of coding results. Based on this conceptualisation, is it possible to apply statistical inferences to decide whether differences in annotations are significantly different to infer that this difference corresponds to a difference in the phenomena.

The elaboration of this general idea will proceed along the following steps:

First, I will provide a mathematical specification of the outcome space and describe an appropriate statistical model for CAAS (5.3.1). Statistical inferences deduce properties of a probability function from an observed sample. These inferences are usually based on statistical assumptions, referred to as the statistical model. For CAAS, I will equate the outcome space with a finite set of argumentation graphs. This outcome space will be rewritten as a cartesian product so that certain independence assumptions hold.

Second, I will describe a specific statistical method that is suitable for our research context

²⁹Cf. Footnote 9.

(5.3.2). As elaborated in Chapter 3, satisfying reliability—in the form of weak phenomenon sensitivity—demands deciding whether observed differences are significant enough to justify the conclusion that there is a difference in the phenomena. Suppose, for instance, the amount of argumentative components is relevant, and we want to investigate whether two texts differ regarding their expected number of argumentative components. To that end, we have to generate a sample of argument maps for each text and decide whether the observed difference between the mean number of argumentative components is significant enough. I will closely follow the description from Section 3.4.2 by applying the method of significance testing. The main conceptual challenge is that the null hypothesis—that is, the hypothesis that both texts are not different—is complex. In contrast to a simple hypothesis, a complex hypothesis is not represented by one specific probability function but by a whole set of probability functions. Consequently, the concept of the p -value must be generalised to such cases.

Finally, the mathematical considerations to describe the outcome space and the statistical model as well as the elaboration of the statistical inference method will reveal a certain complexity (5.3.3). The size of the model's parameter space and the size of the sample space will render the actual calculation of p -values computationally complex. I will sketch some solutions for how to deal with these complexities. However, it is beyond this work to sufficiently solve these problems. Instead of demonstrating that the discussed statistical methods can be applied to realistic cases, I will point to further research.

3. RELIABILITY AND HERMENEUTICAL UNDER-DETERMINATION

Krippendorff characterises content analysis as “a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use” (Krippendorff 2004a, 18). One distinguishing feature of content analysis compared to other empirical approaches is the phenomenon of interest—namely, communication or, rather, its manifestations in the form of text messages, paintings, music, movies or any other entities intended to convey meaning. However, not every analysis of text messages is a proper form of content analysis since content analysis is a scientific method and therefore bound to specific requirements.

The main aim of this chapter is to equip content analysis with the ability to deal with what I call *hermeneutical underdetermination*: Understanding meaning is a process of interpretation. And the results of this process can vary between people. In other words, different content analysts might get different results due to interpretational differences. Hermeneutical underdetermination is a challenge for content analysis as a scientific method since the possibility of subjective differences threatens reliability.

I will proceed along the following steps: First, I will provide a brief description of what content analysis is (3.1). I will then move on to characterise different desiderata of the scientific method in general and elaborate on the role of validity and reliability in content analysis (3.2). Third, I will distinguish between two types of content analysis, namely quantitative and qualitative content analysis, and show that the latter does not prioritise reliability in contrast to the former (3.3). This different stance towards the importance of reliability has ramifications for how both types deal with hermeneutical underdetermination. I will argue that neither is able to reconcile faithfulness to reliability with an irreducible degree of hermeneutical underdetermination. The central tenet of this chapter is that both discussed paradigms of content analysis do not exhaust the set of available options. Instead, I will argue in the last section (3.4) that an irreducible degree of interpretation does not necessarily impede reliability. By drawing on an analogy from the natural sciences, I will sketch a probabilistic conceptualisation of reliability that accounts for hermeneutical underdetermination.

3.1 WHAT IS CONTENT ANALYSIS?

In its early stage, at the turn of the 20th century, content analysis was mainly used for journalistic studies. Quantitative newspaper analysis used subject matter categories to, for instance, survey the literal space that editors gave to specific topics in American newspapers. Researchers compared these quantified results between different newspapers. They investigated subject matter changes over time to substantiate claims about general journalistic trends about, for instance, the growth of yellow journalism (Krippendorff 2004a, 5–6).

The 1930s and 1940s initiated a phase of methodological reflection and interdisciplinary growth. The rise of radio and television led to an interest in understanding how these new forms of mass media influence public belief, preferences and attitudes. Economic stakeholders took an interest in investigating the efficiency of advertising (Merten 1995, 39). With the beginning of fascism, the U.S. used content analysis to identify propagandists and to extract information from propaganda to predict the enemy's actions (Krippendorff 2004a, 8–9).

The increase in relevance incentivised content analysts to reflect on their methods. In 1939 Harold D. Lasswell, who coined the term *content analysis* (Merten 1995, 43), became Chief of the Experimental Division for the Study of Wartime Communications at the U.S. Library of Congress. He and his colleagues contributed to the definitions of important concepts such as *recording unit* and *sampling unit* and established inter-coder reliability as a quality standard to assess the intersubjectivity of research findings (Merten 1995, 41). Berelson and Lazarsfeld (1948) provided the first comprehensive picture of content analysis as a method, which resulted in Berelson (1952), the first content analysis textbook (Merten 1995, 42). He defined content analysis as “a research technique for the objective, systematic, and quantitative description of the manifest content of communication” (Berelson 1952, 18). The first methodological controversies were triggered by Kracauer (1952), who criticised content analysis, particularly Berelson's conceptualisation, for its reliance on quantitative methods and its naive notion of objectivity. At that time, mainstream content analysts largely ignored this criticism and embraced the quantitative paradigm (Krippendorff 2004a, 10–11). But over time these challenges led to a methodological maturation of content analysis, as witnessed by several alternative characterisations,³⁰ and to a variant of content analysis that is less optimistic about the prospects of using quantitative methods. So-called qualitative content analysis was popularised mainly by Mayring (1983) in the German speaking community.

By now, content analysis offers a broad spectrum of techniques that can be used to analyse different types of messages. While content analysis was first used to investigate political and social phenomena that are connected to mass media, it is by now applied in a diverse range of disciplines, such as psychology, anthropology, political science, literature, and linguistics (Krippendorff 2004a, 11–12).

³⁰For an overview, see Titscher et al. (2000), 57–58 and Neuendorf (2002), 10.

3.1.1 DOING CONTENT ANALYSIS

Content analysis is an umbrella term for a set of research methods tailored to analyse texts or meaningful matter in general. For simplicity's sake, let us from now on consider texts as the objects of analysis.

As a research technique, content analysis is goal-orientated: It is used to provide answers to research questions. Before using these methods, the researcher has, therefore, to formulate the research question and possibly some specific hypotheses. The step of defining the research problem is not independent of the decision to even use content analysis. The researcher has, first, to presuppose that the research question can be answered by analysing texts and, second, that content analysis is suitable to provide evidence for their hypotheses.

The heart of any content analysis is the classification of texts, or text segments according to an annotation scheme. The categories of the annotation scheme represent dimensions or variables that are supposed to be relevant for answering the research question. A content analysis contains three general elements, or steps: The *research design* includes the precise formulation of research questions and hypotheses, and the design of the annotation scheme (i.e., the definition of all relevant categories). The *data-making process* includes the selection of the relevant text corpus that is to be analysed, the determination of text segments in the corpus that are going to be categorised (so-called coding units) and the actual categorisation of these coding units. Finally, the gathered data is supposed to provide evidence for the researcher's hypotheses. Accordingly, the final step is to *analyse the data* to draw inferences concerning the research question. These different elements do not need to be traversed in a specific linear order and can be tied to each other by feedback loops. However, later we will identify some essential constraints resulting from demanding a specific type of intersubjectivity.

In contrast to the natural sciences where researchers can draw on already established operationalisations of relevant variables, there are usually no standardised operationalisations for analysing texts. The design of the category system is therefore of crucial importance in every content analysis. A category system is a set of categories with a set of category values for each category. Annotation (also, coding) is the process of assigning these values to coding units in the text corpus, which can comprise whole texts or text segments in texts.³¹ You can think of this process as an annotation of a text. Accordingly, we will call those assigning category values to coding units coders or annotators. I will later in this chapter show that the requirement of reliability demands to separate researchers and coders in a specific research context. In other words, the researcher should not perform the actual annotation. The category system is subject to the following requirement: The values of one category should be mutually exclusive and jointly exhaustive (e.g., Krippendorff 2004a, 155; Neuendorf 2002, 118). Depending on the analytical aim and the nature of the category, its values can represent a nominal, ordinal, interval or even a ratio scale.³²

Often, the category system will contain category values that can be attributed easily—for instance, formal categories (e.g., the author of a text or its publication year) or mere

³¹This is a slight simplification. In many applications, categories are not monadic properties but relations. In other words, instead of assigning a category value to *one* coding unit, the category value will represent a specific relation between two or many coding units.

³²For details, see Krippendorff (2004a, 161–69).

syntactical categories (e.g., the frequency of specific word tokens). If syntactical properties are used to define category values, the categorisation can be left to an algorithm. However, often the research question can only be answered by including semantic and pragmatic categories—that is, categories whose assignment demands an understanding of the text. At this point, human coders are usually used for the annotation. Since the categorisation should satisfy certain scientific requirements such as reproducibility—we will soon examine these desiderata in more detail—the annotation should be rule-based and requires certain cognitive abilities. Human annotators will only agree in their understanding of how to attribute category values to coding units if category values are precisely defined and if annotators are properly instructed to apply the annotation scheme. To that end, the category definitions are, usually, explained and illustrated with paradigmatic examples in an annotation manual, which is used to train annotators.

There are different approaches to designing the category system. As already noted, the categories should be relevant to the research problem. But that leaves room for two conceptually different approaches, which can be combined in practice. A *content-driven approach* devises categories by investigating the relevant text corpus itself. A researcher will often formulate the research question by examining their target phenomena. Tentative hypotheses might emerge as the result of exploratory observations and preliminary impressions. Similarly, categories can be developed by reviewing the target corpus.

In contrast, a *theory-driven approach* develops categories based on theories and established scientific results relevant to the research question. By using already established conceptualisations, the researcher can profit from the given state of scientific knowledge and avoids unnecessary detours and deadlocks.³³

The data-making process involves three steps. First, the researcher has to identify the text corpus that is relevant to answer the research question. Depending on the target phenomenon, this corpus can be huge. If, for instance, the researcher wants to gather evidence for very broad hypotheses about the print media in the previous decade, the relevant text corpus is simply all the print media in the previous decade. Consequently, it is often necessary to confine the analysis to a subset of the relevant text corpus. The researcher has to ensure that this subset allows inferring something about the target phenomenon. Often, this is facilitated by using a statistically representative sample from the relevant text corpus. To do that, the researcher has to define *sampling units* within the relevant text corpus that can be (randomly) selected to form a sample, for instance, newspaper articles (Krippendorff 2004a, 113).

However, the sampling units are not necessarily the *coding units*—that is, the units that annotators categorise. Instead, coding units are often text segments within sampling units. Depending on the granularity of the analysis, coding units can be as small as words or clauses or comprise larger text segments, such as sentences, paragraphs or whole texts. There are, unfortunately, no general rules of how to individuate coding units. Preferably,

³³Often, the theory-driven approach is referred to as a deductive design of categories and the data-driven approach as an inductive design. This terminology might evoke misunderstandings since the theory-driven approach seldomly amounts to simple logical deductions from a theory. It is often hard to pinpoint the theory precise enough to allow for logical deductions. Consequently, a specific theory is often consistent with different category systems or might only be used in a suggestive way to design a category system (see also, Kuckartz 2012, 64–65).

they are determinable by simple rules so that coders who annotate the same sampling unit categorise the same coding units. However, depending on the research question, their individuation might demand an understanding of the text. In these cases, annotators might diverge in their determination of coding units.

Lastly, we must distinguish coding units and *context units* (Krippendorff 2004a, 101). Context units determine the context that annotators can use as further information for categorising coding units. Often, it is necessary to allow annotators to use other information than what they can grasp by considering the coding unit alone. For instance, in the case of anaphoric references, annotators have to consider preceding sentences. A pragmatic and often feasible suggestion is to use the whole text in which a coding unit occurs as its context unit.

The actual annotation will generate the raw data, used for the data analysis. The data analysis includes all inferential steps of answering the research question by using the data. Often, the content analyst will aggregate the raw data in some way before drawing further conclusions. For nominal data, it might be helpful to count occurrences of different values; for more complex data types, averages and other statistical measures might be fruitful for the analysis.

As already noted, these three main elements of a content analysis—research design, data-making process and data analysis—are not necessarily independent. It is, for instance, often indispensable to perform a pre-coding during the design phase: Annotators use a preliminary annotation scheme to categorise a subsample of the text corpus. Thereby, the content analyst can test whether the category system is of any help in answering the research question, whether annotators understand the annotation manual and whether the annotation leads to a reliable data-making process. These pre-codings can then be used to improve the annotation scheme. The final data analysis should, however, only be based on the data generated with the final annotation scheme.³⁴

3.1.2 CATEGORISING AS MEASUREMENT

Speaking of the annotation process as part of a data-making process suggests regarding it as a measurement procedure. Therefore, it is no surprise that the measurement picture of content analysis permeates some prominent definitions of content analysis—most notably Berelson’s definition, quoted above (Krippendorff 2004a, 20). Neuendorf (2002) is another textbook including an explicit endorsement of the measurement picture. She begins by citing Stevens’s influential characterisation of measurement as “the assignment of numerals to objects or events according to rules” (Stevens 1951, 1) and goes on to picture the basic idea of classical test theory (Neuendorf 2002, 111). According to this theory, the result of a measurement m can be conceptualised as a combination of two elements: A true value v_t , which we seek to discover, and disturbing influences, which are the result of random or even systematic errors e . This simple idea is then expressed by the following formula: $m = v_t + e$. The main difficulty in measurement is to minimise the error term e and gather information about it to estimate the true value v_t based on measurements. According to

³⁴This requirement is also shared by proponents of qualitative content analysis (see e.g., Kuckartz 2012, 47; Mayring 2015a, 371, 375).

Neuendorf, we can transfer this picture to content analysis. The true value corresponds to the correct category value as defined by the category system; the error term corresponds to misinterpretations of the annotation manual, inattention and fatigue (Neuendorf 2002, 112).

However, the measurement picture has at least three problems—if we leave aside the categorisation of formal and mere syntactical properties of text messages. First, coding is not necessarily an assignment of numbers to coding units, in contrast to Steven’s definition of measurement. In the case of a nominal scale, we could, of course, label each category value with some number. However, an arbitrary assignment of numbers as mere labels to nominal category values does not represent a proper quantification.³⁵

The second problem of the measurement picture concerns its implicit representationalism—the idea that text messages are independent bearers of content and meaning; the idea that content can be *discovered* by the attentive observer. Krippendorff refers to this idea as the container metaphor of content, which “entails the belief that messages are containers of meaning” (Krippendorff 2004a, 38). He objects that the container metaphor implies a naive picture of content and meaning and misses that texts “do not have single meanings that could be ‘found,’ ‘identified,’ and ‘described’ for what they are” (Krippendorff 2004a, 40). Instead, content is the *result* of understanding and interpreting a text by someone in a specific context. In other words, the container metaphor is oblivious to the interaction between the text and reader; thus, it tends to ignore the relatedness between text and reader.

Finally, the measurement picture presupposes the existence of *unique* true values. Let us concede the following charitable interpretation of the analogy: Talking of true values is not supposed to convey any ambitious ontological commitments. It is, in particular, not meant to imply that true values exist independent of the annotator and the context. Hence, we do not necessarily fall prey to the above mentioned container metaphor of content if we adopt the terminology of ‘true values’. In consequence, we should understand true values as correct values according to an annotation scheme. However, even under this charitable interpretation the measurement picture seems to labour under a uniqueness assumption of value assignments. For instance, Neuendorf (2002, 111–12) explains the variability of reproduced categorisations by coding errors alone. In other words, she assumes that category values have to be defined in such a way that every coding unit can be assigned exactly *one* correct category value.

So it seems we should refrain from thinking of annotation as a measurement. While I believe we should take these problems seriously, I oppose the conclusion of rejecting the measurement picture. Instead, I will embrace the analogy and use it to develop and

³⁵You can, of course, attach numbers to items on a nominal scale. However, what is needed for a proper quantification is a clarification—provided by a specification of measurement procedures—of what the numerical differences between these numbers mean. Otherwise, the ordering of the numbers and their differences have no relevance. It does not express something represented in world of the items. You could use any other set of numbers. That is the reason why in measurement theory the differences between scales are formally couched in terms of invariance-transformations that have no bearing on the underlying informational content. An ordinal scale, for instance, allows any transformation from one set of numbers to another set that does preserve the ordering of the numbers. The underlying informational content is, in this case, exhaustively captured by the ordering. For an introduction to these issues, see Tal (2020); for a rigorous mathematical introduction, see Suppes et al. (2007).

illustrate some of the main ideas of this chapter. In doing so, I do not have to assume that coding is, in fact, a measurement. The fruitfulness of the analogy must be measured by what we learn from it, and what we learn from it, must be justified independently of the analogy. In other words, I do not use the measurement picture to justify the central claims of this chapter but to motivate them. Additionally, I do not propose to transfer ideas and concepts unquestioned from metrology—the “science of measurement and its application” (JCGM 2012, 32)—to content analysis. Rather, we have to critically assess and adapt what we take from there in light of the differences between measurements made with instruments and human annotation.

So why do I think of the measurement picture as a fruitful analogy? One reason is that the raised methodological concerns are discussed not only by content analysts but also by metrologists and philosophers of science in general. By considering the broader metrological context, we might gain relevant methodological insights for content analysis. Additionally, the raised problems, particularly the second one, draw on a simplified or even naive measurement picture that does not represent modern conceptualisations of measurement.

Let me start with the first problem. Content analysts have already discussed whether assigning category values from a nominal scale to objects can be regarded as a measurement. One worry is that if we consider a mere nominal scale—and thereby every higher scale—as a measurement scale, assignments of categories (independent of their type) would generally count as measurements. In other words, whenever someone or some apparatus categorises an object, the categorisation would count as a measurement. This would trivialise the very notion of measurement (Ritsert 1972, 26). This objection does, however, ignore that the distinguishing feature of measurements in comparison to other types of categorisations does not necessarily have to be fleshed out in terms of scale types. Früh (2017, 37–39) argues that nominal-scaled values do not automatically constitute a measurement. But as soon as we count occurrences of observed nominal-scaled values and compare them for different categories or between different texts, we can speak of a measurement. In metrology, the current official view, as witnessed by the *International Vocabulary of Metrology*, is to discard nominal scales as measurement but to regard all higher scales as measurement scales (JCGM 2012, 16). But this view might change.³⁶

In the context of this work, the question is more or less irrelevant. As soon as content analysts can count occurrences of nominal scaled values, they can use statistical methods. In this connection, it is not important whether we regard the assignment of nominal scaled values as a measurement but whether the use of quantitative methods from metrology offers the content analyst any surplus.

The second problem, however, is more pressing since it seems to pinpoint a relevant difference between natural-science measurements and human annotation. In contrast to the container metaphor of content, “texts have meanings relative to particular contexts, discourses, or purposes” (Krippendorff 2004a, 24). They are the result of an interaction between the reader and the text. According to Früh (2017, 108), we can think of this interaction process as a transformation in contrast to a simple transport model. According to the transformation model, content results from an ongoing transformation. The text is

³⁶Mari (2015) considers the stance toward nominal scales as one open issue in metrology.

only the starting point, which is then transformed by the reader into meaning by going through a hermeneutic circle.³⁷ This process depends on the subject, who interprets the text, but is not arbitrary since inter-subjective conventions govern it.

It is crucial to keep in mind that the categorisation of coding units results from an interpretational interaction between the annotator and the text. However, it is a simplification to think—as insinuated by the second problem—that a measurement device simply mirrors some kind of independent reality. Surely, we can hardly think of measurement devices as individuals who interpret their target objects. However, what they share with text interpreting annotators is a process of interaction. According to the model-based account of measurement, measurements are the results of an interaction between a measurement device and the target object. This process must be represented by a theoretical model—often a statistical model—based on non-trivial assumptions about the measurement device (Tal 2020). The model assumptions, and particularly the assumptions about the measurement device, have a bearing on the inferences the researcher can draw about the target object. In this respect, measurement is similar to annotation. In both cases, the result of the interaction hinges on the target object and the measurement device. While there are important differences in the nature of this interaction process, the conceptualisation of this interaction process in metrology might offer some valuable insights to methodologists of content analysis.

Similarly to the first two problems, the third concern—the existence of unique true or correct values—is discussed among metrologists. Not only is there a shift from a naive realist conceptualisation of measurement (Mari 2015), some question even the usefulness of the notion of true values (e.g., Grégis 2015). Instead of uncovering *the true value* by a measurement, metrologists think of true values as values that are consistent with the definition of the quantity to be measured (JCGM 2012, 20). These definitions can allow for so called definitional uncertainties due to the “finite amount of detail in the definition” (JCGM 2012, 25). On this view, the measurement process is thought of as a “process of experimentally obtaining *one or more quantity values* that can reasonably be attributed to a quantity” (JCGM 2012, 16, *emph. added*).³⁸ Consequently, the assumption of unique true values is not an objection to the measurement picture in general but an objection to a very specific and narrow conceptualisation of measurement.

3.2 RELIABILITY AND VALIDITY

Definitions of content analysis emphasise its role as an empirical research method bound to specific scientific requirements. Neuendorf, for instance, characterises content analysis as “a summarising, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing) [...]” (Neuendorf 2002, 10). The distinguishing feature of content analysis in contrast to other forms of text analysis is, therefore, its scientific aspirations. On this view, content analysis has the same primary goal as any scientific investigation, namely “to provide a description or explanation of a phenomenon in a way that avoids the biases of the investigator. Thus, objectivity is

³⁷For the role of hermeneutical methods in content analysis, see Mayring (2015b, 29–32) and Kuckartz (2012, 16–21).

³⁸For a more thorough analysis of these considerations, see Grégis (2015).

desirable” (Neuendorf 2002, 11). Similarly, in another textbook, Riffe et al. (2014) regard “objectivity and reproducibility or replicability” as general scientific traits. They connect these requirements to the idea of avoiding subjective biases: Scientists’ “findings should be objective and not subject to what the researcher believes or hopes the outcome will be” (Riffe et al. 2014, 20).

These borrowed formulations suggest that the relevant scientific requirements draw on various concepts, such as objectivity, intersubjectivity, reliability, reproducibility and validity. But how are they related to each other? Do they correspond to different requirements, or can they be interpreted as reformulating one or at least a limited number of requirements? Is content analysis bound to one specific set of requirements or are there different types of content analysis that differ in their compliance with these requirements? To answer these questions, we, first, have to better understand which ideas these concepts are supposed to convey.

I suggest distinguishing two main ideas that should be conceptually separated. First, there is a notion of scientific objectivity in a strong sense that is closely connected to truth: Scientific findings should uncover facts, which are independent of the individuals who seek to find them. In content analysis, this idea is referred to as validity. Second, there is a notion of objectivity in a weaker sense that is closely connected to intersubjectivity and reproducibility: Scientific findings should be reproducible. If different researchers replicate a research effort, they should come to the same results. The second idea is called reliability in content analysis. Though conceptually different, both notions are connected: Reliability is taken as an indicator of validity. So let us take a closer look at how these requirements are explicated in content analysis.

3.2.1 OBJECTIVITY AS VALIDITY

Scientific objectivity, in a strong sense, is grounded in the idea that science aims at true descriptions of things that are out there in the world (Reiss and Sprenger 2020). Scientists are trying to discover facts of an objective reality—a realm that exists independently of the scientist—by formulating scientific theories that systematise a body of knowledge in a particular domain.

The extent to which scientific theories reach the aim of truth is often explicated with the notion of validity. There are, however, different validity concepts. They all share a connection to the notion of objectivity as faithfulness to facts.³⁹ Here, we will only consider the notion of validity inasmuch as it applies to measurements in content analysis—that is, the validity of coding.

Observations mediate the scientist’s access to theories. Scientists make measurements that produce evidence to choose between different competing theories. Accordingly, to help track true descriptions of the world, the measurement devices have to measure what they are supposed to measure. This requirement is referred to as validity of measurement—or, in the context of categorising coding units, validity of coding. It demands that measurement

³⁹Krippendorff (2004a, 313–38) provides an overview of different validity concepts in general and validity concepts that are relevant in the context of content analysis. His *semantic validity* captures best the validity concept I discuss in the context of the annotation process.

devices provide true testimonials of the reality they are devised to capture.

We already saw that the process of categorising coding units is different from other measurements in a couple of relevant aspects. Annotation involves humans and their understanding of the coding unit's meaning. This "content" is not something to be discovered by impartial observers but the result of an interaction. That does not imply that the very concept of measurement validity is misplaced. Admittedly, to use the expression 'truth' and, in particular, to think of it as being objective in the sense of being independent of the interpreting subject is at least misleading. However, meaning is not arbitrary but bound to certain semantic and pragmatic conventions. Instead of speaking of truth, we should speak of correct categorisations relative to these conventions. The question is whether the idea of correct categorisations is still too strong a requirement.

The challenge in content analysis is that linguistic conventions are often implicit, vague and largely context-dependent. What counts as a correct categorisation might, therefore, be similarly vague and context-dependent—at least, if a coding unit's categorisation depends on its meaning. Instead of judging a categorisation as correct simpliciter, its correctness might come in degrees. Similarly, there might be more than one correct categorisation of a coding unit. So how can the content analyst devise a criterion of correctness? The predominant suggestion is to bound annotators to a standard of coding validity by making the linguistic conventions and the interpretational context explicit. The validity standard is constituted by an annotation manual that includes all necessary coding instructions together with carefully formulated category definitions. The role of explicit and precise category definitions is to specify what counts as a correct categorisation and thereby to narrow down the degree of interpretation.

Ideally, the content analyst can reduce the set of correct categorisations of a coding unit to exactly one (Früh 2017, 113; Krippendorff 2004a, 24). This notion of objectivity as validity and the focal role of precise category definitions were early on advanced by Berelson. According to him, "the requirement of objectivity stipulates that the categories of analysis should be defined so precisely that different analysts can apply them to the same body of content and secure the same results" (Berelson 1952, 16).

The question is whether it is always possible to define category values so precise that there is exactly *one correct category value* for each coding unit. The central subject of this work is to formulate a requirement of reliability that is applicable even if this uniqueness assumption is violated. Before we can approach this challenge, we should, however, better understand the notion of reliability, its relation to similar concepts such as reproducibility and its relation to coding validity.

3.2.2 INTERSUBJECTIVITY AS RELIABILITY

If different annotators categorise a coding unit, we expect them to arrive at the same category value—at least if there is only one correct category value. The reproducibility of coding results by different annotators is a form of intersubjectivity referred to as inter-annotator reliability. In the measurement picture, reliability can be "defined as the extent to which a measuring procedure yields the same results on repeated trials" (Carmines & Zeller 1979 cited by Neuendorf 2002, 28). Considering the specifics of content analysis,

it is perhaps better to follow Krippendorff's interpretivist conception of reliability on which "reliability is the degree to which members of a designated community agree on the readings, interpretations, responses to, or uses of given texts or data" (Krippendorff 2004a, 212).

In this section, I will clarify the notion of inter-annotator reliability as reproducibility by suggesting that inter-annotator reliability is relevant as an indicator of coding validity. By illuminating the role of reliability for the concept of validity, we will be able to better determine the relevant concept of measurement reproducibility in content analysis.⁴⁰ These insights will help to pinpoint important implications for those variants of content analysis that are faithful to reliability (3.2.3) and lay the groundwork for justifying the claim that qualitative and quantitative content analysis do not meet the challenge of hermeneutical underdetermination (3.3).

The considerations in this section proceed in two steps: I begin by describing a general picture according to which the epistemic function of reproductions is their confirmatory power. Accordingly, the extent to which repetitions should be similar and dissimilar depends on what is supposed to be (dis-)confirmed. We will then see that reliability should be understood as a notion of independence: Measurements should be independent of the particular measurement apparatus we take from a specified population of measurement devices. Failing to satisfy this desideratum indicates a lack of validity, and reproductions of the coding effort are supposed to provide (dis-)confirmatory information for this independence claim.

To test a measurement procedure for reproducibility, we have to repeat measurements. However, there is no exact replication of a measurement procedure (Schmidt 2009). At least the point in time will differ when a measurement is replicated. Additional practical constraints will prohibit us from repeating a measurement in every infinitesimal detail.⁴¹ But this inability is not a problem since we do not want exact replications. Krippendorff (2011) explains that "reliability [...] is obtained by comparing measures of the same phenomena obtained under different circumstances or from different devices assumed to measure the same" (Krippendorff 2011, 94).

In the more general context of experimentation, Collins (1985) explains that exact replications of experiments lack the confirmatory information we seek. What we want to achieve by reproducing an experiment is to gather additional and independent evidence for what we already observed; we want to obtain the observed result by different experiments.⁴² To that end, we might vary the experimenter or experimental set-up. An apt analogy might be the role of independent witnesses in a trial. The same witness might be requested to repeat their story to check for internal inconsistencies. However, at some point, the confirmatory power of these repetitions will not further increase. But independent testimonies from others will increase trustworthiness of the initial testimonial if the different reports coincide. The same rationale motivates the reproduction of experimental outcomes. By reproducing an

⁴⁰There are different conceptualisations of measurement reproducibility (and reproducibility in general) without any established systematic terminology (Schmidt 2009). See also Fidler and Wilcox (2018) and Hammersley (1987) for an overview of different notions of reproducibility. The challenge is then to identify the relevant notions of reproducibility for content analysis.

⁴¹This point is already made by Popper (Popper [1959] 2002, 469–70).

⁴²This point is made by Drummond (2009) in the context of computer science.

experiment under different circumstances, the researcher can gather independent evidence of their research results.

Additional confirmational support for an experimental result is thus not provided by repeating an experiment in the exact same way but only by introducing differences to the initial experiment. However, as Collins (1985) explains, successful repetitions should not be too different. At some point, a successful repetition of an experimental outcome becomes spurious and will even cast doubt on the initial outcome. To reuse the analogy: Consider a plaintiff who presents two witnesses of a crime scene who confirm the same story. If, however, it turns out that the second witness was actually not at the crime scene and could thus not have observed what they claim to have witnessed, the deceptive testimonial does not only cast doubt on the second witness but on the testimonial of both witnesses.

These general considerations about the role of experimental repetitions apply similarly to the case of measurement procedures. Suppose a scientist intends to test a measurement procedure for reproducibility. In that case, they have to repeat it under conditions that are to some extent different and, at the same time, to some extent similar. The question is then what we have to vary and what to leave unchanged to maximise the confirmatory power of these repetitions. However, the so far described picture is too abstract to provide a general answer. Rather, the answer will depend on the specifics of the context and the specific intended epistemic function of reproducibility.⁴³ Fortunately, the preceding considerations about the confirmatory power of replications point to the relevant question we have to ask to determine the appropriate level of variation and stability in the context of coding reliability: What exactly is it for what we search confirmation by reproducing a coding effort?

Krippendorff reminds us that the literal meaning of reliability is the ability to rely on something. He elaborates that “in content analysis [...] reliability is the ability to rely on data that are generated to analyse phenomena a researcher intends to study, theorise, or use in pursuit of practical decisions” (Krippendorff 2016, 139). And to rely on data, the researcher has to be sure that the “data have been generated with all conceivable precautions in place against known pollutants, distortions, and biases, intentional or accidental [...]” (Krippendorff 2004a, 211). This understanding of reliability is connected to a notion of objectivity in terms of excluding extraneous and distorting circumstances in the measuring process. In the case of content analysis, the measuring process is the annotation process, and the generated data is the coding data. Krippendorff mentions several examples of such distortions, such as the lack of understanding or misunderstanding of category definitions, the annotation manual or coding instructions in general and idiosyncratic habits or even personal interests of annotators in the outcome of the research (Krippendorff 2011, 94).

Krippendorff’s considerations already pinpoint what is supposed to be confirmed by reproducing a coding effort. To say that a categorisation is free of any extraneous and disturbing influences and to say that it is error-free, is nothing more than to say that it is a correct categorisation and, hence, a valid categorisation. On this view, we should conceptualise coding reliability as an indicator of validity. In other words, the confirmatory

⁴³This context dependency is probably but one explanation for the variety of reliability concepts. See Fidler and Wilcox (2018) for an overview of classifying various reproducibility concepts according to their epistemic function.

information for which we aim in reproducing a coding effort is information that is relevant for judging the validity of the coding process. By varying certain parameters between repetitions and keeping others constant, multiple categorisations of the same material should provide confirmatory information regarding the validity of the coding.

We should, however, still be able to tell both concepts apart. In other words, it should be possible that a measurement procedure is reliable without being valid. To that end, a reliable measurement procedure is not *necessarily* a procedure that is free of any extraneous and disturbing influences. Rather, reliability can be taken as provisional evidence for validity.⁴⁴

Let us now better understand the confirmatory relation between validity and reliability. As often noted and explained above: Reliability is necessary for validity but not the other way around (Krippendorff 2004a, 212–13; Neuendorf 2002, 141). Under the uniqueness assumption—that is, the assumption that there is exactly one correct categorisation for each coding unit—one of two disagreeing categorisations must be incorrect. In other words, if two annotators disagree on the same coding unit, one of them must be wrong. Therefore, the categorisation should satisfy the requirement of (strict) inter-annotator reliability—that is, we should demand that different annotators should categorise the same coding units in the same way.

The same reasoning applies, of course, to repeated codings by one annotator. Consequently, we also demand that repeated codings of a coding unit by one annotator should not differ. This is usually called intra-annotator reliability.

If these requirements are violated, the content analyst has evidence that the coding results are incorrect. In other words, diverging categorisations of the same coding unit allow us to infer a violation of validity. In this sense, reliability is necessary for validity.

Although validity implies reproducibility of categorisation under the uniqueness assumption, this form of intersubjectivity is weaker than validity because annotators may agree on an incorrect categorisation. In other words, a successful replication of a categorisation by two different annotators does not imply that both annotators arrived at a correct categorisation.

However, that does not mean that coinciding categorisations provide nil confirmatory power for validity. Instead, we can regard reproducibility as an indicator of validity. The basic idea is to reinterpret the disconfirming power of failed reproduction attempts by using the broader epistemological paradigm of Popper's falsificationism (Popper [1959] 2002). Falsificationism is motivated by the observation that scientific theories include universal statements, which imply statements over a potentially infinite number of phenomena. The sentence "*All ravens are black.*" is an illustrating example of a universal statement. The statement can be interpreted as saying that every object that is a raven is black. Inasmuch as there are infinitely many objects in the universe, the universal sentence has infinitely many implications. The problem—prominently known as Hume's problem of induction—is that universal empirical statements cannot be directly proven to be true since scientists

⁴⁴Such a conflation of reliability and validity is particularly prevalent in Schreier's definition of coding reliability, who also embraces a measurement picture. She writes that "an instrument *is called reliable* to the extent that it yields data that is free of error" (Schreier 2012, 167, *emph. added*). However, since a lack of errors implies the validity of measurements, her reliability concept coincides with validity.

cannot check every element in an infinitely large set. Additionally, every confirming instance seems to contribute only an infinitesimal small amount of support for the universal statement.

The basic idea of falsificationism is that scientists should not search for confirming instances but for falsifying instances. It only needs one white raven to disprove the claim that all ravens are black. According to this view, science advances by trying to falsify hypotheses instead of proving them. Once a hypothesis is falsified, an alternative should be formulated that withstands all previous attempts to disprove the former hypothesis—that is, an alternative consistent with all observations made so far. Then, the new hypothesis is the target of further falsification attempts. Theories that survive all attempts to disprove them can be rationally accepted. Hence, the fact that a theory withstands serious attempts to disprove it is in itself a good reason to accept the theory. In other words, if comprehensive efforts fail to disprove a theory, they can be considered as indirect confirmation. This acceptance is provisional since future observations might again falsify the currently accepted theory.⁴⁵

We can now reinterpret the role of reproducing coding results from the perspective of falsificationism. By reproducing categorisations of the same coding units—either by different annotators or by the same annotators at different times—the content analyst gathers reliability data, which they use to infer something about the coding procedure: Reliability data can be used to falsify validity. Since coding reliability—understood as coding reproducibility—is necessary for validity, failed attempts to reproduce the coding effort are falsifications of validity. We can, therefore, consider the reproduction of a coding effort as an attempt to disprove the validity of a coding procedure. In this sense, successful reproductions of coding results constitute indirect evidence for validity.⁴⁶

The fact that reliability is necessary for validity gave us reason to interpret the role of replications as attempts to disprove validity. We can now use the falsificationist perspective on reproducibility together with the measurement picture of content analysis to answer the

⁴⁵The described picture is a simple, or even naive form of falsificationism and faced by many challenges. The locus classicus for this prominent debate comprises Kuhn (1962), Lakatos (1968) and Lakatos (1970).

⁴⁶The described falsificationist picture is, additionally, in accordance with the role of failed reproduction attempts in content analysis. According to falsificationism, the scientist should search for a new theory consistent with all previous observations if they have discovered falsifying evidence. In the same way, failing to reproduce the categorisation of coding units is not necessarily the end of a research effort. Instead, failed attempts of reproduction should be used to alter the measurement procedure—that is, to adapt the measurement apparatus so that it withstands further attempts to disprove its validity. The falsifying evidence can be used to identify the sources of the observed discrepancies. Suppose, for instance, the actual categorisation takes annotators a couple of hours. According to the initially devised coding scheme, annotators are asked to categorise the whole material without taking a break. If the content analyst observes that coding discrepancies do not happen at the beginning but occur more and more frequently with passing time, annotator fatigue is a likely explanation for the lack of reliability. In this case, the content analyst should simply try to include breaks, repeat the coding effort under the new coding scheme and check whether reliability increases. In other words, reliability data has two essential roles in content analysis. Suppose the produced reliability data confirms a lack of validity. In this case, it can help to redesign the measurement apparatus by, for instance, changing the coding regime or altering the coding manual and the coding instructions. Hence, failed reproductions provide important insights during the design phase of a research effort. Reliability data that confirms the reproducibility of a measurement procedure is, on the other hand, confirmation of reliability and an indirect indicator of validity. At that point, the actual data-making process can start.

initial question about the specification of what to vary and what to keep constant for the generation of reliability data.

The main tenet of falsificationism in science is to search for falsifying instances of theories and, in particular, their general statements. So we should ask ourselves what exactly do we want to falsify by reproducing a coding effort. In particular, which general statement do we aim to falsify? A preliminary answer is that by comparing multiple categorisations of the same coding unit, we are able to falsify the validity of the coding. If two annotators categorise the same coding unit differently, one of them must be wrong. But we do not expect everyone we put up to this task to categorise correctly. To pick up Collin's point from above, we might even get suspicious about our measurement design if arbitrary people—including people we did not instruct properly—would agree in their categorisation.

One focal idea of content analysis as a scientific method is to establish coding rules in the form of category definitions and an annotation manual. These rules set up the validity standard for the coding process—that is, they determine what qualifies as a correct categorisation—and are used during the annotator training to instruct people on how to annotate according to those rules. So while we do not expect arbitrary people to arrive at correct categorisations, we expect properly trained people to do so. Reproducibility tests are, therefore, attempts to frustrate this more specific expectation. We want to check whether individuals from this particular population—the population of trained annotators—are able to categorise correctly. What we want to falsify—to express it in the measurement picture—is the validity of a measurement procedure. Category definitions, annotation manual, coder training and other specifications determine the measurement procedure—that is, the population of measurement instruments. The specification of the measurement procedure includes, additionally, all other relevant specifics about the coding scheme and all requirements that people have to fulfil to qualify for a coder training—for instance, the necessary educational background and needed language skills.⁴⁷

The expectation is now that an arbitrary chosen “measurement instrument”—that is, a trained annotator satisfying all constraints—will categorise coding units correctly. In other words, the general statement that we try to falsify by reproducing a coding effort is that all measurement devices in the specified population will produce correct measurements. And that answers our initial question: During a reproduction of a coding effort we should not change those parameters that constitute the measurement procedure. For instance, we should only use properly trained people. But we can, and even should vary all parameters that are not explicitly or implicitly determined. Inasmuch as a property varies within the population of measurement devices, we should strive to vary this parameter during the reproduction of measurements. For instance, since our research results should not depend on one specific individual, the specification of the coding procedure does usually not prescribe to use one specific individual, say Richard Powers. Accordingly, we should vary the annotators between reproductions of the coding. Hence, we should check for inter-annotator reliability.

We can now see that reliability is, in essence, an independence claim: Codings should not depend on the particular measurement device we pick from the population of measurement instruments. To put it into a neat formula: The requirement of reproducibility and a specifi-

⁴⁷See Krippendorff (2004a, 127–31).

cation of what to vary and what to keep fixed during reproductions equals independence of the particular measurement device we choose from the population of measurement devices.

This insight is unsurprisingly in line with prominent clarifications of reliability. Krippendorff, for instance, picks up the measurement theoretic understanding of reliability, which demands that measurement results are not to be “*affected by variations in the extraneous circumstances of the measuring process*” (Krippendorff 2011, 94). There are two ways to ensure that possibly distorting influences do not affect the measurement: Either we have to ensure that a disturbing influence does not occur during a measurement (by including a corresponding requirement in the specification of the measurement procedure), or we know and can show through replications of the measurement process that those parameters do not affect the measurement result. If, for instance, we know that fatigue disturbs the measuring process, we have to shield the coding process from this influence by making sure that annotators are well recovered before coding; if, for instance, we hypothesise that interests in the outcome do not disturb the measuring process, we should check that this, at least possibly disturbing influence is, in fact, not disturbing the coding process by reproducing the categorisation with annotators that differ with respect to their interest in the outcome.

This notion of reliability as reproducibility, which turns out to be an independence requirement, provides a conceptualisation of reliability that is distinct from the notion of validity: It is still conceptually possible that all annotators from a specified population of annotators agree in their categorisation of a coding unit while at the same time being collectively wrong in their categorisation.

Additionally, identifying a coding procedure with a population of measurement instruments hints at possibilities to move beyond the simple falsificationist paradigm—at least what regards the inference from reliability data to the reliability of the coding process. If the produced reliability data is based on a representative sample of the whole population, we can ground the justification of reliability on statistical inferences.

3.2.3 RELIABILITY-ORIENTATED CONTENT ANALYSIS

In the following, I will refer to content analysis that considers reliability, as it was explicated in the last section, an essential requirement as *reliability-orientated content analysis*. The elaborated clarification of reliability as a notion of independence implies some important desiderata.

First, content analysts should not themselves annotate the text corpus of interest but instruct others to do so (Krippendorff 2004a, 131). These annotators should be people who were not involved in the design of the category system. Otherwise, there is the danger that these researcher-annotators utilise implicit knowledge that is not part of the explicit description of the coding procedure. In other words, people who participate in the design phase of the research effort are not like any other randomly drawn “measurement instrument” from the population of measurement instruments but exhibit some very specific features. In consequence, observed coding agreements from this class of annotators do not constitute informative evidence for the reliability of the coding process in general.⁴⁸

⁴⁸It is, of course, possible to define the population of measuring instruments in such a way that it is

Second, the coding procedure should be replicable in a practical way. All researchers who want to replicate the data-making process should be given the means. In other words, the measurement instrument should be available to other researchers. I will refer to this type of replicability as repeatability.

Repeatability demands that the measurement procedure with its accompanying methods and all the necessary information for its practical implementation are described transparently and that these descriptions are accessible to other researchers. In other words, the category definitions, the annotation manual, specifics about the annotator training and the criteria for selecting qualified annotator candidates should be made available to the research community.⁴⁹

The detection of possible fraud is hopefully not the primary rationale for this requirement. Rather, repeatability is linked to the demand to test reliability as rigorously as possible. Reliability tests performed by another research team have a much higher chance of detecting a lack of reproducibility. Suppose, for instance, that the research team that performed the content analysis in the first place did successfully check the coding procedure for reproducibility but satisfied a specific requirement implicitly—something they did not include in the specification of the measurement procedure. Another team, unaware of this relevant information, who repeats the coding solely based on the explicit description, will likely not satisfy this requirement and thus not be able to reproduce the annotation successfully. In this example, conformity to repeatability and the actual replication of the annotation uncovers a spurious form of reproducibility—or to give it a positive twist, an incomplete description of an otherwise reliable annotation procedure.

Third, there should be a procedural boundary between the design phase of the research effort and the data-making process. The measurement instrument must not be changed during the data-making process. In other words, category definitions or any other properties that determine the coding procedure must not be altered during the data generation. The reason is quite simple: First, the generated measurements should be based on one particular measurement procedure. Second, this measurement procedure should be checked for reliability.

This requirement does neither imply that reliability tests have no role in the design phase of the measurement instrument nor that there is a predefined border where the design phase ends and the data-making process begins—quite the contrary. As explained above, preliminary attempts to design a coding process should be checked for their reliability, which demands generating reliability data. If this preliminary measurement procedure fails to be reliable, the test results can be used to identify distorting influences and help to increase reliability. However, at some point, the researcher has to decide on a final version of the measurement procedure, determine its reliability and then generate all needed data without further adapting the measurement process. In other words, every alteration of the coding procedure demands a reliability test of the resulting coding procedure and the whole material has to be (re)annotated with the new measurement device.

required to run through the design phase of the research effort. However, most likely, such a measurement procedure will not be reliable since the design phase won't be reproducible.

⁴⁹See, e.g., Krippendorff (2004a, 217), Krippendorff (2016, 143), Riffe et al. (2014, 20) and Artstein and Poesio (2008).

Last, annotators should work independently from each other—that is, not collaborate during the coding. Instead, annotators should work on their own and based on their given instructions and the annotation manual (Krippendorff 2004a, 131). Why is that important? If reproducibility is essential, why don't we let different annotators not work together and let them collectively decide how to categorise a coding unit? So-called consensus coding would lead, obviously, to an agreement in categorisation. However, this kind of agreement has nothing to do with the reliability of a coding procedure. In particular, a collaboration between annotators is not a means to gain reliability. The point is that reliability is not about any form of agreement but about reproducibility or, as explicated above, independence. Let us, therefore, consider about how consensus coding fits into the picture measurement reproducibility.

There are three possibilities for conceptualising consensus coding as part of the coding procedure. Either the specification of the coding procedure forbids consensus coding, requires consensus coding, or is silent on the issue.

In the first case, the agreement of annotators who violate the requirement of independent coding has nothing to do with the reliability of the coding procedure simply because, in this case, collaborating annotators do not constitute a proper measurement procedure as defined by the specification of the measurement process.

If, on the other hand, annotators are required by the specification to decide on categorisations collectively, then one of those annotators alone does not fully represent the measurement device but the whole set of annotators who work together. Accordingly, the collaborative decision on a categorisation is only one single measurement and not the replication of many measurements (Krippendorff 2004a, 217). To reproduce a measurement with another measurement instrument, another set of collaborating annotators would have to repeat the categorisation—including the collective decision making—independently from the former set of annotators.

Since group discussions are likely affected by internal group dynamics, such a repetition would most probably lead to a different coding result. In consequence, if consensus coding is intended as part of the specified coding procedure, the reliability of such an instrument is expected to be low. This is, of course, an empirical hypothesis, which might not hold generally. The main point is that consensus coding needs to be scrutinised by special tests of inter-annotator reliability, which demand to reproduce the collaborative annotation by different groups of annotators. In particular, the agreement that is part of the collective decision has nothing to do with the reliability of such an instrument.

If, finally, the specification of the coding procedure is silent on the issue, both collaborating annotators and independently working annotators are instances of the measurement procedure. Again, there is the suspicion that it matters whether annotators collaborate or not. Accordingly, the content analyst should vary this parameter during the replication to check reliability. Similar to the second case, it is most likely that such a coding procedure is not reliable.

Consensus coding in itself has thus nothing to do with reliability. It even stands to reason that consensus coding is not reliable. In the context of reliability-orientated content analysis, the analyst has therefore two options: Either they use consensus coding and have

consequently check the reliability of this group effort by repeating the consensus coding with a different group of annotators, or they dispense with consensus coding and use the independently generated coding results as reliability data.

3.2.4 THE GRADED PICTURE OF RELIABILITY

So far, we have thought about reliability as an all-or-nothing concept. Either a categorisation is reproducible, or it is not. And if it is not reproducible, we can infer that the measurement procedure is invalid. This type of reproducibility is a very demanding requirement. On this picture, it needs only two disagreeing annotators from the whole population of measurement devices to reject the coding procedure as invalid. The all-or-nothing conceptualisation of reliability evaluates coding procedures against an idealised standard in which humans do not make any mistakes. Consequently, it would be hard to develop reliable coding procedures—besides those that can be performed by algorithms.

From a pragmatic perspective, it is, therefore, advisable to search for a suitable weakening of this requirement without giving up the very idea of reliability (Krippendorff 2011, 95). Fortunately, there is a straightforward way to do so. Paradigmatically, annotators are asked to categorise many coding units. Consequently, we can say more about the success of replications than simply stating whether there are disagreements or not. Instead, based on many coding units, we can determine the number of agreements and disagreements among them. On this idea, reliability comes in degrees. The natural suggestion is then to introduce a graded notion of reliability, that is, a reliability measure.

The formulation of reliability measures poses two challenges. First, we have to mathematically conceptualise such a measure—that is, we have to suggest a method of calculating reliability values given our reliability data. Second, we have to consider how this quantity relates to validity and how we judge a coding procedure to be *sufficiently* reliable.

Following the initial consideration, we could take percentage agreement as a measure of reliability. If the reliability data is based on two annotators who categorise n coding units and if they agree in their categorisation on n_A of those units, the observed percentage agreement A_o would be calculated by $A_o = \frac{n_A}{n}$.

This simple measure is, however, problematic for the following reason:⁵⁰ A certain extent of agreement is to be expected even if annotators categorise by using a chance mechanism. If, for instance, there are only two categories, and two annotators categorise by flipping a coin, the probability for a perfect agreement on any finite number of coding units is 0.5. Pure chance agreement “is commonly equated with the complete absence of reliability” (Krippendorff 2011, 97). In other words, an observed percentage agreement cannot be taken as evidence for reliability without taking into account the probability of agreement by chance alone. What is more, this probability depends obviously on the number of the categories. The more categories an annotation scheme has, the less likely is agreement by chance. Scott (1955) frames the issue therefore as a bias problem by saying that percentage agreement “is biased in favor of dimensions with a small number of categories” (322).

⁵⁰For additional problems, see Artstein and Poesio (2008), Krippendorff (2011) and Lombard, Snyder-Duch, and Bracken (2002).

Consequently, content analysts advise against using bare percentage agreement as reliability measure and devise other measures that take chancy agreement into account. The basic idea is to provide a quantity expressing how well the agreement is above chance. There are several different reliability measures, which differ in their scope of application and their assumptions regarding the agreement by chance. The basic idea of these measures can be expressed in the following simple general form:⁵¹

$$rel(A_o, A_e) := \frac{A_o - A_e}{1 - A_e}$$

When observed percentage agreement A_o is indistinguishable from what we expect by chance alone (A_e), the reliability value equals zero—thereby indicating minimum reliability.⁵² If the observed agreement is perfect ($A_o = 1$), and thusly maximally reliable, the reliability becomes one. Thus, reliability measures that are conceptualised by the above formula measure reliability by using two reference points—one of pure chance agreement and the other of perfect agreement.⁵³ Everything between these two reference points expresses the extent to which the agreement is above chance.

Suppose the content analyst has chosen one specific reliability measure that suits their coding design. The question is then how reliable the coding procedure has to be to count as sufficiently reliable. Naturally, we should demand that the reliability value is above zero. Otherwise, the observed agreement can be reached by chance alone. However, how well above chance should annotators perform? In other words, what is the necessary threshold the reliability value has to exceed?⁵⁴

The question of specifying numerically precise thresholds and justifying them as adequate standards of reliability is, however, not fully answered.⁵⁵ There are several suggestions for acceptable levels of interannotator agreement—either with respect to a specific coefficient or as a general recommendation for many coefficients. For instance, Krippendorff (2004a, 241) suggests for his coefficient α that data should only be trusted if α exceeds a value of $\alpha = 0.8$ and that reliabilities between $\alpha = 0.667$ and $\alpha = 0.8$ should only be used for drawing tentative conclusions. The question is which threshold values are suited for what purposes. As emphasised by Krippendorff (2004a, 242), acceptable levels of agreement

⁵¹This formula expresses reliability in terms of agreement. Krippendorff (2004a, 222) prefers an expression in terms of disagreement.

⁵²The calculation of A_e rests on an underlying statistical model—that is, a hypothesised mechanism that produces categorisations and agreements by chance alone. The statistical model should not be interpreted as annotators categorising by chance (Krippendorff 2016, 140). Instead, it is a chance mechanism we use as a reference, which annotators should outperform. Different reliability measures use different probabilistic assumptions. The most simple suggestion is to assume that the probability of putting an item into one category is uniformly distributed over all categories. Many reliability measures proceed, however, differently. They use the observed frequencies of how annotators categorise to estimate parameters of the probability distribution. For an overview of different reliability measures and their probabilistic assumptions, see Artstein and Poesio (2008).

⁵³Cf. Krippendorff et al. (2016, 2348).

⁵⁴Here, the underlying idea is to calculate one reliability value based on the given reliability data and judge the reliability of the coding procedure based on this single point estimate. A more elaborate way is to use a confidence interval around that estimate. Threshold criteria can then be formulated with regard to this confidence interval. See, for instance, Krippendorff (2004a, 242).

⁵⁵As noted, for instance, by Artstein and Poesio (2008, 576) and Neuendorf (2002, 143).

should depend on what is at stake. If the costs of drawing incorrect conclusions are severe, reliability thresholds should be higher than in cases of marginal damage.

3.3 HERMENEUTICAL UNDERTERMINATION

In this section, I will characterise two different types of content analysis that are often distinguished in the literature, namely quantitative and qualitative content analysis. The main aim is to explicate how both variants meet the challenge of hermeneutical underdetermination. This comprises situations in which coding units have more than one correct category value, or in other words, situations where different annotators might end up with different categorisations even though none of them has made any error.

My considerations will proceed along the following lines: First, I will motivate that the difference between quantitative and qualitative content analysis has at best contingently something to do with diverging preferences to use or not to use quantitative methods. Second, and more importantly, I will argue that qualitative content analysis does not prioritise the requirement of reliability in contrast to quantitative content analysis. According to another often-mentioned difference, quantitative content analysis is confined to manifest content whereas qualitative content analysis includes latent content as well. The discussion of this distinction, will, fourth, provide a better understanding of the category system's role to narrow down the degree of interpretation in the categorisation process. Based on these insights, I will, finally, elaborate on how both forms deal with hermeneutical underdetermination and argue for the main claim of this section: Neither quantitative, nor qualitative content analysis provides methods that are simultaneously faithful to the requirement of reliability while at the same time being able to deal with cases of an irreducible degree of hermeneutical underdetermination.

3.3.1 QUANTITATIVE VS. QUALITATIVE CONTENT ANALYSIS

From the beginning of content analysis, starting with Berelson's exposition, several different objections were raised against content analysis. This methodological controversy resulted in two types of content analysis.

The *received view of reliability-orientated content analysis* (or, in short, *received view*) is based on the above described graded notion of reliability. It is often referred to as *quantitative content analysis*. While this form of content analysis avoids far too naive notions of objectivity and does not fall prey to the container metaphor of content, it is still faithful to the desideratum of scientific objectivity in the form of reliability and validity as described above. The second form of content analysis is called *qualitative content analysis* by its proponents and has, as I will show, a somewhat looser stance towards the role of reliability. We will see that this difference can be explained by how proponents of quantitative and qualitative content analysis deal with hermeneutical underdetermination—the problem that different annotators might come to diverging categorisations of coding units due to interpretational differences. Based on this understanding, I will motivate a hitherto not discussed form of content analysis, which is still bound to the desideratum of reliability but differs from the received view in essential points.

Before describing the differences between both forms in more detail, let me start with some cautionary remarks.

The methodological dispute that led to these different forms involved several recurring misunderstandings, some of which resulted from non-charitable interpretations of the various contributions to the debate. Consequently, the dispute was at times the opposite of a fruitful and constructive exchange of ideas and distracted from the critical underlying methodological issues. Fortunately, modern content analysts are less quarrelsome and more charitable towards diverging methodologies. Some go as far as questioning the relevance of the difference between both forms of content analysis—especially if it is framed as a difference between quantitative and qualitative methods.⁵⁶ While I agree that it is futile to cast methodological differences between both forms in terms of a qualitative-quantitative dichotomy since it might invite unnecessary misunderstandings,⁵⁷ I claim that there are important differences between both types of content analysis.

A related problem is a certain degree of fuzziness in the different expositions of the distinguishing features of qualitative content analysis. I will base the following considerations mainly on one specific account of qualitative content analysis—namely, on Kuckartz (2012)—without intending to show whether his exposition is representative or paradigmatic.

Finally, I refrain from a recapitulation of the methodological debate but merely flesh out the main points that are important for this work since the following considerations are driven by methodological concerns and not by exegetical accuracy.⁵⁸

It is instructive to begin by trying to understand what is not a distinguishing feature between quantitative and qualitative content analysis.

Proponents of qualitative content analysis suggest explaining the differences by contrasting quantitative and qualitative methods.⁵⁹ So what is so quantitative about quantitative content

⁵⁶For instance, Krippendorff concludes that the “quantitative/qualitative distinction is a mistaken dichotomy” (Krippendorff 2004a, 87) and “question[s] the validity and usefulness of the distinction between quantitative and qualitative content analysis” (Krippendorff 2004a, 16) since “qualitative approaches to text interpretation should not be considered incompatible with [quantitative] content analysis” (Krippendorff 2004a, 89). Similarly, Früh (2017, 40) opposes the quantitative-qualitative dichotomy since it presupposes that quantitative steps cannot complement qualitative steps. And Mayring, the founder of the modern form of qualitative content analysis, suggests combining qualitative and quantitative steps of analysis (Mayring 2001; 2015b, 53).

⁵⁷That is why I prefer to use another label for quantitative content analysis. I use the term *received view of reliability-orientated content analysis* since it is the currently dominant form of content analysis laid out in several influential textbooks (most notably, Krippendorff (2004a), Neuendorf (2002) and Früh (2017)).

⁵⁸The interested reader can consult Hansen et al. (1998, 94–98), Merten (1995, 50–57), Lisch and Kriz (1978, 48–49) and Ritsert (1972, 19–31) for an overview of this debate.

⁵⁹Several other properties are suggested to distinguish both forms of content analysis (e.g., Kuckartz 2012, 46–47). However, some of these properties draw on very narrow conceptions of content analysis. They can therefore not be offered as distinguishing features between qualitative and quantitative content analysis. Kuckartz (2012, 27), for instance, compares qualitative content analysis with what he calls classical content analysis, which is confined to counting word tokens. While restrictions to syntactical properties are not excluded in quantitative content analysis, annotators will often need to consider semantic and pragmatic properties of coding units. Hence, the modern form of quantitative content analysis is not restricted to count words. Another suggestion is to flesh out the difference in terms of how the content analyst designs category systems and formulates their hypotheses. Kuckartz (2012, 46–47) explains that in qualitative content analysis,

analysis? Content analysts formulate research questions, design category systems, interpret results and argue how their observations are relevant in answering their research questions. All these steps are obviously qualitative. The target phenomenon itself—that is, text or other meaningful matter—is qualitative (Krippendorff 2004a, 87). Similarly, the resulting data of the codings are paradigmatically qualitative and not quantitative since the introduced category values are often based on nominal scales. However, the data aggregation and analysis usually include some form of counting and a consecutive statistical analysis, which are quantitative steps. Content analysts count how often a specific category value occurs among the categorised coding units and compare these frequencies among different texts (Hansen et al. 1998, 95). So, should we understand the difference between both forms of content analysis by their different stances towards the role of frequencies and the use of statistical methods? Do qualitative content analysts not count how often annotators used the same category value during the coding?

Kuckartz reminds us to scrutinise the relevance of frequencies. Simply because something can be counted, gives us no reason to regard the resulting frequencies as relevant for answering a research question (Kuckartz 2012, 55). However, it is a mistake to assume that the quantitative content analyst buys into such a naive picture of praising the countable. In contrast, they agree that the evidential relevance of frequencies for answering a research question hinges on the specific research context (Krippendorff 2004a, 58–62). On the other hand, would the qualitative content analyst refrain from using observed multiplicities of codings if they are informative? Surely not (Lisch and Kriz 1978, 49). But if the usefulness of frequencies depends on the research context and, in particular, on the research question at hand, we might explain the difference in terms of partially diverging research interests: While the quantitative content analyst is (only) interested in research questions that can be answered with the help of counting observed category values, the qualitative content analyst is (also) interested in research questions for which such frequencies are irrelevant.

This interpretation is in line with textbooks of qualitative content analysis. Both, Kuckartz and Mayring emphasise that the use of statistical methods can be helpful in qualitative content analysis but should not be regarded as mandatory and may, depending on the research context, play only a minor or even no role (Kuckartz 2012, 47; Mayring 2015b, 53). But then, I did not find any hint that the quantitative content analyst *demand*s to use statistical methods. In other words, both forms of content analysis can be performed with or without using statistical methods. What remains is perhaps a less sharp distinction: We might say that the quantitative content analyst is *more often* interested in research questions for which frequencies are relevant, among other things, and that the qualitative content analyst is *more often* interested in research questions for which frequencies play a minor or even no role. The demarcating power of this fuzzy characterisation is at least unsatisfactory. Perhaps it is better to search for an alternative or, at least, additional distinguishing features.⁶⁰

the researcher will often analyse the target material to accomplish both tasks. Similarly as before, modern quantitative content analysis can proceed in the same way. It is, in particular, not forbidden to follow the above mentioned content-driven approach (see page 42) to design categories (Krippendorff 2004a). If the quantitative content analyst satisfies the above-mentioned desiderata, they can (and often should) use the target material in a way as described by Kuckartz.

⁶⁰Another suggestion to flesh out the difference proceeds in terms of the investigation's target. According to this story, the quantitative content analyst is interested in general hypotheses. The qualitative content

3.3.2 RELIABILITY AND QUALITATIVE CONTENT ANALYSIS

The received view of content analysis regards reliability, particularly inter-annotator reliability, as a necessary criterion. Content analysts are required to calculate reliability values of their coded material, which have to exceed the recommended thresholds. In contrast, qualitative content analysis advances a less strict view on reliability. While it is acknowledged as a valuable property, it is not regarded as necessary (Mayring 2015b, 54). This looser stance towards reliability is evidenced by the absence of some desiderata that are implied by the requirement of reliability, which I mentioned above.

First, the qualitative content analyst does not demand a strict separation between annotators and researchers. Most often, researchers categorise coding units instead of instructing other persons to do the coding (Kuckartz 2012, 211). As argued above, letting researchers perform the coding of the target material is a threat to reliability.

Even more noteworthy, qualitative content analysts advise using consensus coding as a preferred method to reach agreement among annotators (Mayring 2014, 114; Schreier 2012, 174; Kuckartz 2012, 105). As argued above, consensus coding is a threat to reliability and demands additional and very specific reliability assessments—none of which are advanced by these authors.⁶¹ Hence, inasmuch as qualitative content analysts rely on consensus coding, they most probably violate the reliability requirement.

To the extent that reliability is not required in qualitative content analysis, researchers cannot be sure whether annotator's subjective decisions influence the generated data. Alternative requirements are advanced to compensate for this drawback. As already noted, qualitative content analysts advance some form of consensus coding; other recommend—either as an addition or an alternative to consensus coding—the requirement of (hermeneutical) transparency (Lisch and Kriz 1978, 46), which requires that subjective decisions and interpretations should be made transparent.

3.3.3 MANIFEST AND LATENT CONTENT

We saw that the difference between qualitative and quantitative methods results in a, at best, partial and fuzzy demarcation between both forms of content analysis. The divergence in the appraisal of reliability, on the other hand, provides a significant and much more selective characterisation of the differences. The question is why the qualitative content

analyst, on the other hand, confines their considerations to single cases without trying to generalise their findings to a broader population. This suggestion is in line with the described diverging preferences to use statistical methods since statistics can be regarded as the central mathematical toolbox for generalising from observed properties of samples to properties of larger populations. Even if this description is adequate, it inherits, however, the fuzziness trouble. Additionally, I doubt that the qualitative content analyst agrees with this description. At least Mayring (2015b, 20) claims that generalisation is possible even without the use of statistical methods.

⁶¹Kuckartz (2012, 211) labours under the assumption that using reliability coefficients is motivated by the aim to reach agreement between annotators. Consensus coding is, admittedly, a means to reach agreement. However, this view misunderstands the role of agreement for reliability. On the received view, observed agreement can serve as evidence to assess the reliability of a research effort. Accordingly and in contrast to Kuckartz's view, observed agreement is not a final end, but instrumentally relevant for reliability. What is more (and as elaborated above), agreement should not be achieved by any means. In particular, it has to be the result of independent coding to be evidentially relevant for assessing reliability.

analysis regards reliability as dispensable, although it is a fundamental desideratum of science in general, as argued above.

Qualitative content analysts do not doubt the importance of reliability. It stands, therefore, to reason that they often cannot satisfy reliability because adhering to reliability would boil down to giving up something else that is important to them. In other words, the qualitative content analyst believes that there is a trade-off between reliability and something else.⁶²

It is instructive to elaborate on another distinction that is often used to clarify the difference between both forms of content analysis. According to this story, quantitative content analysis is confined to mere manifest content, whereas qualitative content analysis is able to include latent content as well. Understanding this distinction will help us to pinpoint the precise trade-off and will, additionally, contribute to our understanding of the category system's role in the connection with hermeneutical underdetermination of categorisation.

Let us, first, understand the difference between manifest and latent content. Manifest content covers those properties of text segments that are “on the surface and easily observable” (Potter and Levine-Donnerstein 1999, 259) and “elements that are physically present and countable” (Gray and Densten 1998, 420). Conversely, latent content must be searched for “between the lines” and cannot be observed directly but must be inferred indirectly by interpreting the text. Latent content is, on this view, not diametrical to manifest content but rather something more complex because it piggybacks on manifest content: Understanding latent content depends on understanding manifest content. However, this characterisation is admittedly too fuzzy. There are at least two different interpretations that provide a more precise distinction.

A narrow conceptualisation of manifest content confines it to mere syntactical or physical properties of texts—properties that are objectively present and that can be immediately observed. These may include occurrences and frequencies of specific word tokens, larger grammatical structures and other properties that can be identified by mere mechanistic or simplistic algorithmic means—in other words, properties that do not involve meaning. As soon as the annotator had to invoke their understanding of the given text, it would amount to latent content. According to this view, the distinction between manifest and latent content boils down to whether the categorisation of coding units demands the comprehension of meaning—that is, the consideration of semantical and pragmatic aspects.

However, the narrow conceptualisation of the distinction is only a foil to understand the

⁶²Kuckartz, in contrast, advances another reason why the reliability requirement cannot be satisfied in qualitative content analysis. He argues that chance-corrected reliability measures cannot be used in qualitative content analysis. He observes that in qualitative content analysis, annotators often have to identify coding units (Kuckartz 2012, 211). As a consequence, a reliability measure should not only measure differences between categorisations but has, additionally, to account for differences between segmentations of the text into coding units. He then goes on to suggest a reliability measure that is based on a percentage agreement and argues that any form of chance correction is negligible (Kuckartz 2012, 212–16). There are two problems with his argumentation. First, even if a chance correction is negligible for his suggested reliability measure, there might be other reasonable chance-corrected reliability measures. In fact, there are some more recent attempts to provide such measures (see, e.g., Fournier and Inkpen (2012), Krippendorff (2004c), Krippendorff et al. (2016), Mathet, Widlöcher, and Métivier (2015)). Second, even if the use of chance-corrected reliability would not be advisable in qualitative content analysis, it is premature to conclude to do without the reliability requirement altogether—what he suggests by advancing consensus coding as an alternative (see above).

following points. If we want to employ the distinction as a distinguishing feature, we need a more inclusive understanding of manifest content since the narrow understanding doesn't do any justice to quantitative content analysis: It is simply not true that quantitative content analysis is confined to manifest content in the narrow sense (Krippendorff 2004a, 20).⁶³

On a *broader understanding*, manifest content is not confined to syntactical properties but includes meaning as well, as long as the meaning is shared among competent speakers. On this view, manifest content covers meanings on which everyone will agree. If, on the other hand, different people do not agree on a meaning due to their diverging interpretations, the corresponding content is latent. This suggestion to clarify the manifest-latent distinction is more charitable since it does not make the mistake of assuming that the categorisation of manifest content does not involve interpretation.⁶⁴ The distinguishing feature is, on this reading, not whether the annotator has to go through a process of interpretation to categorise but whether annotators can agree on what is meant.

This broader understanding of manifest content raises a worry that explains the qualitative content analyst's preference to include latent content during the coding. Jürgen Kriz pointedly formulated the problem. He begins by observing that text messages allow, in the majority of cases, for different interpretations. Apart from very specific contexts that are governed by formalised or rigorous linguistic conventions, the understanding of meaning is usually hermeneutically underdetermined (Kriz 1978, 38). He labours under the assumption that meaningful manifest content—that is, manifest content that surpasses mere syntactical properties of coding units—is only to be found in formalised or highly conventionalised parts of our language. The worry is that if the quantitative content analyst is restricted to manifest content to satisfy coding reproducibility, they are at the same time restricted to uninformative or even trivial findings since meaningful manifest content is seldomly found. He surmises that the quantitative content analyst prefers “objective nonsense over meaningful subjective content” to satisfy the requirement of reliability (Kriz 1978, 46, translation mine). That is why the qualitative content analyst is interested in latent content even if it threatens the reliability of the coding process. The presumed trade-off is, therefore, between reliability and the informational content of the generated data.⁶⁵

⁶³Berelson (1952) invoked this narrow conception by saying that “the only sense in which ‘manifest content’ exists is in the form of black-marks-on-white” (19). However, he referred to a hypothetical argument without committing himself to this narrow understanding of manifest content, quite the contrary. He writes that “content analysis proceeds in terms of what-is-said” (16), which, to his understanding, includes meaning.

⁶⁴It was, in retrospect, unfortunate that Berelson (1952) introduced this terminology when he defined content analysis as “a research technique for the objective, systematic, and quantitative description of the manifest content of communication” (Berelson 1952, 18). He borrowed the manifest-latent distinction from psychoanalysis, where the manifest content of a dream is equated with what actually happens in the dream. The latent content of the dream is the meaning that the psychoanalyst is supposed to uncover by interpreting the dream. It is not surprising that subsequent authors took this analogy all too seriously and accordingly criticised Berelson's definition for its apparent restriction to manifest content in the narrow sense.

⁶⁵This trade-off is often framed as a trade-off between reliability and validity. For instance, Krippendorff writes: “In the pursuit of high reliability, validity tends to get lost. This statement describes the analyst's common dilemma of having to choose between interesting but nonreproducible interpretations that intelligent readers of texts may offer each other in conversations and oversimplified or superficial but reliable text analyses generated through the use of computers or carefully instructed human coders” (Krippendorff 2004a, 213). The question is to what kind of validity he refers. It is surely not the concept I discussed above—that is, the validity of categorisation. The latter is used to distinguish between correct and incorrect categorisations of coding units. In this case, the category definitions *determine* what counts as valid. In

While these considerations—especially the identification of the relevant trade-off—go in the right direction, the confinement to manifest content (in the broader sense) does not necessarily preclude the content analyst to gather informationally relevant data (w.r.t. a research question). Informationally relevant manifest content is more prevalent in our natural language than Kriz suggests. The problem with his objection is that it underestimates the role of the category system. The definition of category values, along with clarifications and examples, is, among other things, introduced to narrow down the degree of interpretation. If we understand better how definitions of category values can lead to converging categorisations among annotators, we will better understand how informationally relevant manifest content can be found in our natural language.

Let me first note that defining category values is not intended to transform latent content to manifest content by redefining vague, ambiguous or fuzzy terms that appear in coding units. The application of the category system is not supposed to replace or alter the annotator's understanding of text messages. As Kriz (1978, 45) correctly points out, the annotator has to approach the meaning of text messages in the same way as any other competent speaker. They are neither in a better position to determine some average or ordinary meaning—whatever that might mean—nor to better pick out the most appropriate interpretation of the text message's meaning.⁶⁶

But how can the category system narrow down the interpretational leeway in the categorisation of coding units if category definitions do not narrow down the fuzziness in the meaning of coding units?

Let me start by noting that the definite determiner in 'the meaning of a coding unit' is a little misleading. We should better think of a coding unit's meaning in terms of different meaning aspects. Each aspect corresponds to a specific question that might be raised to query its meaning. We might, for instance, ask whether a coding unit contains a description of a car and its owner. We might further ask whether the owner's attitude toward their car is described. Then, we might ask whether the coding unit contains a description of the car's age and so on. Obviously, there are uncountable possible meaning aspects. We can now define the *overall meaning* of a coding unit as the entirety of all meaning aspects—that is the answers to all these questions. Additionally, we can define the *shared meaning* of a coding unit with respect to a specified group of people as those meaning aspects to which everyone in this group will agree.

Hence, the shared meaning of a coding unit is determined by those questions on which

contrast, Krippendorff discusses the extent to which a category system *is itself* valid. Instead, I take it that Krippendorff's mentioned validity notion is a form of measurement validity, broadly characterised as the requirement that we should measure what we want to measure—namely, something relevant or informative concerning the research question. To be more precise, the produced data of the coding process is supposed to be evidentially relevant to the research question. In other words, it must be possible to interpret coding results as confirming (or disconfirming) evidence for the research hypotheses in question. The content analyst is therefore not free to use any categories whatsoever. Some will be more relevant than others. The data can, in this sense, be “interesting” or too “oversimplified and superficial” to answer the research question. The trade-off that Krippendorff describes is, therefore, a trade-off between reliability and the evidential relevance, or informational content of data generated with a specific category system.

⁶⁶Surely, the annotator is advised to use informative context knowledge to understand text messages, but that does hold similarly for others and does not clarify the role of the category system in connection with hermeneutical underdetermination.

the answers of different annotators coincide. But then, annotators can diverge in their understanding of a coding unit's overall meaning while there is, at the same time, a shared meaning. For instance, one annotator might understand a coding unit as describing a car that is loved and driven by its owner. In contrast, a second annotator understands it as a description of a car driven by its owner, who has no emotional connection to it. Both annotators have different interpretations, but there is still a shared meaning: A description of a car its owner drives. Hence, we might characterise the shared meaning as the common denominator of all different interpretations—or, to use another mathematical metaphor, the set-theoretic cut of peoples' interpretations.

We might, therefore, agree with Kriz (1978) that the *overall meaning* of most natural language expressions is hermeneutically underdetermined—in the sense that different people will come to diverging interpretations—but add that this does not preclude them from having a shared meaning. Natural-language expressions can have a shared meaning (and thereby non-trivial manifest content), even if there are diverging interpretations of their overall meaning. Consequently, a coding unit can have shared meaning and latent content at the same time.

We are now able to understand the ramifications of category definitions for hermeneutical underdetermination of categorisation. Instead of redefining the meaning of coding units, category definitions pin down those meaning aspects that are relevant for categorising coding units. As Früh (2017, 118) puts it, the category system provides a search strategy, which guides the coding process. In this way, the category system determines a particular context within which the annotator analyses the meaning of coding units (Krippendorff 2004a, 24).

Not all meaning aspects of a coding unit are necessarily relevant for the categorisation. In other words, annotators do not have to consider the overall meaning of a coding unit. The question is whether the categorisation can be performed by considering the manifest content (shared meaning) alone or whether its latent content is relevant as well. While annotators will usually not coincide in the identification of all meaning aspects of a coding unit, they will agree on some meaning aspects. If, now, the shared meaning of a coding unit comprises all meaning aspects that are relevant for the categorisation, properly instructed annotators are expected to converge in their categorisation.⁶⁷ Hence, the hermeneutical underdetermination of the coding unit's overall meaning does not imply an underdetermination in the categorisation of coding units.

I already argued that inter-annotator reliability—interpreted as reproducibility of coding results—is an important desideratum of scientific inquiry. But why, we might ask, is the confinement to manifest content a problem? If reliability is so important, the content analyst might simply confine the definition of category values by relying on shared content only. Surely, every coding unit has at least some shared content—that is, some meaning aspects on which every annotator can agree. In other words, why is hermeneutical underdetermination of categorisation a problem if the researcher is free in defining their category values? What does *irreducible degree of interpretation* mean if every coding unit has at least

⁶⁷This important point is also made by Früh, who frames it in a slightly different way. In discussing the above considerations of Kriz (1978), he notes that we “have to distinguish between the definition of categories and the identification of indicators in coding units” (Früh 2017, 117, translation mine).

some shared meaning?

The point is that the research question puts some constraining conditions on the definition of category values. Category values cannot be arbitrarily defined once the research question is set. Not every conceivable category system is informative or evidentially relevant for a specific research hypothesis. Accordingly, it can happen that a research question cannot be answered satisfactorily if the researcher confines the definition of category value to manifest content. In other words, the meaning aspects that are relevant to answer the research question may exceed the shared content of coding units. In the design phase, researchers should revise their category system to decrease the degree of hermeneutical underdetermination. However, the research question constitutes a constraining boundary condition for the prospects of simultaneously maximising reliability and evidential relevance by improving the category system. There will be a point when the content analyst has found a category system whose reliability cannot be further improved without aggravating its evidential relevance. If, at this point, there is still a non-vanishing degree of hermeneutical underdetermination, we have a case of an *irreducible degree of hermeneutical underdetermination* (w.r.t. a research hypothesis). The question is how to meet this challenge of hermeneutical underdetermination.

3.3.4 THE CHALLENGE OF HERMENEUTICAL UNDERDETERMINATION

In the last section, I argued that the main distinguishing feature between quantitative and qualitative content analysis is their different stance towards the role of reliability. I moved on to explain that this difference is rooted in diverging views of how to balance reliability against informational content of coding results. Finally, I used the distinction between manifest and latent content to shed some light on the role of the category system in connection to hermeneutical underdetermination.

Based on these considerations, we are now in a position to clarify how both paradigms deal with hermeneutical underdetermination of categorisation. Do they presuppose that there is exactly one correct category value for each coding unit—that is, no hermeneutical underdetermination—or can they deal with situations in which this uniqueness assumption is violated?

Let's start with qualitative content analysis. The qualitative content analyst does not shy away from category definitions that can lead to different correct coding results among independently working annotators. Rather, they emphasise the role of interpretation in the coding process and suggest using hermeneutical methods borrowed from other disciplines (Kuckartz 2012, 16–21; Mayring 2015b, 29–32).⁶⁸ This is, additionally, in line with characterising qualitative content analysis as a method that is not confined to manifest content. If we take the suggested explication of the latent-manifest distinction, latent content is by definition hermeneutically underdetermined. In other words, if category definitions draw on latent content, different annotators may come to different coding results without making any errors during the categorisation. Hence, the involvement of latent content is inconsistent with uniqueness assumption.

⁶⁸But that does not amount to an anything-goes attitude towards categorisation. While there are no correct or incorrect categorisations per se, they can be more or less adequate (Kuckartz 2012, 20).

So it seems that the qualitative content analyst does not labour under the uniqueness assumption. Different annotators can come to different categorisations that might be equally adequate. This interpretation is, however, in some tension with the role of consensus coding: Why bother to let annotators work together to reach an agreement if they can come to equally adequate categorisations by themselves? As I argued above, the resulting agreement has nothing to do with reliability and reproducibility. The leading idea is, perhaps, that the collaborative effort is supposed to identify the most adequate categorisation among those that annotators choose individually. But then there seems to be an implicit commitment to the uniqueness assumptions. Inasmuch as annotators should come to a collaborative decision as to what the most adequate categorisation is, they are supposed to decide on *one* categorisation. In contrast to quantitative content analysis, the correct or most adequate category value is thusly not determined by the category system alone but by the category system together with a process of consensus coding.

For the following considerations it is not important to decide this issue. We can leave it at two possible interpretations: On one view, the qualitative content analyst does not presuppose uniquely determined category values. On the alternative view, the qualitative content analyst commits themselves to the uniqueness assumption implicitly by advancing consensus coding. As argued above, consensus coding alone doesn't have anything to do with reliability. That is, on both of these interpretations, the reliability requirement is of no concern. Hence, the qualitative content analyst does not provide a method that combines faithfulness to the requirement of reliability with a concession to hermeneutical underdetermination.

The picture in quantitative content analysis is more straightforward, or so it seems. As elaborated above, the uniqueness assumption is used to explicate the relation between reliability and validity. Inasmuch as reliability is supposed to be necessary for validity, the quantitative content analyst commits themselves to the uniqueness assumption. Only if we assume that there is one unique correct category value, a difference in the categorisation of one coding unit implies that one of the annotators must be wrong.

This view is, additionally, in agreement with a characterisation of quantitative content analysis as confined to manifest content. If the definition of category values draws on shared content only, then there is necessarily only one correct category value for each coding unit. The shared content of a coding unit comprises by definition only those meaning aspects on which properly instructed annotators will agree.

On the latter view, quantitative content analysis is committed to the uniqueness assumption and can, thusly, not be applied in contexts where latent content is necessary to answer a research question. However, the question is whether there is another interpretation of quantitative content analysis. In particular, I do not want to take a stance on whether quantitative content analysis is committed to the confinement of manifest content. So let us assume that it is not, and let us ask whether the graded picture of reliability is committed to the uniqueness assumption. Can we interpret the introduction of chance-corrected reliability measures as a concession to the uniqueness assumption, or, in other words, as an expansion of quantitative content analysis to latent content?

It is instructive to have a look at how reliability thresholds are justified to answer this question. Unfortunately, justifications are sparse and partly opaque. Some go as far as

thinking that “deciding what counts as an adequate level of agreement for a specific purpose is still little more than a black art” (Artstein and Poesio 2008, 576).⁶⁹ The most explicit expositions I found, are formulated by Krippendorff, who writes that “the choice of reliability standards should always be related to the validity requirements imposed on the research results, specifically to the costs of drawing wrong conclusions” (Krippendorff 2004a, 242; see also Krippendorff 2004b). According to this picture, lower reliability values increase the chance of erroneous conclusions.

It seems that Krippendorff invokes ideas that are similar to the statistical method of hypothesis testing. In hypothesis testing the acceptance and rejection of hypotheses depends on the willingness accepting a specific probability to reject a true hypothesis, or to accept a false hypothesis, respectively. In other words, what is rationally acceptable and rejectable depends on the probability of being wrong. And, similarly as suggested by Krippendorff, the corresponding thresholds are usually justified by recurrence to the personal or societal costs of being wrong.

The problem is that Krippendorff does not elaborate on the precise relationship between reliability thresholds and probabilities to “draw wrong conclusions”. He does not tell us how diverging categorisations as measured by reliability values result in erroneous conclusions. I suppose the general picture is this: Disagreements in coding results are connected to the chance of incorrect categorisations. This interpretation is in line with the broader view of reliability, which conceptualises reliability as an indicator of categorisation validity. And incorrect categorisations are likely correlated with false conclusions based on these categorisations. In this way, reliabilities are indirectly related to the chance of drawing wrong conclusions.

The precise probabilistic relationship between incorrect categorisations and wrong conclusions will depend on the broader research context, in particular on the research hypotheses in question. But this does not need to concern the justification of reliability thresholds. Rather, we can interpret them as a mere indicator of categorisation validity. On this view, there is an inverse relationship between reliability values and the chance of incorrect categorisations. Low reliability values correlate with high probabilities of incorrect categorisation.

Is this picture committed to the uniqueness assumption? Even without elaborating on the precise relationship between reliability values and incorrect codings, it should be clear that the chance of incorrect coding results depends not only on the reliability value itself but also on the range of correct category values. Take the limiting case of categorisation by chance alone—that is, a reliability value of zero. Let’s further assume a very simplistic chance mechanism of a uniform distribution: The chance of categorising a coding unit with a specific category value is the same for every category value and equals 1 divided by the number of categories. In this case, the chance of a correct categorisation is 0.5 if half of the category values represent correct category values. Similarly, if a tenth of all category values represent correct category values, the chance is 0.1. Reliability values above zero can be interpreted by saying that a certain share of all categorisations is as “good” as a pure chance mechanism (Krippendorff 2004a, 226). Consequently, the considerations for the case of nil reliability apply (with some modifications) similarly to other reliability values. For any fixed reliability value, we will, therefore, observe roughly the following

⁶⁹See also Neuendorf (2002, 143).

relationship: If the number of correct category values increases, the chance of incorrect categorisations decreases. Hence, the chance of a correct categorisation can generally not be calculated based on the reliability value alone. The range of correct category values has to be taken into account too.

But if reliability thresholds should be chosen in relation to the chance of incorrect codings, and if these chances depend on the range of correct category values, reliability thresholds should also depend on this range. In other words, reliability thresholds should depend on the context—a context that is determined by its degree of hermeneutical underdetermination. However, I did not find any explicit consideration of this context dependence in the methodological literature. The most charitable interpretation is that these thresholds are implicitly based on one particular assumption about the extent of hermeneutical underdetermination for all contexts. It stands to reason that this implicit assumption is simply the uniqueness assumption. First, the uniqueness assumption is already used in connection with the simple relationship of reliability being necessary for validity; second, the uniqueness assumption is at least framed as an ideal for which the content analyst should strive (Krippendorff 2004a, 24; Früh 2017, 112).

Again, we do not need to decide this issue. Either the graded picture of reliability is (at least implicitly) based on the uniqueness assumption or it is based on some specific assumption about the extent of hermeneutical underdetermination. Either way, the quantitative content analyst misses to explicate the discussed context dependency and does not introduce the corresponding context-dependent reliability thresholds. Surely, the degree of underdetermination hinges crucially on the introduced category system and will therefore vary between different research contexts. Hence, similarly to the qualitative content analyst, the quantitative content analyst does not provide a method that combines faithfulness to the requirement of reliability with a concession to (context-dependent) hermeneutical underdetermination.

3.4 PHENOMENON SENSITIVITY

In the last section, I argued that the two prominent paradigms of content analysis do not provide methods that unite faithfulness to the reliability requirement with a concession to hermeneutical underdetermination of categorisation. It seems that the content analyst has to decide: Either they comply with the reliability requirement without being able to deal with cases of an irreducible degree of hermeneutical underdetermination, or they include such cases on pain of violating the reliability requirement.

The central tenet of this work is that both discussed paradigms of content analysis do not exhaust the set of available options. Instead, I will argue in the following sections that an irreducible degree of interpretation does not necessarily impede the reliability of the data-making process. The preceding discussion suggests that one possibility is to introduce context-dependent reliability thresholds based on chance-corrected reliability measures. I will, however, not pursue this line of reasoning. Chance-corrected reliability measures assess to what extent annotators perform better than chance. Instead of measuring the performance of annotators *in comparison to* a chance mechanism, I will devise the following reliability considerations from a different perspective of chancy categorisation: I

will assume that the space of different, but equally correct categorisations can be modelled by a probability distribution and will ask under which circumstances the researcher can reliably infer something about the phenomena of interest *in spite of* chancy categorisation.

The subsequent considerations will proceed along the following steps: I will, first, introduce the notions of strong and weak phenomenon sensitivity as an additional framing of the reliability concept (3.4.1). Then, I elaborate on how the reliability requirement is met within the natural sciences (3.4.2). It will turn out that the theory of errors utilises the notion of weak phenomenon sensitivity that can be used as an analogy by using the measurement picture of annotation. I will end this chapter by using this analogy to develop an alternative probabilistic conceptualisation of reliability that allows for hermeneutical underdetermination (3.4.3).

3.4.1 PHENOMENON SENSITIVITY AND RELIABILITY

Quantitative content analysts are generally interested in an observer-independent reality. They strive to gather knowledge about the social phenomena they study, and their measurements should enable them to infer something about them. To that end, the quantitative content analyst has to be sure that a difference between coding results corresponds to a difference in the phenomena and is not the result of divergent interpretations. In other words, coding results should be sensitive to differences in phenomena.

This notion of *phenomenon sensitivity* is a property of the measurement process and is closely related to the concept of reliability that I introduced above. As Krippendorff (2016) puts it: “Measures of replicability need to assure researchers that the variance of the generated data is explainable by the differences that annotators detected among the phenomena they recorded” (139). Krippendorff suggests that reliability is closely connected to the requirement of phenomenon sensitivity: By successfully reproducing a measurement on *one* phenomenon, the researcher can be sure that differences in measurement results on *different* phenomena can only be explained by a difference between these phenomena. In fact, both concepts are more or less equivalent. Reliability is sufficient for phenomenon sensitivity and the other way around.

Let us, first, clarify how reliability implies phenomenon sensitivity. The basic idea is this: If there would be other causes for an observed difference than the differences between the different phenomena, the same causes would lead to a violation of reliability. In more detail: We can safely assume that there are only three (not necessarily exclusive) possibilities to explain differences in measurements on different phenomena. Either the difference results from the differences between these distinct phenomena, from differences between the measurement devices, or an involved chance mechanism. Suppose now that the requirement of phenomenon sensitivity is violated. In other words, let’s suppose that a difference in measurement results of different phenomena is the result of differences between the measurement devices or the result of a chance mechanism. But then these differences should similarly produce differences in repeated measurements of one and the same phenomenon. This consideration implies that if we can successfully reproduce measurements on one phenomenon, we can exclude the last two possible explanations. In other words, successful reproductions of measurement results on one phenomenon leave the researcher only one explanation for differences between measurements on different

phenomena: They have to be explained by differences between them. Hence, the suggested conceptualisation of reliability implies phenomenon sensitivity.

Reliability is not only sufficient for phenomenon sensitivity but also necessary. In other words, phenomenon sensitivity is only satisfied if the reliability requirement is met. Suppose that two measurements on the same phenomenon lead to different measurement results. These differences cannot be explained by differences between the phenomena of interest because there is only one phenomenon. Hence, the difference must be explained by something else, for instance, by differences between the measurement instruments. Whatever the precise cause, it will at least occasionally produce differences between measurements of different phenomena. Hence, there will be differences in measurements of different phenomena that can be explained by something other than the differences between these phenomena. Thus, a violation of the reliability requirement will also lead to a violation of phenomenon sensitivity.

The standard of phenomenon sensitivity is matched in content analysis if observed differences between categorisations correspond to differences in the phenomena and not between annotators. The concept I explained so far can be called *strong phenomenon sensitivity*: A measurement instrument satisfies strong phenomenon sensitivity if *every* difference between annotation results is only explainable by a difference in the phenomena and henceforth not by a difference in interpretation. In other words, *every* difference in the measurement results *implies* a difference in the phenomena.

Obviously, strong phenomenon sensitivity cannot be satisfied in the case of hermeneutical underdetermination. In other words, strong phenomenon sensitivity implies the uniqueness assumption (i.e., that there is only one correct category value for each coding unit). If two annotators categorise one coding unit correctly but differently, this difference cannot be explained by a difference in the phenomenon of interest because there is only one phenomenon. Naturally, the difference must be explained by interpretational differences.

The question is whether there is some weakening of strong phenomenon sensitivity that can deal with hermeneutical underdetermination and that will still allow the researcher to use annotation results to identify differences in the phenomena. The basic idea of this work is to utilise two ways of weakening strong phenomenon sensitivity: Instead of demanding that *every* difference in measurement results *implies* a difference in the phenomena, we only require that *some* of differences between measurements *statistically imply* differences in the phenomena. *Weak phenomenon sensitivity* is satisfied if there is a subset of differences in the set of all possible differences in the measurement results that allow statistically valid inferences to corresponding differences in the phenomena.

The rest of this work is devoted to elaborate on this idea by specifying under which conditions the researcher can draw such statistical inferences and to explain the notion of statistical inference in more detail.

3.4.2 THE ANALOGY FROM ERROR THEORY

In content analysis, strong phenomenon sensitivity cannot be achieved in cases of hermeneutical underdetermination. But is weak phenomenon sensitivity an attainable goal? It will be helpful to reassess the measurement picture of content analysis to answer this

question. I will, first, clarify how measurements in natural sciences comply with phenomenon sensitivity. Then, I will move on to transfer all relevant notions to content analysis to formulate conditions under which the categorisation of coding units can satisfy weak phenomenon sensitivity.

Even in the natural sciences, strong phenomenon sensitivity cannot be reached. Suppose a natural scientist intends to measure the magnitude of a quantity in a specific target system. To do so, they will usually use a measurement apparatus that can indicate its different states by, for instance, a pointer that can move across a scale or a display that provides numbers. The measurement apparatus is brought into some interaction with the target system, after which it will end up in an equilibrium state, where the pointer or display will remain stable. The resulting measurement value—the measurement indication (JCGM 2012, 37)—is supposed to inform the scientists about the quantity they intend to measure.

But is every such indication value to be equated with the quantity value of the target system? Suppose, for instance, that the least significant digit on the readout of a digital volt meter corresponds to hundredths of a volt ($1/100V$) and that the display provides a measurement indication of 2.00 volt. Should the scientist trust this value? Is the quantity value they intend to measure really 2.0000 volt rather than, say, 2.0001 volt? The variability of measurement results aggravates this worry. If the resolution of a measurement device is sufficiently high, the scientist will yield different indication values even if they replicate the measurement under conditions of repeatability—that is, if they use “the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time” (JCGM 2012, 23). The scientist should, therefore, not fully trust the given indication value to correspond to the quantity value they intend to measure. Consequently, every single measurement indication should be considered a mere estimate with a certain degree of uncertainty.

But how should we further qualify this degree of measurement uncertainty? The basic idea is to express the uncertainty by an interval $v \pm \Delta$ around the measurement indication v . This so-called confidence interval is to be understood to contain values the scientist can be sufficiently certain to coincide with what they intend to measure. In this way, the confidence interval is supposed to quantify the uncertainty of the measurement. A satisfactory theory of measurement uncertainty that is based on this schematic suggestion must provide answers to the following questions:

1. **Justification of uncertainty:** How can the scientist determine and justify a level of uncertainty (i.e., a confidence interval)?
2. **Interpretation of uncertainty:** How should they interpret the uncertainty that is expressed by a confidence interval?
3. **Uncertainty & phenomenon sensitivity:** How do uncertainties relate to phenomenon uncertainty? How can the scientist use their knowledge of uncertainties to judge whether an observed difference between two measurement indications can be taken to imply a difference in the measured quantities—that is, a difference in the phenomena?

A preliminary conceptualisation of confidence intervals takes the variability of measurement results as a starting point to answer the first question. The measurement of a

quantity is repeated multiple times under conditions of repeatability. The resulting sample is then used to determine the uncertainty of measurements associated with one particular measurement apparatus.

Confidence intervals of different measurements can then be compared to each other to decide whether observed differences are significantly different. In other words, based on the sketched notion of uncertainty, a theory of uncertainty can formulate conditions under which measured differences allow the scientist to conclude that there is a difference in the phenomena. As we will see shortly, these conditions cannot be expected to be generally satisfied. As a consequence, the scientist has to do with weak phenomenon sensitivity.

Let us now look closely at one prominent theory of measurement uncertainty. The classical way to answer the above questions is based on the theory of errors (Edgeworth 1888; Taylor 1996). According to this theory, the uncertainty of measurements is conceptualised by resorting to the concept of measurement errors.⁷⁰ This theory postulates that the quantities to be measured do not change under conditions of repeatability and that, thusly, the variability of multiple measurements under such conditions must be interpreted as the result of measurement errors. On this view, the observed indication value v_i can be understood as the composition of the true quantity value v_t and a measurement error e : $v_i = v_t + e$.

The scientist has to determine the extent of the measurement errors to quantify the uncertainty of measurements. There are two different sources of errors that correspond to two different types of errors. An error is called systematic if its magnitude and direction do not change for different measurements (Taylor 1996, 94) or if its variation is predictable (JCGM 2012, 22). A typical source for such errors is the miscalibration of measurement devices. For instance, the scale of a mercury-in-glass thermometer might get out of place so that it constantly overestimates the temperature by five degrees Celsius. Naturally, systematic errors cannot be successfully chased by repeating measurements under conditions of repeatability since the error would be the same for every measurement. The only way to find them is to compare measurement results to some known or hypothesised standard to provide more accurate values (Taylor 1996, 97). Random errors are—as the name suggests—deviations from the true value resulting from accidental disturbances. These deviations are unpredictable and uncontrollable but thought of as the result of a chance process.

The difference and interplay between both types are often illustrated with the archer analogy (Tal 2020; Taylor 1996, 94–95). Consider the series of pictures in Figure 3.1, which depict bullet holes produced by an archer or gunman who aims to hit the bull's eye. In this analogy, the centre of the target represents the true value. The closeness of hits to the bull's eye represents the accuracy of a measurement procedure. The different pictures in Figure 3.1 illustrate the different error types and their composition. Even if the archer is generally good at hitting the target, there might be disturbing influences, for instance, fluctuations of atmospheric conditions that vary from shot to shot randomly and force the arrow off target. As a result, multiple shots will be distributed around the bull's eye (Figures 3.1a and 3.1b). The spread of the resulting distribution represents the extent of random errors in the analogy. If, on the other hand, the archer has a general tendency to miss the target into a specific direction, the centre of the distribution will not coincide with

⁷⁰Taylor (1996) even uses both concepts interchangeably.

the bull's eyes (Figures 3.1c and 3.1d). This replicable tendency corresponds to systematic errors in the analogy. As the figures illustrate, both types can occur simultaneously and vary in their extent.

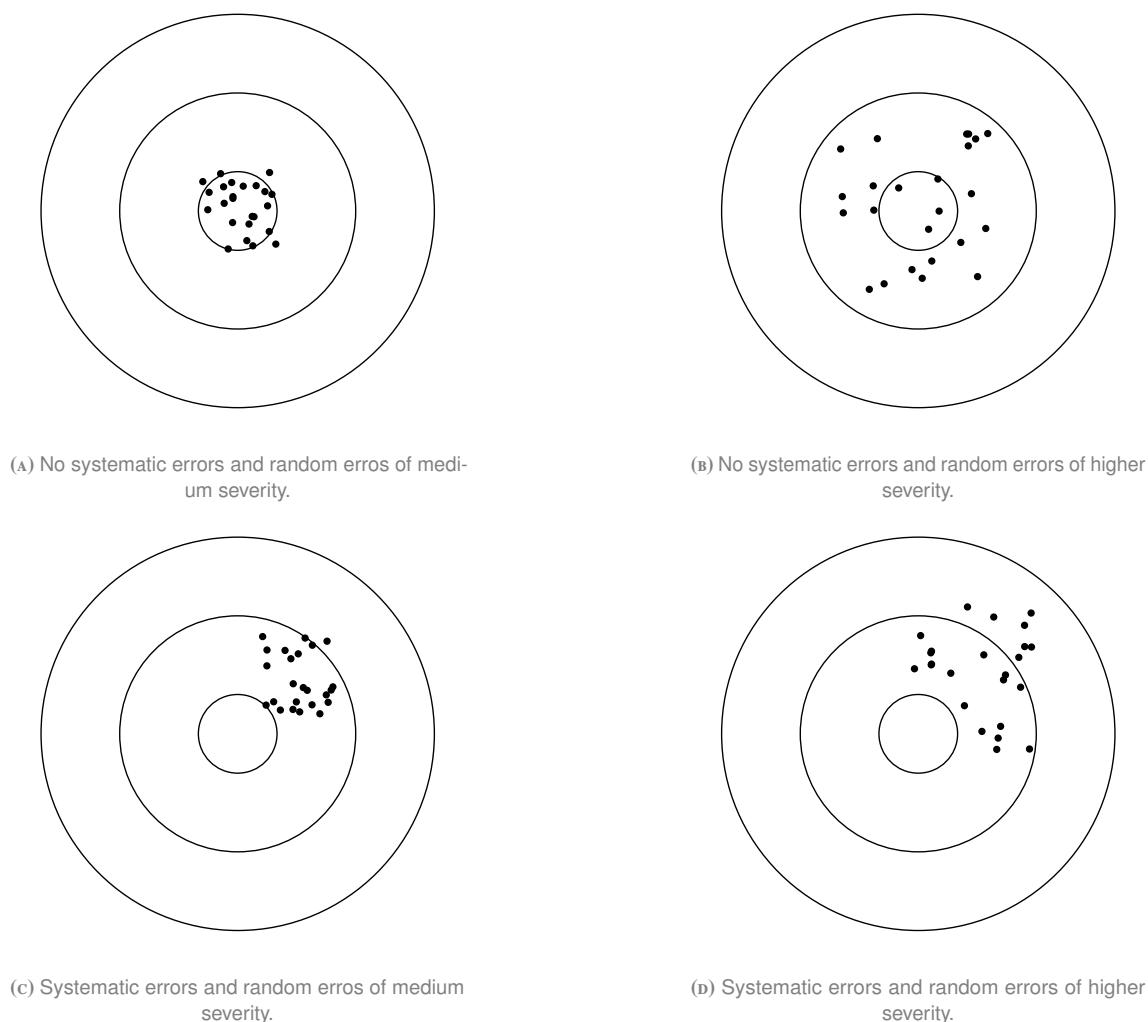


ABBILDUNG 3.1
Random and systematic errors in the archer analogy.

Additionally, the analogy illustrates an epistemological worry associated with systematic errors. If we do not have an external truth standard to compare measurements values—that is, if we erase every indication of where the bull's eye is in the picture, we cannot compare the different pictures concerning the question of systematic errors (Taylor 1996, 95). The problem is now that, at least for continuous quantities, the exact true values are unknowable or that the question of systematic errors reoccurs for every reference value that serves as a hypothesised truth standard (Tal 2020).⁷¹

Fortunately, we do not have to worry about systematic errors since we are only inte-

⁷¹One possible consequence of this worry is to drop the error-based theory of measurement uncertainty. The uncertainty approach to measurement is an alternative theory that dispenses with the notion of true values, or at least regards them as fictitious idealisations (for an overview, see de Courtenay and Grégis 2017; Grégis 2015).

rested in understanding the role of phenomenon sensitivity. If we compare two different measurements—that is, multiple measurements not performed under conditions of repeatability—but under the same measurement procedure, the systematic error will be the same for both measurements. By comparing two measurement indications (of different measurements), we are interested in whether there is a significant difference between both measurements. In other words, we are interested in the expression $v_i - v_j$, in which the systematic errors will cancel each other out. Hence, systematic errors can be neglected for questions of phenomenon sensitivity.⁷² For simplicity of exposition, we will assume there are no systematic errors in the following.

On the view elaborated so far, the uncertainty of measurements pertaining to the phenomenon sensitivity is determined by the random error. Naturally, the error-based approach conceptualises random errors in a probabilistic fashion. In other words, the error-based approach is grounded in a statistical model of measurement uncertainty. The random-error part of a measurement value is thought of as a random drawing of a value from the population of all possible random errors. The probability distribution of this population is usually assumed to be bell-shaped or, in other words, normally distributed around the true value.⁷³ As we will now see, the determination and justification of the uncertainty—that is, answering the first question—is tantamount to determining all relevant parameters of this distribution.

Two parameters characterise the normal distribution: The expectation value which corresponds to the true value v_t and the variance σ^2 of the distribution, which measures the spread of the normal distribution (see Figure 3.2). Both parameters are usually not known prior to any measurement. Instead, they are estimated by replicating a measurement under conditions of repeatability. Let x_1, x_2, \dots, x_n the resulting measurement indications of such a sample. The measured mean of this sample $\mu_{\bar{x}} := \frac{1}{n} \sum_i x_i$ is used as a point estimate of the true value v_t . This point estimate can be regarded as the best estimate of the true value v_t in the following sense. First, the sample mean is a maximum likelihood estimator of the true value: The normal distribution (under any fixed σ^2) maximises the probability of the observed sample under the assumption that $v_t = \mu_{\bar{x}}$ (DeGroot and Schervish 2012, 420–21). Additionally, the point estimate will converge in probability towards the true value with an increase in the sample size n (DeGroot and Schervish 2012, 352–53).

However, any such estimation will most probably not coincide with the true value. In other words, any particular sample mean will most probably deviate from the true value. Due to the probabilistic modelling of the random error, deviations of observed sample means from the true value are probabilistically distributed. That motivates conceptualising the uncertainty of the estimated true value by using the distribution of sample means.

We can now formulate the relevant question in the following way: What is the distribution of sample means of a given normal distribution? The beautiful answer to this question is

⁷²It is probably no coincidence that the archer analogy is also used in content analysis (Krippendorff 2004a, 214; Neuendorf 2002, 114). In content analysis, the distribution's spread corresponds to reliability, whereas the distance of the distribution's centre to the bull's eye corresponds to the validity of coding. This picture is, thusly, in line with neglecting systematic errors for questions of phenomenon sensitivity since phenomenon sensitivity is just another framing of reliability (see above).

⁷³I use the term 'distribution' as an abbreviation for 'probability distribution' or 'probability density function' in the case of a continuous variable.

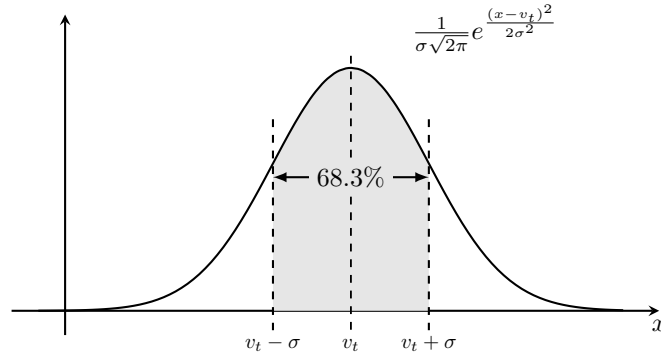


ABBILDUNG 3.2
Normal distribution around v_t with variance σ^2 .

perplexingly simple: The distribution of sample means of a normal distribution is itself a normal distribution around the true value. What is more, the variance of the sample distribution $\sigma_{\bar{x}}^2$ is mathematically connected to the variance σ^2 in the following way: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ (DeGroot and Schervish 2012, 311).⁷⁴

The confidence interval, which is supposed to quantify the uncertainty of the estimated true value, can now be expressed in the following form:

$$v_t = \mu_{\bar{x}} \pm c \frac{\sigma}{\sqrt{n}}$$

But how should we answer the second question—that is, how should we interpret this interval? Prior to observing a particular sample the probability of a sample mean $\mu_{\bar{x}}$ being between an interval around the true measurement value v_t can be expressed by $P(v_t - c \frac{\sigma}{\sqrt{n}} \leq \mu_{\bar{x}} \leq v_t + c \frac{\sigma}{\sqrt{n}}) = \gamma$. In this expression, γ and c are parameters for further specification, which are not independent. For $c = 1$, the probability becomes 0.68 ($\gamma = 0.68$) and the corresponding confidence interval is called the 68% confidence interval. For the 95% confidence interval, c becomes 1.96.

The preceding probability statement can be easily mathematically transformed into:

$$P(\mu_{\bar{x}} - c \frac{\sigma}{\sqrt{n}} \leq v_t \leq \mu_{\bar{x}} + c \frac{\sigma}{\sqrt{n}}) = \gamma$$

This expression should, however, not be interpreted as a probability of v_t being within the confidence interval of an observed sample.⁷⁵ The error-based approach is based on a frequentist interpretation of probability, which confines the assignment of probabilities to (sequences of) repeatable events resulting from a chance set-up. For any *given* observed sample \bar{x} there is no such chance set-up that can produce a sequence of true measurement values. Instead, for any given observed sample there is one true measurement value,

⁷⁴The variance σ^2 is usually not known and must be estimated by using the observed standard variance of the sample $\frac{1}{n-1} \sum (x_i - \mu_{\bar{x}})^2$ (Taylor 1996, 140).

⁷⁵Confidence intervals are prone to be misunderstood. For some prominent misreadings and their discussion, see Dickson and Baird (2011).

which is either within the confidence interval or not. But there is a frequentist long-run interpretation of the last expression in terms of creating confidence intervals containing the true measurement value: If we repeat the described procedure of estimating the true value by producing a sample of size n , the proportion of calculated 68% confidence intervals containing the true measurement value will tend (in probability) toward 0.68.

Another (complementary) interpretation of the confidence interval is linked to the statistical method of significance testing. Significance tests are used to decide whether observed data is statistically consistent with a probabilistic hypothesis—usually called the null hypothesis. The central idea is to calculate how likely the observed data is under the assumption of the probabilistic hypothesis in question. If this probability is sufficiently low, the data is considered significantly different and said to refute the null hypothesis.

Let's look at this rough picture in more detail using a particular example. Suppose a scientist has devised a measurement procedure to measure the speed of light. The scientist hypothesises that the measurements are normally distributed around the value $c = 300.000\text{km/s}$ with a variation σ^2 . To test this null hypothesis, the scientist generates a sample of n measurements and calculates its mean $\mu_{\bar{x}_{\text{observed}}}$. If, as hypothesised, the measurements are normally distributed around c , the mean is similarly normally distributed around c (DeGroot and Schervish 2012, 311). The question is now whether the deviation of the observed sample mean from c is significantly different. If it is, the data is regarded as refutational evidence against the null hypothesis.

To answer this question, the scientist will use the null hypothesis to calculate the probability of observing a sample mean $\mu_{\bar{x}}$ whose deviation from c is at least as high as the observed sample mean $\mu_{\bar{x}_{\text{observed}}}$. In other words, they calculate

$$P(|c - \mu_{\bar{x}}| \geq |c - \mu_{\bar{x}_{\text{observed}}}|; h_o)$$

under the assumption of the null hypothesis h_o . This is the so-called p -value of the observed sample. It measures the likelihood of observing as least as extreme deviations of sample means from c as the observed sample mean. The introduced notion of extremity motivates to regard observed sample data with higher p -values as less statistically consistent with the null hypothesis.

Finally, the scientist has to decide on a threshold for p -values—the so-called significance level α of a significance test—to decide whether an observed sample is considered significantly different. A typical value is $\alpha = 0.05$. If observed p -values are smaller than the significance level ($p < \alpha$), the observed sample is regarded as refutational evidence against the null hypothesis.

We can now come back to the interpretation of confidence intervals that are calculated by a sample of measurements under conditions of repeatability. One can show that the 95% confidence interval $\mu_{\bar{x}} \pm 1.96\sigma/\sqrt{n}$ contains those values for which the observed sample mean would not be significantly different at a significance level of $\alpha = 0.05$ (DeGroot and Schervish 2012, 540). In other words, every value within the confidence interval corresponds to a null hypothesis about the true measurement value that would not be rejected by the observed sample mean.

We now have an interpretation of confidence intervals that is motivated by the method of significance tests. On this view, we can use confidence intervals to compare observed measurement values with hypothesised true values. All values within the confidence interval correspond to hypothesised true values that a significance test would not reject.

Based on this interpretation of confidence intervals, we can now answer the third question—namely, the determination of conditions under which differences between measurements of two quantity values are significant enough to infer differences in the phenomena. So far, I did not suggest how to understand the notion of “significant difference” in the context of phenomenon sensitivity. The second interpretation of confidence intervals based on significance tests provides such a conceptual elaboration, which we will use now.

However, there is a small difference, which demands adding one further step to our previous considerations. So far, the given interpretation of confidence intervals allows us to compare one observed sample mean and its uncertainty to one (or more) particular hypothesised value(s). The (repeated) measurement of one quantity value together with its uncertainty that is represented by a confidence interval, enables the researcher to decide whether the measured quantity value is significantly different from one (or more) hypothesised value(s). The third question, however, concerns the comparison of measuring two quantity values, both of which are associated with uncertainties. The difference is that both values are to some extent uncertain and both corresponding confidence intervals have to be considered in deciding whether the quantity values are significantly different. In other words, instead of comparing one confidence interval with one (or more) particular value(s), we want to compare two confidence intervals.

To be more precise, we will think of the situation as follows: We want to measure two quantity values v_1 and v_2 , which might be different. Drawing on the error-based account, we assume measurements to be normally distributed around v_1 and v_2 with variances σ_1^2 and σ_2^2 . As before, we base our measurements for each quantity on a sample of measurements. That is, we measure each quantity value n times under conditions of repeatability. The resulting samples x_1 and x_2 can now be used to calculate the means $\mu_{\bar{x}_1}$ and $\mu_{\bar{x}_2}$ and their associated confidence intervals.

How do we decide whether our observed sample means are sufficiently different to allow us to conclude that the two quantity values are different? It is possible to answer this question based on significance testing. We simply have to reframe the described general picture in the following way: We want to know whether $v_1 \neq v_2$. From the perspective of significance testing, we ask whether our observed data is sufficient to reject the null hypothesis that $v_1 = v_2$, in other words, that $v_1 - v_2 = 0$. We further know that the distribution of the difference between both sample means is normally distributed around $v_1 - v_2$ with the variance $\sigma_{1-2}^2 = \sigma_1^2/n + \sigma_2^2/n$ (DeGroot and Schervish 2012, 306–7). Hence, our null hypothesis is equivalent to saying that the difference between both sample means is normally distributed around 0. We can now use our observed sample means to calculate a 95% confidence interval for the difference between both sample means:

$$(\mu_{\bar{x}_1} - \mu_{\bar{x}_2}) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}$$

According to the second given interpretation of confidence intervals, we can now say that if 0 is not contained in the confidence interval, the hypothesis that both measurements are not different is rejected at a level $\alpha = 0.05$. In this case, the difference between observed sample means allows us to reject that $v_1 = v_2$, and, hence, allows us to infer that both quantity values are different.⁷⁶

3.4.3 A RECIPE FOR CONTENT ANALYSIS

The central tenet of this work is to use the described theory of measurement uncertainty as an analogy to devise a method that is able to deal with hermeneutical underdetermination in the categorisation of coding units. Since the method of significance testing helps the natural scientist to satisfy weak phenomenon sensitivity, the content analyst will be similarly able to fulfil this desideratum—even if annotators disagree in their categorisations due to interpretational differences. In other words, I claim that violating the uniqueness assumption does not necessarily lead to unreliable data in content analysis.

There are, obviously, important differences between measurements in the natural sciences and the categorisation of coding units. So, how do we translate the picture of confidence intervals to meet the challenge of hermeneutical underdetermination in content analysis? The described elaboration of the error-based theory of measurement uncertainty contains the following general elements:

1. *A general probabilistic conceptualisation of the measurement procedure:* A probabilistic description of the measurement process, which is tied to a specific interpretation of these probabilities.
2. *The formulation of a statistical model:* Given the general probabilistic conceptualisation, the measurement process is modelled by specific probabilistic assumptions.
3. *Formulation of conditions to satisfy weak phenomenon sensitivity:* Specification of statistical inference patterns that allow the scientist to assess the evidential relevance of observed differences.

In the following, I will use the measurement picture of content analysis. I will provide a sketch of a measurement theory for content analysis by elaborating on the differences to the error-based approach. I will proceed along the three mentioned elements of a measurement theory.

In the error-based approach, probability of measurement results is interpreted as an objective concept. On this view, probabilities are tied to systems in the world—so-called chance set-ups—instead of viewing probability primarily as a degree of evidential support or a degree of belief. The general conceptualisation of such a measurement process should

⁷⁶As before, the variances σ_1^2 and σ_2^2 must usually be estimated by using the standard variances of the samples (see footnote 74). Note also that the general description to measure two quantity values is usually conceptualised in a simplified way (Taylor 1996, 101–2): First, the measurement apparatus is assumed to determine the variance fully. Consequently, the variance is the same for the different quantity values the scientist intends to measure ($\sigma_1 = \sigma_2$). The scientist will then generate one larger sample to estimate the variance of the random error. Additionally, every subsequent measurement of a quantity value will be performed by only one measurement. In other words, the sample size will be $n = 1$, and the mean of the corresponding sample will be equal to the one measurement indication.

therefore contain a description of the corresponding chance set-up: A specification of outcomes resulting from randomly drawing a sample from a population.

In the case of the error-based approach, the outcomes of the chance set-up are random errors of a particular magnitude. As elaborated above, the variability of observed measurement indications (under conditions of repeatability) is understood as the result of measurement errors.

While I suggest sticking to an objective concept of probability in content analysis, the corresponding chance set-up cannot be tied to an interpretation of outcomes as errors. I argued that in the case of hermeneutical underdetermination, the variability of categorising coding units (under conditions of repeatability) is not necessarily an indicator of invalidity. Rather, annotators might disagree on their interpretation of coding units without making errors. In a similar vein, the error-based approach labours under the assumption that there is one true quantity value for each measurement. In contrast to that, there might be different and equally correct category values for each coding unit. In other words, the assumption of interpreting variability as the result of errors, conceptualised as deviations from a true value, does not hold in content analysis.

So, how should we understand the underlying chance mechanism in content analysis? The measurement picture of content analysis can help to answer this question. In this picture, categorising coding units is understood as a measurement procedure defined by a population of appropriately trained annotators. The actual categorisation of coding units can be understood as picking one particular annotator from this population and letting them categorise some coding units. Inasmuch as the coding instructions leave the categorisation hermeneutically underdetermined, different annotators might come to different categorisations. A probabilistic conceptualisation of coding variability can therefore be understood as the result of randomly picking an annotator from the population of trained annotators who is requested to categorise coding units.

There is another crucial difference, which demands using other probabilistic models in the case of content analysis. So far, we understood the measurement of quantity values as the measurement of numerical quantities on a continuous scale. The usual probabilistic assumption is to model the corresponding distribution of random errors by the normal distribution. In the case of content analysis, category values are seldomly conceptualised as continuous values but often as values on a nominal scale. Accordingly, we need to employ probabilistic models suitable for categorical data.

Let us, finally, come to the third element. How can the content analyst satisfy weak phenomenon sensitivity?

To answer this question, we, first, have to better understand the nature of the phenomena the content analyst seeks to investigate. In the error-based approach, the phenomena that the scientist intends to capture are true quantity values. So by asking whether a difference in measurement results corresponds to a difference in the phenomena, they ask if measurements should be taken as evidence of a difference between true quantity values. In the case of hermeneutical underdetermination, we have no analogue to the one true quantity value. There is not one correct interpretation but different interpretations that are correct. Consequently, we cannot equate one particular category value with the phenomenon of

interest. So what is then the phenomenon the content analyst intends to capture?

The following observation will help to identify the phenomena the content analyst is interested in: We expect that qualified annotators categorise correctly. But not every correct category value might have the same chance of being used by a randomly chosen annotator from the population of appropriately trained annotators. However, we expect that the conceptualised chance mechanism corresponds to a particular distribution of correct category values. Naturally, a simple suggestion is to regard this distribution as the phenomenon of interest. Accordingly, a difference in phenomena corresponds to a difference between the distributions of correct interpretations.

This suggestion is admittedly a little bit abstract. But it is a natural generalisation from the case of satisfying the uniqueness assumption to the more general case of hermeneutical underdetermination. If there is only one correct category value, the phenomena of interest might be conceptualised in terms of frequencies. In this case, it makes sense that the researcher is interested in *the* number of coding units that fall under a certain category. For instance, they might be interested in the phenomenon of hate speech in a specific medium. If there is for every coding unit one correct answer as to whether it is hate speech, then there is also one correct number of occurrences in a specific corpus. If, however, the categorisation is hermeneutically underdetermined, then there is not one correct number but many numbers. Fortunately, the suggested probabilistic conceptualisation does not lead to an anything-goes attitude: Not every possible number of occurrences is on an equal footing. Rather, every number has a specific probability attached. So instead of trying to uncover the correct frequency of a specific category value, the researcher can, for instance, determine the most probable frequency or the expected frequency—that is the probability-weighted mean—of a specific category value. More generally, by identifying the distribution of interpretations with the phenomenon of interest, every property of this distribution and every statement that is implied by it might be of interest to the researcher.

Now comes the important point to elaborate the third element: There is a major disanalogy between the error-based approach of measurements and the suggested conceptualisation in content analysis. The phenomena of interest in the former are conceptualised as true quantity values that correspond to parameters of an unknown probability distribution. In the latter case, the resulting probability distributions of correct category values are the phenomena of interest. But despite this disanalogy, we can apply the same statistical methods to decide whether observed differences—that is differences between samples—can be taken as evidence for a difference in the phenomena. In the case of the error-based approach, I elaborated that we have to use the sample data to estimate the true quantity values and their uncertainties. That turned out to be tantamount to estimating the probability distributions that give rise to the random fluctuations around the true quantity values. If in content analysis, as I suggest, the distributions themselves are the phenomena of interest, we can use the same statistical methods to estimate these distributions. Consequently, every statistical method that is apt to estimate the corresponding distributions and to determine differences between distributions will be suitable to formulate criteria of weak phenomenon sensitivity in the context of content analysis.

Significance testing is, for instance, as suitable in the context of content analysis as in the error-based approach since it is not bound to a particular interpretation of probability

distributions. Significance testing is, in particular, not committed to interpreting population means as true values and deviations from the mean as random errors.

The general inference pattern is therefore the same in both pictures: In the error-based approach, samples are used as a basis for statistical inferences to parameters of the distribution of random errors. Similarly, the content analyst should use samples of categorisations generated under conditions of repeatability—what we formerly referred to as reliability data—as a basis for statistical inferences to determine the distribution of different interpretations.

The last considerations provide but a mere sketch of how to devise a probabilistic conceptualisation of categorisation that can deal with hermeneutical underdetermination. Without a further specification of the category system, it is, in particular, impossible to suggest a specific probabilistic model. The selection of such a model, in turn, is a prerequisite to pinning down the corresponding patterns of statistical inference, which are used to formulate conditions for weak phenomenon sensitivity. I will approach these issues in Chapter 5.

Let me end this chapter with some clarifications that help to understand the main differences between the suggested approach to deal with hermeneutical underdetermination and the received view of reliability in content analysis—that is, the discussed graded notion of reliability that relies on chance-corrected reliability coefficients.

According to the received view, researchers have to assess whether their data-making process is not corrupted by annotators who do not fully adhere to their coding instructions. Content analysts have to ensure that coding results are sufficiently better than chance.

To that end, different annotators produce reliability data by independently categorising a representative subset of the text corpus of interest. The resulting reliability data is used to estimate the inter-annotator reliability. If the calculated reliability values exceed the chosen threshold, the data-making process can be regarded as reliable. In this case, researchers can trust data that is produced by this data-making process and can consider observed differences as evidence for a difference in the phenomena.

In some way, the suggested alternative view takes an opposite stance toward the role of chance: Instead of striving to exclude categorisation by chance, it intends to estimate the likelihood of correct category values. Not that it assumes that individual annotators categorise by chance, but it models the space of all admissible interpretations as probabilistically distributed. The central idea is to estimate the probability of getting a specific interpretation by randomly drawing one annotator from the population of trained annotators.

Accordingly, the role of reliability data is different in both approaches. The received view regards reliability data as evidentially relevant for judging whether annotators perform better than chance. In the alternative picture, reliability data is used to estimate the distribution of different interpretations.

The most important difference concerns the role of thresholds. Similar to the received view, the alternative picture uses thresholds to determine the significance of observed differences. As elaborated above, significance levels are used to decide whether the scientist can regard measured differences as evidence for a difference in the phenomena. However, this

relationship between thresholds and significance is more complicated than in the received view.

In the received view, the trustworthiness of the data-making process does not depend on measurements once the content analyst is able to produce reliability data whose reliability coefficient exceeds the chosen reliability threshold. Subsequent differences in measurements are regarded as significant differences without any further qualifications. In this sense, the trustworthiness of the data-making process is fixed given a specific reliability threshold and the produced reliability data.

In contrast, the scientist using the suggested alternative approach has to decide for every measured difference anew whether the observed difference is significant. In the concrete discussed case of measuring a difference on a continuous scale, the question of whether an observed difference is significant depends on the chosen significance level, the observed difference between sample means and the variance of the distribution of means. This suggests that in the case of content analysis, the significance of a difference will depend similarly on the chosen significance level, the observed difference in terms of the relevant quantity and the shape of the distribution of interpretations.

Instead of attributing reliability to the whole data-making process, the formulation of significance criteria amounts more to a conditional reliability of the data-making process. Capturing the degree of interpretation enables the content analysts to determine which differences in measured data are trustworthy—in the sense of being significant enough to be considered as evidence for a difference in the phenomena. In this way, the content analyst can satisfy the requirement of weak phenomenon sensitivity: Quantifying the degree of interpretation enables them to formulate conditions under which differences are significant.

In particular, the dependence on the distribution of interpretations is one advantage compared to the received view. I argued above (68–70) that the received view is not able to deal with hermeneutical underdetermination because the selection of threshold values for reliability coefficients should depend on the range of correct interpretations. In other words, it should depend on the expected degree of interpretation. Since the received view does not incorporate such a context dependence, I suggested that it is implicitly committed to the uniqueness assumption.

The alternative view, on the other hand, incorporates such a context dependence. Since the distribution of interpretations is a measure of the degree of interpretation, the question of whether differences are significant depends obviously on the extent of hermeneutical underdetermination. In particular, the view is not committed to the uniqueness assumption.

The connected and more general problem of the received view is its lack of providing a sufficient clarification of reliability coefficients—one that is able to justify the choice of specific reliability thresholds. The general idea is that the choice of reliability thresholds should depend on the researcher's willingness to accept a certain chance to draw false conclusions from their data. Above, I argued that content analysis offers no satisfactory elaboration of this rough idea (68–70).

The suggested alternative view, on the other hand, offers a precise probabilistic elaboration of exactly this idea. A specific significance level corresponds to a certain probability of

regarding something as significantly different (i.e., to reject the null hypothesis that there is no difference in the phenomena) even though there is no difference in the phenomena—a so-called false positive result. For instance, a significance level of 0.05 corresponds to a probability of 5% to end up with a false positive. In other words, in the long run, a significance level of 0.05 will tend to produce false positives with a rate of 5%.

I introduced the basic idea of how reliability-orientated content analysis can meet the challenge of hermeneutical underdetermination. The underlying assumptions are weaker for the data-making process than those used by the received view. But they are still quite demanding.

One central prerequisite is that different coding results must be comparable. In significance testing, *p*-values are used to decide on the significance of measurement differences. The *p*-value of a measurement *m* corresponds to the probability of observing measurement results at least as extreme as *m*. In the case of a continuous quantity value, we can use the distance of a measurement value to the sample mean. In the case of measuring a categorical variable, the content analyst has to use another way to decide on the extremity of coding results. The predominant role of counting offers different possibilities to comply with this constraint. In Chapter 5, I will introduce such a measure for the case of CAAS.

Additionally, the effort to satisfy the criteria of weak phenomenon sensitivity will be more challenging than establishing reliability according to the received view.

First, according to the received view, the reliability data, which is used to decide on the reliability of the measurement procedure, is generated on a representative sample of a text corpus, often using only two different annotators. The role of reliability data is a little bit different in the alternative approach. It is used to estimate the distribution of interpretations. To that end, the annotators that produce the reliability data should make up for a representative sample of the population of trained annotators. It is not difficult to realise that $n = 2$ will often be a too poor basis.⁷⁷ In consequence, the content analyst is required to use much more annotators as compared to what is usually done in content analysis.

Second, and more crucially, the distribution of interpretations will often depend on the particular text as much as it depends on the annotators and the coding instructions. In other words, the degree or variance of hermeneutical underdetermination might differ between texts. The case of argumentative analysis is a paradigmatic example in this connection. Texts can differ severely in their argumentative clarity and transparency. Sometimes arguments are explicitly referred to by indicator words such as ‘argument’ and ‘reason’. Such linguistic cues leave little leeway to interpret the argumentative role of the corresponding text segments. But in others texts, it might be difficult to understand the author’s intentions. In these cases, annotators might struggle to distinguish justifications from, for instance, mere explanations, illustrations or motivations. The resulting degree of interpretation can therefore differ from text to text.

This constitutes a relevant difference to measurements in natural sciences, where the random error is often assumed to depend only on the measurement apparatus. In consequence,

⁷⁷I cannot substantiate this claim here. It may suffice to note that in the error-based approach, the estimated variance of the random error can be decreased by increasing the sample size (see p. 77).

the extent of random errors—the analogue of the degree of interpretation—has to be estimated only once.⁷⁸ A similar fortunate situation prevails if we stick to the received view of content analysis. Once the trustworthiness of the data-making process is established, it needs only one annotator to fully categorise the text corpus. If, however, the distribution of interpretations might change from text to text, we need to estimate this distribution anew for every text. In other words, instead of letting many annotators categorise a smaller but representative text sample—as the received view suggests—the whole text corpus has to be categorised by many annotators.

Though the dependence on texts is a challenge, it is not an ultimate reason against my suggested approach. For every research effort, it must be somehow decided—not necessarily by the researcher—whether the expected research outcomes are worth the effort. At the very least, the suggested approach can satisfy weak phenomenon sensitivity in the face of hermeneutical underdetermination. To my mind, the only alternative is to stick to qualitative content analysis on pain of violating reliability.

Additionally, the assessment of interpretational distributions should be viewed as an interesting empirical question. Research into these distributions might reveal indicators or proxies for the degree of interpretation. The degree of interpretation will probably not vary to an arbitrary extent between different texts but will depend on other factors such as the educational background of those who produce the text, possibly the text type and additional circumstances of the text production. The assessment of the degree of interpretation and how it depends on such factors, which are external to the text, is not only interesting from the perspective of establishing weak phenomenon sensitivity but might be a worthwhile enterprise in itself.

⁷⁸See also Footnote 76.

4. HERMENEUTICAL UNDERDETERMINATION IN ARGUMENTATION ANALYSIS

The analysis of argumentation structure is a hermeneutical process and can result in different interpretations. The extent of this divergence will, of course, depend on many different aspects and can vary from context to context; it is particularly prevailing if the argumentation we want to understand and analyse is complex—that is, if it consists of many arguments and reasons which are related to each other in a justificatory way. But the degree of interpretation depends not only on the argumentation at hand but also on the definitions, rules, and conventions we introduce to analyse argumentation structure.

I will use the terminus *hermeneutical underdetermination* to denote cases where the used definitions, rules and conventions of understanding language allow different interpretations. As a result, different persons may come to different interpretations without violating these rules and conventions or misunderstanding given definitions. The set of all these admissible interpretations determines what I call the *degree of interpretation*. This specific type of intersubjective interpretational degree is, in principle, consistent with a vanishing degree of interpretation on the subjective level. Different people can come to different admissible interpretations without being aware of alternative possible interpretations. For them, their subjective interpretation might be without alternatives.⁷⁹ In this work, the relevant questions concern the variability of interpretations between different individuals. Whether this intersubjective degree of interpretation is complemented with or rooted in an (intra)subjective degree of interpretation is of minor importance for questions of reliability. In the following, I will use the terminus *degree of interpretation* and *hermeneutical underdetermination* to refer to this intersubjective degree of interpretation.

Hermeneutical underdetermination in understanding argumentation is not surprising. Argumentation analysis demands understanding the meaning of language. Since the comprehension of spoken or written utterances is a hermeneutical activity, which can lead to different interpretations, so can the analysis of argumentation structure. Whether this

⁷⁹Let me illustrate this point with a hypothetical toy example: The rules that determine the set of admissible interpretations may include a rule to choose one particular interpretation whenever the other rules do not suffice to fix one interpretation in particular. For instance, such a rule might map the interpreter's month of birth to one particular interpretation. For the individual subject, there is, in such a case, only one interpretation and no subjective leeway to choose an interpretation out of many. In this sense, there isn't a subjective degree of interpretation. But different individuals will come to different interpretations due to sketched selection rule. In this sense, there is an intersubjective degree of interpretation, or as I called it, hermeneutical underdetermination.

hermeneutical underdetermination is problematic will depend on the particular context and aim of argumentation analysis.

Hermeneutical underdetermination is particularly worrying in social-research contexts. Suppose a social researcher intends to analyse the argumentation structure in their textual data. It may be data that comprises, for instance, transcribed interviews, argumentative statements, transcribed group discussions or persuasive texts. The researcher usually analyses their data to answer a specific research question. If, however, the data allows for different interpretations, the results might also depend on the subjective interpretive decisions of the researcher. In this way, hermeneutical underdetermination can threaten the reliability of the research results.

One suggestion to accomplish high reliabilities is to narrow down the degree of interpretation. From the perspective of reliability-orientated content analysis, we should investigate the sources of hermeneutical underdetermination and find ways to minimise it. Argumentation-theoretical accounts of argument analysis are natural candidates as a means to narrow down the degree of interpretation. They seek to clarify central notions of argumentation analysis and provide analysts with methods and guiding principles to understand and evaluate argumentation.⁸⁰ The following questions are, therefore, of utter importance in this connection: How large is the degree of interpretation in argumentation analysis? Does argumentation theory offer concepts and criteria to minimise the degree of interpretation?

The degree of interpretation is a contingent property of the corpus in question. Theoretical considerations and small illustrative case studies drive the main argumentation in this chapter. While the results cannot be generalised to any corpus whatsoever, the preliminary conclusions I reach at the end are sobering. While argumentation theory systematises the analysis and evaluation of argumentation and provides more precise concepts than our ordinary language, it cannot exclude hermeneutical underdetermination. In particular, the method of reconstructing arguments does not necessarily offer an advantage in terms of narrowing down the degree of interpretation. Depending on the text and preferred analytical perspective, the analyst should expect a non-vanishing degree of interpretation. In this case, different analysts will end up with different interpretations. If the problems I discuss represent paradigmatic cases, there is no good reason to prefer the sophisticated reconstructive argumentation-theoretical accounts over non-reconstructive accounts, at least not from the perspective of maximising reliability.

From the vantage point of content analysis, this is good and bad news: If reconstructive accounts offer no advantage over non-reconstructive ones in terms of hermeneutical underdetermination, the researcher can do without the former if there are no other concerns than reliability. The coding instructions can thus be based on a slim background from argumentation theory. Since employing reconstructive methods requires extensive training, the instruction of annotators would be comparably less time-consuming. On the other hand, we would have to live with hermeneutical underdetermination in CAAS. How to deal with this hermeneutical challenge in CAAS, will be addressed in Chapter 5.

⁸⁰Since the claims I substantiate in this chapter apply to argumentation analysis in general and are not confined to CAAS in particular, I will refer to persons analysing argumentation structure as ‘analysts’. Later on, when I summarise relevant implications for CAAS (4.5), I will, similarly to the last chapter, refer to them as ‘annotators’.

In the following, I will distinguish three different types of hermeneutical underdetermination (4.1) and three different paradigms of argumentation analysis (4.2): non-reconstructive analysis, applied formal logic and informal logic. I will then move on to assess the extent of hermeneutical underdetermination for the three paradigms. Non-reconstructive analysis does not heavily rely on concepts and techniques drawn from applied argumentation theory but is, for the most part, based on our pre-theoretic understanding of argumentation. Non-reconstructive analysis will often allow for different interpretations for the prevailing lack of unique linguistic cues in texts (4.3). Applied formal logic and informal logic are reconstructive paradigms: Both introduce concepts and methods to transform ordinary-language arguments in their premise-conclusion structure.⁸¹ I will argue that these more elaborate tools of argument analysis do not necessarily lead to unique interpretations of argumentation structure (4.4). Finally, I will describe the most important implications of this chapter for devising an annotation scheme for CAAS (4.5), which you will find in Appendix A.1.

4.1 VARIETIES OF HERMENEUTICAL UNDERDETERMINATION

An analysis of argumentation structure proceeds roughly in the following way: The starting point is a text that puts forward or reproduces an argumentation. It can be a persuasive text, a transcribed speech, a forum discussion, a transcribed group discussion, or what have you. The first task is to identify argumentative units—that is, text segments that express an argumentative component.⁸² These can be claims, arguments, reasons, objections or something else. The next step is to analyse their justificatory relations to each other and possibly their internal premise-conclusion structure. Justificatory relations encompass relations that are relevant for the justification of a statement. For instance, x being an objection to y can be regarded as a justificatory relation between x and y since it can be interpreted as x justifying that y is false.

Even if different analysts utilise the same argumentation theory and agree on the relevant types of argumentative components and justificatory relations, they can come to different results when analysing the argumentation structure. If they did not violate any requirements of adequacy—that is, if they did not do something wrong in their analysis—the differences might stem from their different interpretations of the given textual data. The abstract characterisation of argumentation structure (see 2.5.3) motivates distinguishing the following types of hermeneutical underdetermination.

4.1.1 NODE AMBIGUITY

To begin with, analysts might disagree during the first step of the analysis. This ambiguity pertains to the identification of argumentative units. Since they correspond to nodes in

⁸¹I will use the term *ordinary-language argument* as referring to utterances or text segments that express arguments as competent speakers express them under normal circumstances. Usually, ordinary-language arguments are different from their reconstructions in a premise-conclusion form. However, both are expressed in natural language. Accordingly, I will use the term *natural-language argument* as an umbrella terminus for ordinary-language arguments and their reconstructions. Natural-language arguments should be distinguished from their formalisations, which are not formulated in a natural language (see below).

⁸²See Footnote 9 on page 13 for an explanation of the used terminology.

an argumentation graph, I will refer to it as *node ambiguity*. It includes two subtypes of undetermination.

First, analysts might come to different conclusions as to whether a particular text segment formulates an argumentative component, in other words, whether it is intended as justificatory relevant (*relevance ambiguity*). For instance, one analyst might interpret some text segment as intended as an argument, whereas another as something else, say, a mere illustration with no argumentative relevance. Consider the abstract example in Figure 4.1. The analyst A_1 identified three text segments (S_1 , S_2 and S_3) as argumentative units, which express three argumentative components as visualised in the adjacent argumentation graph of A_1 . If, now, a second analyst A_2 does not interpret the third text segment S_3 as an argumentative unit—they might regard it as a mere explanatory remark without being intended as a justification by the author—we have a case of node ambiguity (more specifically, a case of relevance ambiguity).

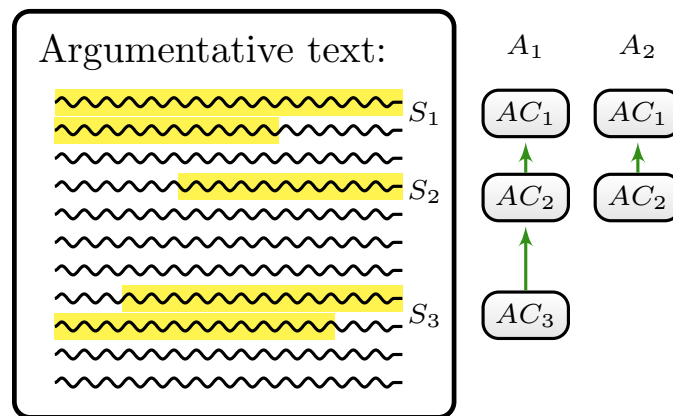


ABBILDUNG 4.1

An abstract illustration of *node ambiguity* (more specifically *relevance ambiguity*). In contrast to A_1 , the analyst A_2 does not interpret the text segment S_3 as an argumentative unit.

Second, analysts might differ in categorising the justificatory role of an argumentative unit (*node-type ambiguity*). For instance, if the used argumentation model distinguishes between central claims and reasons, analysts might differ in whether they interpret a text segment as a central claim or a reason.

If node ambiguity prevails, the resulting argumentation graphs might differ in the number of nodes or what they represent.

4.1.2 UNDERDETERMINATION OF GRANULARISATION

Even if analysts agree that a specific text segment is justificatory relevant and agree on the argumentative role of the text segment, they might disagree on the number of argumentative components expressed in a particular text segment. For instance, one analyst might interpret the text segment as expressing one reason, whereas the other identifies two. The abstract example in Figure 4.2 can be interpreted in this way. Analyst A_1 interprets the text segments S_2 and S_3 as expressing two distinct argumentative components AC_2

and AC_3 , while the second analyst (A_2) interprets these text segments as one compound argumentative component (their AC_2).

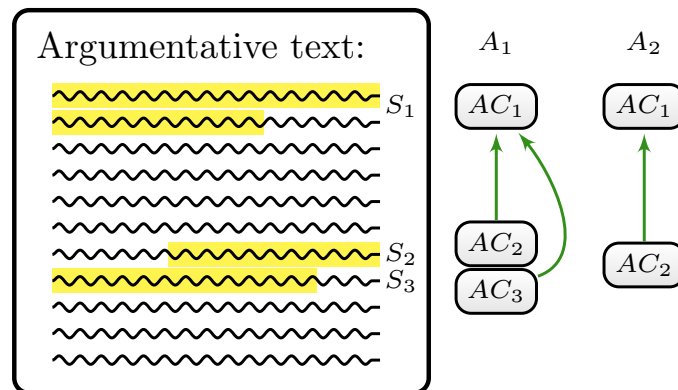


ABBILDUNG 4.2

An abstract illustration of *underdetermination of granularisation*. Analyst A_1 interprets S_2 and S_3 as two distinct argumentative components (their AC_2 and AC_3) and analyst A_2 interprets them as one compound argumentative component (their AC_2).

A related type of ambiguity concerns the combination of atomic argumentative components into more complex ones. Take, for instance, arguments, which can be thought of as having an internal structure consisting of different premises and one conclusion: Suppose that two analysts agree that each element of a set of text segments represents a reason in favour of the main claim. Both analysts might interpret these reasons as premises but aggregate them in different ways to arguments. The first analyst might interpret the whole set of reasons as one argument, whereas the other divides the set of reasons into two different arguments for the main claim.

Both types of ambiguity concern the individuation of argumentative components and their granularisation, henceforth called *underdetermination of granularisation*. We can think of this type of hermeneutical underdetermination as a question about where a claim, an argument or a reason starts and ends. It might range over the part of one sentence, over a whole sentence or many. Additionally, the text segments that express one argumentative component might be non-contiguous. They might be scattered over different parts of the text.

Underdetermination of granularisation can lead to differences in the overall amount of nodes in the argumentation graph. In this chapter, I will devote much attention to this type of ambiguity since argumentation theory provides several concepts and criteria to analyse the individuation of arguments and reasons.

4.1.3 RELATION AMBIGUITY

Finally, analysts might disagree in their analysis of the intended justificatory relations between argumentative units, what I will refer to as *relation ambiguity*. This kind of underdetermination comprises three subtypes: Analysts might either disagree on the existence of a justificatory relation between argumentative units (*edge-existence ambiguity*) or on the interpretation of the precise nature of a relation (*edge-type ambiguity*) or on the relata

of this relation (*sink-source ambiguity*). Edge-type ambiguity prevails if two different analysts identify a justificatory relation between two argumentative units but disagree on the type of that relation. One might interpret it as a support, while the other thinks it is an attack. Sink-source ambiguity concerns the sources and targets of justificatory relations and usually comes either as a sink or a source ambiguity. For instance, two analysts might interpret a specific text segment as expressing an objection but disagree on the target of the objection. In Figure 4.3, the first analyst A_1 interprets the argumentative unit S_3 to express an objection against the argumentative component expressed with S_1 . In contrast, the second analyst A_2 interprets S_3 to express an objection against the argumentative component AC_2 . Here, the analysts agree on the type and source of the relation but disagree on the target of this relation.

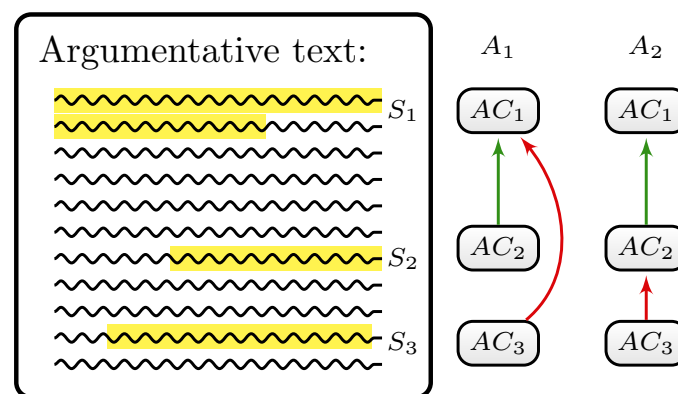


ABBILDUNG 4.3

An abstract illustration of *relation ambiguity* (more specifically, *sink ambiguity*). Analyst A_1 interprets S_3 as expressing an objection against S_1 and analyst A_2 interprets it as expressing an objection against S_2 .

4.2 VARIETIES OF ARGUMENTATION ANALYSIS

In the following, I want to compare different accounts and how they help to minimise the degree of interpretation. I distinguish between *reconstructive accounts* and *non-reconstructive analysis*. I will first discuss the latter and then go on to consider two reconstructive accounts—namely, *applied formal logic* and *informal logic*.

4.2.1 NON-RECONSTRUCTIVE ANALYSIS

Usually, a competent speaker of a language can understand the structure of argumentation in texts or oral communication if they have sufficient background knowledge of the given topic. They realise whether some utterance is meant to be a justification, a reason, an argument, an objection, or what have you. The mastery of argumentation theory is thus no prerequisite to understanding the structure of argumentation. Instead, the skill of understanding argumentation is something we learn in our everyday social environment and, of course, at school. Admittedly, some of us do better than others. But the important point is that we do not need sophisticated theories of argumentation to understand the structure of argumentations. Inasmuch as we can successfully argue with each other—in

the sense of not constantly violating established rules of conversation—we can understand the arguments and reasons of others.

I will call the analysis of ordinary-language argumentation that merely employs our pre-theoretic skills of understanding argumentation *non-reconstructive analysis*. The term *non-reconstructive* is motivated by the fact that we usually do not bother to make the internal premise-conclusion structure of arguments explicit. Under normal circumstances, we do not reconstruct arguments. In particular, we do not utilise those explications of our ordinary-language concepts introduced by argumentation-theoretic accounts to analyse argumentation based on their reconstructions. In contrast to a mere non-reconstructive analysis, the aim of an *argumentation theory* is to provide the necessary conceptual resources for a thorough reflection by introducing new concepts or adapting ordinary-language concepts to increase their precision by introducing general criteria to guide argumentation analysis and by making the aims of argumentation analysis explicit.

Before I go on to broadly characterise two different accounts of argumentation theory, let me admit that the distinction between non-reconstructive analysis and argumentation-theoretical accounts is a little bit fuzzy. It is also important to realise that the different approaches are not exclusive. Quite the contrary, without employing our ordinary-language ability to understand language, we cannot use the more sophisticated concepts of argumentation theory. Rather, an argumentation-theoretic analysis of argumentation is always based on a non-reconstructive analysis. The picture is not one of contradiction but, if anything, a picture of tension. Argumentation theory adds something to our ordinary-language abilities and adapts or corrects some concepts, which might, in the process of analysing argumentation, lead to some revisions of a preliminary non-reconstructive analysis.

4.2.2 RECONSTRUCTIVE ANALYSIS

Argumentation theory is but one analytical perspective on argumentation. I will use the term *argumentation theory* as an umbrella term to denote the spectrum of different theories, methods, and concepts that aim to analyse and evaluate ordinary-language argumentation. It encompasses different and partly conflicting accounts.⁸³

Argumentation theories that rely on reconstructing arguments are particularly suited to analyse argumentation structure. They include applied formal logic, informal logic and Pragma-Dialectics. In contrast to non-reconstructive analysis, reconstructive accounts demand reconstructing arguments in their premise-conclusion form. To better understand this central feature, I will briefly present the basic ideas of reconstructive analysis and then go on to introduce two representatives, namely applied formal logic and informal logic.

A reconstructive analysis is primarily concerned with analysing and evaluating ordinary-language argumentation in terms of rational strength in contrast to rhetorical power (Feldman 2014). Rational strength is the extent to which arguments offer good reasons to believe their conclusions. How persuasive the presentation of arguments is, depends not only on their rational strength but also on how they are presented to an audience—their

⁸³There is no agreed-upon usage of the label *argumentation theory*. Eemeren et al. (2014) understand it in the same way as I do. Sometimes it is used in a narrower sense, for instance, by Walton (2009) as a synonym for informal logic.

rhetorical power, which, for instance, includes how original, engaging, entertaining, and well-formulated the presentation is.

The analysis of rational strength can be roughly divided into three tasks:⁸⁴ The first task is to identify *argumentative units* in a given text—that is, text segments that are intended to be justificatory relevant, such as central claims, arguments, reasons and premises. This can be considered an annotation task. The analyst or annotator has to distinguish text segments that represent argumentative components from text segments with other functions, such as illustrations or mere explanations. The identification of argumentative units relies on our ability to understand ordinary-language argumentation—that is, non-reconstructive analysis—and draws on recognising linguistic cues that we find in the text.

We usually have to consider the whole text for this task. Whether a text segment represents an argumentative element is often not only determined by the text segment itself but also by the surrounding text—in the following called *cotext*. Additionally, it might be helpful to consider a broader context, including all other relevant information outside the text for its analysis. However, it is often not feasible or even impossible to consider the broader context.

Reconstructive accounts do not necessarily offer any systematic means to foster this task. Instead, they provide rules of thumb and explain the role of linguistic cues.⁸⁵

The second task is to provide a transparent, unambiguous and precisely formulated presentation of the arguments by formulating *reconstructions of the arguments*. The analyst usually starts with those components that they identified in the text and rewrites them into a list that makes the premise-conclusion form of the argument transparent:

```
(1) ... (1st premise).
(2) ... (2nd premise).
...
(n) ... (nth premise)
-- from 1-n --
(n+1) Hence, ... (conclusion).
```

Further streamlining is often necessary, which involves reformulating explicitly stated

⁸⁴The categorisation of tasks varies from author to author. Walton (2009), for instance, adds a fourth task—the task of invention, which “is to construct new arguments that can be used to prove a specific conclusion” (2). Often, authors distinguish only two core tasks, namely analysis and evaluation (e.g. Brun and Hadorn 2009; Feldman 2014). In this picture, the task of identifying argumentative units is part of the first task. In content analysis, it is common to distinguish the task of segmentation—that is, the identification of relevant text segments—from the actual annotation of text segments by some category system. This is the motivating reason for the suggested tripartite picture.

⁸⁵Comprehensive lists of relevant linguistic cues (also called *indicator words*) and their classification can be found in introductory textbooks of informal logic (see, for instance, the first Chapter in e.g. Bowell and Kemp 2010; Copi and Cohen 1990; Govier 2013; Walton 2006), applied formal logic (e.g. Brun and Hadorn 2009, chap. 8; Feldman 2014, chap. 5) and Pragma-Dialectics (e.g. Eemeren, Grootendorst, and Henkemans 2002, chap. 3; Eemeren and Henkemans 2016, chap. 3). A more thorough and theory-based treatment of linguistic cues is provided by Pragma-Dialectics (see, for instance, Eemeren and Grootendorst 1984, chap. 5).

sentences and adding elements that are not explicitly mentioned in the text. These can comprise implicit premises or even implicit conclusions.⁸⁶

This streamlining will vary between the different accounts and will be elaborated on in the subsequent sections. However, one central tenet that most accounts share is a desideratum of completeness: The analyst should reconstruct arguments in their complete form. To what completeness exactly amounts varies among the different accounts. They all share the requirement to make all necessary premises and inference rules explicit. Completeness is deemed important for the evaluation of arguments. Only if an argument reconstruction is complete, the analyst can precisely pinpoint weak points in the argument.

The third task of argument analysis is to *evaluate the rational strength* of arguments based on their reconstruction. The question of rational strength is divided into two aspects: the acceptability or plausibility of the premises and the inferential adequacy of the link between premises and conclusion. The inferential adequacy concerns the justificatory link between premises and conclusion—that is, the question of whether the premises provide a sufficient degree of support for their conclusion. Premise acceptability and inferential adequacy are, to some extent, independent of each other, as we will see shortly. There are then two main types of criticism of an argument. First, one can object to one or more premises by showing that they are implausible or unacceptable for some reason. Second, one can question the justificatory link between the premises and the conclusion. Since different argumentation theories endorse partly different views on how to understand the inferential link between premises and conclusion, it varies to what such criticism can amount.⁸⁷

While we usually start with identifying justificatory relevant text segments, there is no particular or binding order in which we have to complete the three tasks (e.g., Betz and Brun 2016, 42). In particular, the task of reconstructing and evaluating the argument are often intertwined. In the following, I will shortly describe two prominent reconstructive paradigms of argumentation theory—applied formal logic and informal logic—in more detail before I go on to assess whether they can help to reduce the degree of interpretation.⁸⁸

⁸⁶Implicit premises are sometimes also called unstated, unexpressed, missing or hidden premises. Following Hitchcock (1985), the termini ‘unstated’ and ‘unexpressed’ are biased toward an understanding of implicit premises as something the arguer had consciously in mind when devising or formulating the argument. However, this is no necessary condition for a sentence to be an implicit premise. In addition to that, Gough and Tindale (1985) argue that the term ‘missing’ suggests that the argument lacks something that the analyst has to add by themselves, which conceals the fact that implicit premises are no creations of the analyst, but are inferred based on the contextual understanding of the given utterance. Accordingly, an implicit premise is already “between the lines” in some way and is not missing in a strict sense. Implicit premises must also be distinguished from the broader genus of implicit assumptions, which comprise not only premises but also presuppositions, for instance, sentences whose truth must be presupposed for the argument to make sense. Compare Ennis (1982) and Govier ([1987] 2018) for more information on the difference between premises and presuppositions.

⁸⁷Besides these main types of argument evaluation, there are other possibilities to assess an argument critically. One could, for instance, question the conclusion itself (Walton 2009). Though it is perhaps more apt to regard the latter type of criticism not as an objection to the argument but to regard it as another argument contradicting the conclusion of the former.

⁸⁸There are many other paradigms having origins in rhetorics, linguistics and discourse theory (for an overview, see Eemeren et al. 2014). I will confine the discussion of dealing with hermeneutic underdetermination to applied formal logic and informal logic and will not specifically elaborate on a third very prominent reconstructive account, namely Pragma-Dialectics as it was developed in Eemeren and Grootendorst (1984), Eemeren and Grootendorst (1992) and Eemeren and Grootendorst (2004). The reason is simple.

4.2.3 APPLIED FORMAL LOGIC

Applied formal logic (in short, *applied logic*) is an account that draws primarily on formal logic to reconstruct ordinary-language arguments.⁸⁹ This account is tightly connected with the doctrine of reconstructive deductivism. According to this view, most, or for some advocates even all ordinary-language arguments have to be reconstructed as deductively valid arguments. Deductive validity is but one possibility to spell out the notion of inferential adequacy. Since deductive validity is a comparably simple notion, the applied logician will try to reconstruct arguments as deductively valid as long as there are no profound reasons not to do so.⁹⁰ This is usually achieved by adding implicit premises and conclusions and by reformulating explicitly mentioned sentences.

Deductive validity refers to what we mean by saying that a conclusion is the logical consequence of a set of premises, or, in other words, that the conclusion logically follows from the premises. But what exactly does it mean that x is a logical consequence of y ? One way of characterising the notion of logical consequence is to connect it to the interrelationship between the truth values of the premises and the conclusion. We will say that an argument is called *deductively valid* (in short: *valid*) if the conclusion cannot be false if the premises are true. For valid arguments, the truth of the premises guarantees the truth of its conclusion.⁹¹

Since applied logicians seek to reconstruct arguments as deductively valid, one main challenge is providing criteria that help decide whether a reconstruction is deductively valid. To that end, the applied logician draws on formal logic. Formal logic introduces formal languages, rules to construct well-formed sentences in that language and conditions for the formal validity of arguments (or inferences) consisting of sentences in that formal language. For instance, the zero-order logic uses constants such as ' p ' and ' q ' for atomic sentences, which can be combined with certain connectives such as ' \wedge ' to form more complicated sentences (' $p \wedge q$ '). These formal languages are then equipped with a semantics

Pragma-Dialectics is a multidisciplinary approach based on insights from theories about communication and interaction, such as speech-act theory and inspired by approaches of formal dialectics. Its distinguishing feature, as compared to other accounts, is its focus on procedural aspects of argumentation. Since I focus more on the structural aspects of argumentation in this work, there is no particular benefit to including a separate discussion of Pragma-Dialectics here (see also 2.5.3). Additionally, since Pragma-Dialectics shares different features of both applied formal logic and informal logic when it comes to argument reconstruction, it will perform similarly with respect to hermeneutical underdetermination. Admittedly, this is a mere hypothesis that is beyond the scope of this work. Whether Pragma-Dialectics offers some additional means for content analysis will depend on the type of text to be analysed. Pragma-Dialectics might be particularly useful to analyse texts that represent dialogical situations, for instance, transcripts of oral discussions. Whether the dialogical perspective provides any benefit to analyse non-interactive texts, for instance, a persuasive text of a particular author, is questionable (compare J. Anthony Blair 2015).

⁸⁹Whereas the term *informal logic* is widespread, the term *applied formal logic* and its abbreviated form *applied logic* is not established terminology. Betz (2010) uses the same terminology. Interestingly, informal logic started under the label *applied logic* (Ralph H. Johnson 2006).

⁹⁰Advocates of applied formal logic include Betz (2010), Bucher ([1987] 2019), Copi, Cohen, and McMahon (2014), Feldman (2014), Luckhardt and Bechtel (1994), Salmon (1963) and Tetens (2004).

⁹¹Although the validity of an argument has something to do with the truth values of its sentences, we do not have to judge the truth of either the premises or the conclusion to decide whether an argument is valid. The property of deductive validity is only a conditional statement: If the premises are true, the conclusion must also be true. For an introduction to the notion of deductive validity, see Shapiro (2006) and Sundholm (2006).

(i.e., a specification of how truth values of atomic sentences are mapped to truth values of more complicated sentences), which allows one to determine which inferences are truth-preserving in that formal language. For instance, the semantic of the connective ‘ \wedge ’ is introduced by defining that the formal sentence ‘ $p \wedge q$ ’ is true if and only if both the truth values of both ‘ p ’ and ‘ q ’ are true. Accordingly, an inference from ‘ p ’ and ‘ q ’ to ‘ $p \wedge q$ ’ is deductively valid in that formal language.

So far, a formal language, its introduced semantics and validity notion do not have to do anything with our natural language. It is, therefore, also better to speak of formal validity (in short, f-validity) in the case of a formal language. But the aim of formal logic is not to introduce any formal language whatsoever but a formal language whose f-validity concept mimics to a preferably large extent the validity concept of natural-language arguments.

Consider the following argument, which is surely deductively valid:

- (1) Cassandra is wearing a green sweater.
- (2) Cassandra has short hair.
- from 1 & 2 --
- (3) Hence, Cassandra is wearing a green sweater, and Cassandra has short hair.

The validity of this argument is explained by how we use the term ‘*and*’ in our language. In contexts where the *and* is not intended to express a specific temporal order (as in “*Peter came to the party, and then came Clara.*”), two true sentences that are combined to a conjunctive form by using the word ‘*and*’ always result in a true sentence (i.e., are truth-preserving). In this way, the semantics of the conjunctive ‘ \wedge ’ in the formal language mimics the semantics of *and* in our natural language. The form of the valid inference in the natural-language argument mirrors the form of the f-valid inference from ‘ p ’ and ‘ q ’ to ‘ $p \wedge q$ ’.

The basic idea of using a formal language to decide on the validity of natural-language arguments is to generalise this simple example: A natural-language argument is said to be deductively valid if its logical form mirrors the form of an f-valid inference in a suitable formal language. Identifying a sentence’s logical form and deciding whether its form mirrors the form of a sentence in the formal language is usually conceptualised as a task to find an adequate formalisation of the natural-language sentence—that is, an adequate translation of the natural-language sentence into the formal language. If the sentences of an inference can be formalised into a f-valid inference in the formal language, the natural-language argument is considered valid.⁹²

Consequently, a significant effort in reconstructing arguments consists of analysing the logico-semantic form of natural-language sentences by formalising them to judge the

⁹²Whether this basic idea of providing an explication of validity succeeds hinges on what is considered an adequate formalisation and a suitable formal language. In very rough terms, the idea is that both the rules of formalisation and the chosen formal language should result in validity verdicts that cohere best with our intuitions about validity. For an elaboration of this idea, see Brun (2014). For challenges and problems that pertain to the task of formalisation, see Brun (2023).

validity of arguments. This process requires experience and time.⁹³

4.2.4 INFORMAL LOGIC

Informal logic is “a branch of logic whose task is to develop non-formal standards, criteria, procedures for the analysis, interpretation, evaluation, criticism and construction of argumentation in everyday discourse” (Johnson and Blair, 1977, as cited in R. H. Johnson and Blair 2002, 358). Though sharing similar aims with applied formal logic, the informal logician is sceptical about the benefits of using formal methods to assess natural-language arguments. However, the rejection of formal methods should best be understood as a contingent and partially obsolete feature of informal logic. The sceptical attitude towards formal methods is best explained by a lack of formal methods for non-deductive reasoning during the beginnings of informal logic. In the 1960s, formal logic was largely confined to formal deductive logic. By now, there is a variety of formal accounts to model non-deductive reasoning,⁹⁴ which attracted the attention of informal logicians.⁹⁵

The distinguishing feature of informal logic is better explained by its stance towards the role of deductivism in analysing and evaluating ordinary-language arguments. Deductive validity is admittedly the strongest possible inferential connection between premises and the conclusion of an argument. The question is whether all or most ordinary-language arguments have to be judged against the standard of deductive validity. According to informal logic, other possibilities to spell out the concept of inferential adequacy are more relevant for analysing ordinary-language argumentation. On this view, a weaker inferential link between premises and conclusion can still provide sufficient support for the conclusion. Arguments that are not valid but satisfy the desideratum of inferential adequacy are called non-deductive arguments.⁹⁶ In a non-deductive argument, the conclusion can still be false if the premises are true as opposed to valid arguments. On the informal logicians’ view, most ordinary-language arguments are not to be understood as implicitly deductively valid but as non-deductive.

If someone accepts the premises of a deductively valid argument, they are rationally required to accept the conclusion. It would be irrational not to accept the conclusion since the conclusion must be true if the premises are true. This rationality constraint does not depend on what else this person accepts as true. If they gather additional information that does not invalidate the premises of the argument, they are still rationally required to accept the conclusion because a valid argument is valid independent of the truth values of other sentences. We could even add these other sentences to the argument without changing

⁹³For a thorough discussion of formalising natural-language arguments, see Brun (2003) and Reinmuth (2014). An alternative strategy that avoids formalisation is to advance a semantic explication of validity that could be more directly applied to natural-language arguments. See, for instance, Hitchcock (1985) and Hitchcock (2011) for such a suggestion.

⁹⁴For an overview, see Prakken and Vreeswijk (2002).

⁹⁵E.g., Walton and Gordon (2015).

⁹⁶This is a mere stipulative definition and does not answer the question of what types of non-deductive arguments exist. Sometimes, the term ‘inductive’ is taken to be equivalent to ‘non-deductive’ in this connection. Often, however, inductive arguments are understood in a narrower sense. They are non-deductive arguments whose premises make the conclusion probable. Other often mentioned non-deductive argument types include presumptive, conductive, and abductive arguments (J. Anthony Blair 2015; Groarke 2019).

this situation. This property is called *monotonicity*: Adding additional premises to a valid argument cannot turn it into an invalid argument.

Non-deductive arguments are different: They are defeasible. Whether it is rational for a person to accept the conclusion of a non-deductive argument depends on the additional information they have. Paradigmatic examples are statistical arguments or the use of non-strict generalisations such as “*Birds can fly*.” Such a sentence cannot be interpreted as a universal generalisation of the form “*All birds can fly*.” but as a non-strict generalisation—a generalisation that allows for exceptions.⁹⁷ It should be rather understood as something like “*Under normal circumstances, birds can fly*.” If you know that Tweety is a bird and know nothing more about Tweety, it is rational to accept that Tweety can fly. But this inference is defeasible and provisional. Getting to know that Tweety is a penguin is perfectly compatible with Tweety being a bird. However, this additional information will invalidate the former implication that Tweety can fly. Informal logic rests on the assumption that most ordinary-language arguments are, in this sense, defeasible.⁹⁸ On such a view, applied formal logic seems to distort ordinary-language arguments since it demands reinterpreting non-strict generalisations as universal generalisations to arrive at valid reconstructions.

The informal logician does not only favour another conceptualisation of inferential adequacy but is generally sceptical of the deductivist’s way of evaluating ordinary-language arguments. Both accounts share the view that the rational strength of an argument depends on two things: the adequacy of the premises and the quality of the inferential link. But instead of judging arguments by their soundness—that is, the truth of their premises and validity of the inferential link—the informal logician employs the alternative concept of informal cogency, which can be spelt out in different ways. The most prominent accounts include the ARS criteria, fallacy theory, defeasible inference rules and argument schemes (J. Anthony Blair 2015). The ARS criteria and argument schemes are the most relevant for the upcoming sections of this chapter.

Ralph H. Johnson and Blair (1977) suggested an elaboration of informal cogency in terms of their now prominent three ARS criteria, which invoke the concepts *acceptability*, *relevance* and *sufficiency*. Similar to applied logic, the premises of an argument have to be evaluated by their acceptability.⁹⁹ The concepts of relevance and sufficiency serve to judge the inferential link. On this view, good arguments have premises that are relevant to their conclusion, which means that they provide at least some support for the conclusion.

⁹⁷Some authors prefer the terminus *defeasible generalisation* (e.g. Walton 2006). However, that might provoke some misunderstandings. First of all, both universal generalisations and non-strict generalisations are defeasible—in the sense that they can be shown to be false by appropriate evidence. The fact that the generalisation is defeasible is, therefore, not the relevant distinguishing feature. What is more, the discussed allowed exceptions of non-strict generalisations are not taken as evidence against the general statement. Tweety being a penguin—and hence a bird—is not a defeating instance of the statement that birds can (normally) fly. Therefore, the counterexample does not defeat the general statement but rather the inference.

⁹⁸Walton (2009), for instance, writes that “many of the most common arguments in legal reasoning and everyday conversational argumentation that are of special interest to argumentation theorists fall into this defeasible arguments class” (6).

⁹⁹Often, the formal logician serves as a foil, who, according to some received description, invokes truth as the relevant condition of premise adequacy. For different reasons, the informal logician regards the truth of premises as neither sufficient nor necessary for the adequacy of the premise (see Groarke 2019 for an overview).

If the support provided by all the premises is enough to establish the plausibility or acceptability of the conclusion, the premises count as sufficient. To what exactly relevance amounts remains often unexplicated or is only explained with the help of similarly intuitive concepts.¹⁰⁰

An alternative or at least complementary picture to these criteria is connected to the use of argument schemes. Argument schemes describe abstract forms of recurring argument patterns that are used in everyday argumentation. They can help reconstruct and evaluate arguments. Walton's approach to argument schemes combines them with critical questions, which are used to scrutinise the acceptability and relevance of the premises.¹⁰¹ To reconstruct an argument, the analyst has, first, to recognise a formulated argument as an instance of a particular argument scheme and, then, substitute the placeholders in the argument scheme with the specifics of the argument at hand. The resulting argument reconstruction is to be evaluated by raising critical questions associated with that argument scheme and questioning the premises' acceptability. If an argument conforms to the pattern of an argument scheme and both the acceptability and critical questions can be given a positive answer, the argument may be regarded as a good or conclusive argument.

4.3 UNDERDETERMINATION IN NON-RECONSTRUCTIVE ANALYSIS

Non-reconstructive analysis of argumentation structure draws on our competence as natural-language speakers to understand argumentation. Under normal circumstances, we can correctly identify utterances as claims, reasons and arguments and correctly understand their justificatory relations to each other. In contrast to reconstructive accounts, identified arguments will not be reconstructed in their premise-conclusion form. The analyst will use the cotext and rely on linguistic cues to understand the argumentation structure. The question is whether non-reconstructive analysis allows for different interpretations of argumentation structure.

Admittedly, an arguer may be perfectly unambiguous about their claims, arguments and reasons. It is possible to make the argumentation structure maximally explicit by including unambiguous linguistic cues as signposts along the following lines:

'My main claim is that ... My first reason for this claim is that ... This can be illustrated by ..., which I do not intend as an argument or reason, but simply as an illustration. A possible objection to my first reason is that ...'

This simple example suggests that being maximally explicit may be tedious and overly explicit.

¹⁰⁰Govier (2013), for instance, distinguishes negative and positive relevance and explains that a "statement A is **positively relevant** to another statement B if and only if the truth of A counts in favour of the truth of B. This means that A provides some evidence for B, or some reason to believe that B is true" (149). Freeman (2001) explicates the relevance notion in probabilistic terms: "A statement A is positively relevant to a statement B just in case A's being true increases, however slightly, the likelihood of B's being true. A is negatively relevant to B just in case A's being true increases the likelihood that B is false" (413). However, he does not explain further how to gauge these probabilities. In the subsequent sections, I will discuss the notion of relevance in greater detail.

¹⁰¹See, e.g., Walton, Reed, and Macagno (2008) and Walton (1996b).

Fortunately, our natural language provides means to communicate more effectively by utilising communication rules and charitable rules of interpretations. The basic idea goes as follows:¹⁰² We can assume that people are generally interested in being understood by their audience. In consequence, we can assume that they adhere to rules of communication that facilitate efficient communication to the best of their knowledge and ability. The audience, on the other hand, can then interpret what was uttered under the charitable presupposition that the arguer did not violate any communication rules. For instance, one such rule demands that an utterance must be appropriately connected to preceding utterances (Eemeren and Grootendorst 1992). Hence, there is no need to formulate whether an utterance is connected to preceding utterances explicitly. In understanding an arguer, we can simply presuppose that it does.

Therefore, an arguer can omit linguistic cues, without necessarily hampering the understanding of argumentation structure. However, this kind of implicit communication can go wrong. In deciding whether to provide explicit linguistic cues, the arguer has to choose their words carefully by using assumptions about what the ones reading the text need to understand the argumentation structure. They must anticipate whether their hints are sufficient for the audience to understand the argumentation. The arguer may underestimate what level of explicitness is needed for that. Since extensive explicitness conflicts with other desiderata of text production, such as brevity of what is written and also aesthetic aspects of the writing, the arguer might provide less explicitness than is needed to understand the argumentation structure unambiguously. Depending on the argumentative text at hand, we can, therefore, expect to be confronted with a considerable degree of interpretation in analysing argumentation structure.

4.3.1 NODE AMBIGUITY

These rather general considerations about the presence or absence of linguistic cues are relevant to all mentioned ambiguity types. Let us now more thoroughly consider the case of node ambiguity—that is, the case of hermeneutical underdetermination that pertains to identifying argumentative units and specifying their precise justificatory role. In CAAS, argumentative units usually express claims, reasons, objections or whole arguments. So, to what exactly do these termini refer?

The basic purpose of offering arguments is to justify a claim. To put forward an argument is a speech act with a specific intention, namely to provide reasons—usually called premises of the argument—for the truth of some claim. An argument is then a complex premise-conclusion nexus: a set of premises intended to provide sufficient support for its conclusion. Often, the arguer will try to convince their audience of some statement by presenting an argument in favour of it. However, an arguer might have other intentions. They might, for instance, present an argument to subsequently show its weakness by putting forward counterarguments. As a consequence, it is not necessarily the intention of the person who presents the argument to argue in favour of the argument's conclusion. We should, therefore, better say that an argument is a set of statements that are *presented* as having the function to justify another statement. Therefore, the analyst who seeks to identify

¹⁰²For a more thorough exposition of the role of communication rules in argumentation, see Eemeren and Grootendorst (1992), esp. Chapters 3–5.

argumentative units in texts should search for text segments presented by the author as justifications for something else.

This functional conceptualisation characterises an argument by virtue of its intended function. The point to bear in mind is that an argument is an argument independent of how it measures up to this function. It is the intended function that is relevant.¹⁰³ This characterisation differs to some extent from our everyday usage of the term. Sometimes, we object to a statement by saying '*This is not at all an argument!*' although the statement had been put forward to justify a claim. By this exclamation, we express that the alleged argument does not succeed in justifying its conclusion. The suggested understanding of the term *argument* is, therefore, to some extent weaker than the ordinary-language understanding of the notion.¹⁰⁴

But what is the relation between arguments and reasons? A reason for a claim can be regarded as a premise in a corresponding argument with the claim as its conclusion. Often, we can think of one reason for some claim as expressing one argument. However, sometimes, one argument can comprise different reasons for a claim because an argument aims at establishing sufficient support for its conclusion. Since we can regard something as a reason even if it does not provide sufficient support for a claim, we might have to aggregate more than one reason into one single argument. There is then not necessarily a one-to-one correspondence between reasons and arguments.

Now, how do we identify argumentative units? The audience will often need linguistic cues for discerning a text segment as justificatory relevant because argumentative texts contain more than just arguments and reasons. An author may motivate arguments, explain their context, provide illustrations and examples, and even tell an interesting surrounding story, to name just a few other roles a text segment may have.

However, linguistic cues can be ambiguous; depending on the cotext, they may signify different roles (Eemeren and Grootendorst 1984, 113; Hovy 1995). The reader might stumble over linguistic cues without knowing with certainty what they indicate. For instance, in introductory textbooks of argumentation theory, much effort is devoted to explaining the difference between arguments and justifying reasons, on the one hand, and explanations, on the other hand (see, for instance, Howell and Kemp 2010, 20–22; Govier 2013, 13–20; Ralph H. Johnson and Blair 1994, 18–19; Copi, Cohen, and McMahon 2014, 18–20).

Consider the following example:

Many believe GGE [germline gene editing] research ought not to be permitted

¹⁰³This is in line with the usual understanding of the notion in informal logic (see, for instance, Govier 2013, 8–9), in *Pragma-Dialectics* (see, for instance, Eemeren 2001, 19) and in applied formal logic alike.

¹⁰⁴This weak conceptualisation is helpful for the argumentation theorist in general and the content analyst in particular (see also 2.5.4). The argumentation theorist wants to distinguish between the analysis and the evaluation of argumentation. To that end, they need a notion of argument that does not presuppose the evaluation. Similarly, the content analyst may either be content with analysing the argumentation structure itself without evaluating the argument or, at least, they should try to distinguish these aspects as well. The reason is that the evaluation of arguments will be much more contentious than the mere analysis of the argumentation structure. We should, therefore, expect less reliability in an evaluative categorisation of argumentation than in a categorisation that merely pertains to the structure of argumentation.

or funded because it is unsafe. (Gyngell, Douglas, and Savulescu 2017, 504)

A sentence of the form ‘*x because y*’ often indicates a justificatory reason relation—with *y* being a justification for *x*. It can, however, also express a causal explanation. Whereas arguments and justifying reasons provide justifications, explanations are formulated to provide understanding (Govier 2013, 15). For instance, a *causal* explanation provides an explanation by describing a cause. The difference between justification and explanation is crucial to the task of identifying arguments and reasons. The purpose of providing an explanation for *x* is different from the purpose of providing a justification for *x*. If we seek an explanation for *x*, we usually do not question the truth of *x*. Instead, the truth is already established—or at least not the point in question—and we ask for a cause that could explain *x*. But if we justify some claim, we intend to establish the truth of that claim by providing reasons for it.

The question is then whether the authors provide a causal explanation of why many people *believe* that germline gene editing research should be prohibited or whether they invoke a justification for the prohibition of such research. The surrounding cotext provides clarification:

In the above section, we argued that there is a significant case in favour of pursuing some types of GGE. In this section, we analyse some of the arguments that have been offered against the pursuit of GGE to determine whether they undermine this case. [...] Many believe GGE [germline gene editing] research ought not to be permitted or funded because it is unsafe. (Gyngell, Douglas, and Savulescu 2017, 504)

The authors explicitly state that they discuss arguments that undermine GGE. They describe these arguments to refute them later in the text. Therefore, the sentence in question must be interpreted as an objection to GGE and not as a causal statement.

Since the authors discuss arguments to which they do not necessarily subscribe, they have to indicate their stance toward these arguments. The many-believe construction is formulated precisely for this purpose. It helps the audience to understand that this argument from risks is being put forward by others but, at this point, not by the authors. Therefore, the argument’s conclusion is not about what many *believe*, but about the prohibition of GGE research.¹⁰⁵

Hence, the following reconstruction of the argument:

<First reconstruction>

- (1) Germline gene editing research is unsafe.
- justified by 1 --
- (2) Germline gene editing research ought not to be permitted or funded.

¹⁰⁵Often, as in this example, the literal meaning of an utterance is different from what is meant with the utterance. Under normal circumstances, we have no difficulty disambiguating the meaning in such cases. For an analytical perspective of such cases in terms of implicit and indirect speech acts in the context of argumentation, see Eemeren and Grootendorst (1992), Chapter 5.

The example illustrates that the same linguistic cue can designate different functions. The term ‘*because*’ can be used to indicate an explanation or a justification. The same applies to other linguistic cues such as ‘*since*’, ‘*therefore*’ and ‘*for*’ (Copi, Cohen, and McMahon 2014). As a consequence, there is no comprehensive list of linguistic indicators that uniquely designate a specific role. For a correct interpretation of the ‘*because*’ phrase, we had to consider the cotext of that sentence that helped us understand the authors’ intentions. Thus, the identification and categorisation of justificatory relevant text segments hinges on the surrounding text elements and is not independent of the author’s intentions.¹⁰⁶

There are other types of explanations than causal explanations that can also be misunderstood as justifications. Elaborative explanations, for instance, are used to clarify a point or explain its meaning. This is often achieved by providing examples and specifications, which can, however, also be used to justify a statement.¹⁰⁷ Analogies can similarly cause interpretational leeway. An author can use an analogy to illustrate and explain an abstract idea or to justify something (Copi, Cohen, and McMahon 2014, 488–89).

4.3.2 RELATION AMBIGUITY

Identifying and categorising justificatory relations between argumentative components can be an additional source of interpretational differences between analysts. What I call relation ambiguities can arise during the process of identifying a justificatory relation (*edge-existence ambiguity*), during the process of categorising an identified justificatory relation (*edge-type ambiguity*) and during the process of identifying the relata of the relation (*sink-source ambiguity*).

Before we can estimate this degree of interpretation, it is useful to get a better grasp of the notion of a justificatory relation. In the preceding sections, we already got familiar with justificatory relations. Both the concept of a reason and the concept of an argument should be understood as a relation. Although we often speak of something being a reason simpliciter, a reason is always a reason for something else. The relata of this reason-relation are statements or their corresponding propositions. A canonical form of a reason should be, therefore, expressed as ‘*p is a reason for q*’ with ‘*p*’ and ‘*q*’ being placeholders for statements. Similarly, I characterised an argument as a set of statements presented as a justification for another statement. Here, the justificatory relation is one between a set of statements—the premises of the argument—and the argument’s conclusion. In a similar vein, an objection is always an objection to something else.

All these relations share that one of their relata is used to justify or cast doubt on the other relatum. A preliminary description of what a justificatory relation is characterises it in terms of how the truth, plausibility or acceptability of its relata are connected. On this view, a relation is a justificatory relation in virtue of its relevance to determining the truth, plausibility or acceptability of its relata. Take, for instance, the relation of contradiction: Two statements are contradictory if they cannot be both true. If one of them is true, the

¹⁰⁶Further elaborations and practical hints about identifying argumentative components in texts can be found in Feldman (2014), Chapter 5 and Fisher (2004), Chapter 2.

¹⁰⁷Scholman and Demberg (2017) provide empirical findings showing that examples and specifications are used in these different ways and that annotators often disagree on their categorisation as elaboration and justification.

other must be false—that is, the truth of one logically implies the falsity of the other. A statement ‘*p*’ and its flat negation ‘*It is not true that p.*’ form the most simple pair of contradictory statements; they cannot both be true simultaneously.¹⁰⁸ Either ‘*p*’ is false or its negation. Knowing that one of them is true is, therefore, obviously relevant to the truth of the other. In this sense, the relation of contradiction seems to be a justificatory relation.

But something is not quite right with the given preliminary description of justificatory relation. At least, the formulation is misleading and has plausible alternatives. The problem is that we might confuse questions of how truth values are connected with questions of what the arguer thinks about how truth values are connected. Remember that I introduced arguments as being constituted by what the arguer *intended* as a justification. An argument is an argument in virtue of its premises being *presented* as providing support for the conclusion—quite independent of whether the premises, in fact, support the conclusion. The latter aspect belongs to the task of evaluating arguments, which is different from the task of understanding an argumentation or analysing its structure. I suggest that we should be consistent and understand all justificatory relations as grounded in whether something is *presented as being relevant* for the truth, plausibility or acceptability of something else. Consider, for instance, the following text snippet:

Peter claims that the left threatens the freedom of science. But this cannot be true. The value of freedom is a focal cornerstone of the left.

The second sentence explicitly lays out how the author understands the role of the third sentence: It is presented as showing the first sentence to be false. To understand the structure of the given argumentation, we do not have to judge whether the third sentence has the claimed bearing on the truth of the first sentence. It is, indeed, doubtful whether both sentences contradict each other. Admittedly, there is some tension between both sentences. But the actual relevance of the third sentence for the truth, the plausibility or acceptability of the first sentence, is a matter of further discussion and evaluation. The point I want to stress here is that this question does not need to concern the analyst. We can confine the analysis of argumentation structure to questions concerning whether the arguer presented some statements as being relevant for the truth, plausibility or acceptability of some other statements.¹⁰⁹

I, therefore, prefer the following characterisation of justificatory relations: A relation is a justificatory relation only if it is *presented as connecting* the truth, plausibility or acceptability of its relata.¹¹⁰ On this view, it is more about the intended function according to the interlocutor and not about whether the function is satisfied.

This characterisation is admittedly too general. If we want to categorise justificatory relati-

¹⁰⁸However, there are also proponents of dialetheism. According to this view, there are true contradictions—that is, true sentences of the form ‘*p and it is not true that p.*’ This philosophical view is given a precise formal language in systems of paraconsistent logic (for an overview, see Priest, Berto, and Weber 2018).

¹⁰⁹My stance is clear on this issue: Argumentation structure should be in the first instance about what the arguer presents as being justificatory relevant. However, this stance has no bearing on this chapter’s main claims or this work in general. That is why I use rather hedged formulations here.

¹¹⁰This is at best a partial explication of the notion. It does not formulate a sufficient condition since an interlocutor might present truth-value connections for reasons different from describing justifications. They might, for instance, try to explain the meaning of a sentence by providing another sentence that is, in their view, logically equivalent.

ons, we have to be more specific about the functions an arguer has in mind. Justificatory relations will be about convincing someone and about justifying something or about presenting something as a justification. There are two fundamental justificatory relations, which are called attack and support. We will introduce the support relation by saying that *p supports q* if the truth, plausibility or acceptability of *p* is presented as providing support for the truth, plausibility or acceptability of *q*. In this sense, arguments and reasons express a support relation. Similarly, we will say that *p attacks q* if the truth, plausibility or acceptability of *p* is presented as providing support for the truth, plausibility or acceptability of *p* being false. In this sense, an objection expresses an attack relation.

Two further questions remain to characterise justificatory relations. The first question is whether there are justificatory relations that cannot be categorised as support or attack relations. The second question is whether it is helpful to introduce subcategories of both relations.

Instead of further manoeuvring in the abstract, let us consider a concrete example of an argumentation and try to identify justificatory relations:

[1] Governments should employ measures of affirmative action (AA) to help minorities that are under-represented in higher education, business and politics. [2] There are, of course, objections. [3] Some claim that affirmative action is unnecessary because countries evolve organically towards equality as long as countries accord protection from discrimination in the law to women and minorities. [4] Under this assumption people from these communities will eventually begin to succeed and this then cascades through society. [5] Others argue that affirmative action can even damage the development of a country; [6] for example, South Africa is a country that needs to develop economically for the well-being of all its people (black unemployment is very high), but its priority is AA which means that it is putting less qualified people into jobs and thereby jeopardising development. [7] This helps neither majorities nor minorities. [8] I will argue that these objections are not convincing and do not outweigh the arguments in favour of affirmative action. [9] First, under-representation is bad in practice as well as in principle. [10] Countries are damaged by not having access to the brightest talent from the largest pool. [11] They also benefit from a greater level of diversity; [12] e.g. representation in parliaments is improved by having more women. [13] Second, ... (Newman, Sather, and Woolgar 2013, 111–12)¹¹¹

The first sentence in this example expresses the main claim (*C* in Figure 4.4) in that it is justified and attacked by different considerations. According to this main claim, measures of affirmative action should be employed to reach an aim—namely, to help under-represented minorities. We could leave it at that. But to my mind, there is an additional admissible interpretation of the first sentence. Since the arguer is explicitly in favour of the main claim (see sentence 8), the first sentence might not only express a means-end relation—as

¹¹¹Most of the formulations are direct quotes from Newman, Sather, and Woolgar (2013), who present these arguments and other arguments in form of a pro and con list. I took the liberty of not indicating what I've changed to enhance the readability of the text snippet. I added linguistic cues, numbers of sentences and slightly changed the main claim.

indicated by the linguistic cue ‘to’—and that the means should be used but additionally, a reason or an argument (R_1) in favour of the main claim. According to this second interpretation, then, the first sentence formulates a practical syllogism: The measure should be employed because the intended end is important. The arguer implicitly presumes that the current under-representation of minorities is a problem that needs fixing and identifies affirmative action as a means to end under-representation.

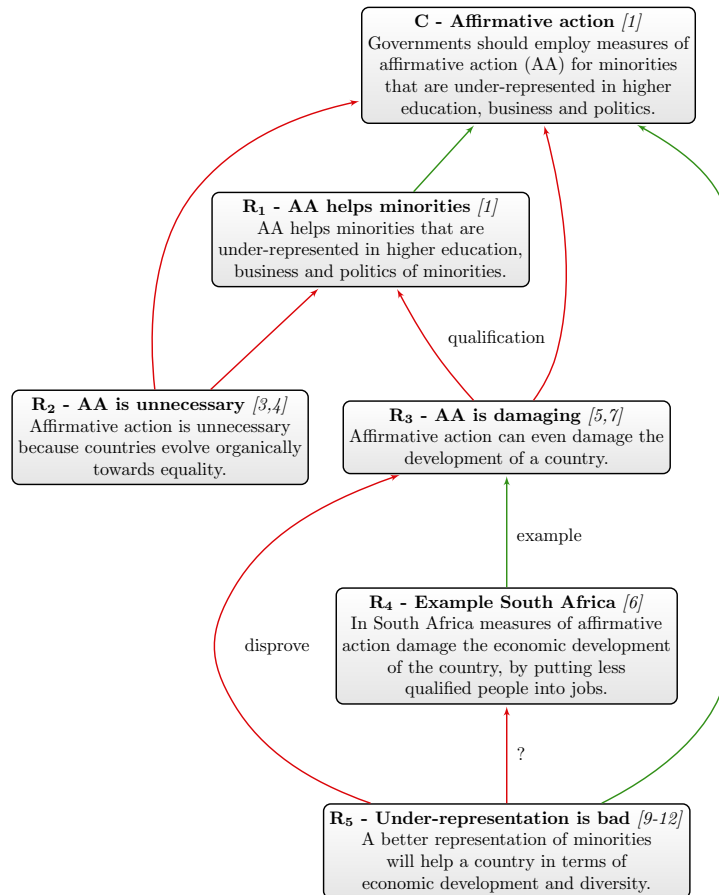


ABBILDUNG 4.4

Argumentation structure of the example. Green edges represent support relations; red edges represent attack relations.

... [2] There are, of course, objections. [3] Some claim that affirmative action is unnecessary because countries evolve organically towards equality as long as countries accord protection from discrimination in the law to women and minorities. [4] Under this assumption people from these communities will eventually begin to succeed and this then cascades through society. (R_2) ...

The second sentence indicates that the author will present objections—that is, attacking considerations. While that much is clear, this hint is silent on the specific target of these objections. Accordingly, the objection formulated in the third and fourth sentence (R_2) can be interpreted as attacking the main claim or the reason R_1 . The latter interpretation might be more plausible but hinges on interpreting the first sentence as a reason. If the analyst decides against such an interpretation, they would have to interpret the objection as attacking the main claim. These considerations show that there are different interpretations as to where the attacking relation from R_2 points (*sink ambiguity*).

... [5] Others argue that affirmative action can even damage the development of a country; [6] for example, South Africa is a country that needs to develop economically for the well-being of all its people (black unemployment is very high), but its priority is AA which means that it is putting less qualified people into jobs and thereby jeopardising development. [7] This helps neither majorities nor minorities. (R_3 and R_4) ...

The objection expressed in sentences 5–7 (R_3) intends to undermine that affirmative action will help minorities (sentence 7). The conclusion of this objection is an empirical statement stating that affirmative action will not help minorities. It is, therefore, plausible that the target of this objection is an argument that assumes that affirmative will help minorities. This is exactly the argumentation of the suggested second interpretation of the first sentence as a reason (R_1), which motivates the attack relation between R_3 and R_1 .

Again, it should be admissible to alternatively interpret R_3 as an attack against the main claim since it is admissible to interpret the first sentence only as a claim and not as expressing a reason. Hence, R_3 can be, similarly to R_2 , interpreted as either attacking the main claim or R_1 (*sink ambiguity*).

We can now ask whether a further characterisation of the precise nature of the attack relation is possible. There are at least two suggestions to further categorise the attacking relation from R_3 to R_1 . First, it could be subcategorised as qualifying the aspects of how affirmative action is helping or, rather, not helping minorities: The objection may seek to establish that affirmative action can damage minorities in their well-being since such measures might economically damage a country. In this sense, affirmative action is, in at least one aspect (well-being), not helping, even if it might help in other aspects (e.g., equality). According to this interpretation of the objection, it should be further qualified in which ways affirmative action is helping minorities. A second possible subcategorisation of the relation understands the needed qualification differently: According to this interpretation, the objection R_3 shows that R_2 is perhaps too general in its formulation. The formulated example (Sentence 6) shows that affirmative action does not help minorities *in all* countries. At best, affirmative action helps under specific conditions, which have to be spelt out, or so the objection goes.

While these and possibly other attempts to provide more specific subcategories for the attack relation might be plausible, it should be clear by now that they depend on many further assumptions, which are in no way uniquely indicated by the text. These interpretations might, therefore, be admissible but are not obligatory (*edge-type ambiguity*).

The sixth sentence suggests a rather clear case of how to subcategorise a support relation. It is formulated as an example that does not have an elaborative role but a justificatory role. It helps the reader to understand the intended meaning of ‘*can even damage*’—namely, that it is meant as there being existing cases where affirmative action has damaging consequences—and supports the reason R_3 by describing such a case (R_4).

... [8] I will argue that these objections are not convincing and do not outweigh the arguments in favour of affirmative action. [9] First, under-representation is bad in practice as well as in principle. [10] Countries are damaged by not having access to the brightest talent from the largest pool. [11] They also

benefit from a greater level of diversity; [12] e.g. representation in parliaments is improved by having more women. (R_5)

Sentences 9–12 are interesting in terms of individuating reasons. Do they express one or many reasons? For now, I want to put aside this question and regard these sentences as expressing one complex reason (R_5). Again, there are different possibilities to interpret its justificatory relation to other argumentative elements. While the eighth sentence indicates that the following sentences should be interpreted in favour of the main claim, it does not tell us whether these considerations should be interpreted as directly supporting the main or indirectly supporting it by attacking objections to the main claim.

Without any further hints, different interpretations come to mind. First, we might interpret it as a reason that supports the main claim. Similar to R_1 , it can be interpreted as a practical syllogism: Ending under-representation is an aim we should pursue for different reasons (sentences 10–12), and affirmative action is a means to that end. Second, it can be interpreted as an attack against R_3 since it suggests that affirmative action will help countries (sentence 10) instead of damaging them. Finally, since it remains unclear whether the proponent of this reason allows exceptions to the “rule” expressed in sentence 10, R_5 might even attack R_4 .

These interpretive considerations are by no means conclusive or without alternatives. But what they suggest is the following: First, the analyst might often encounter sink-source ambiguity. In the example, sink-source ambiguity was partly the result of node ambiguity. If the analyst cannot uniquely determine whether a text segment has to be interpreted as a reason, there might be corresponding ambiguities in determining the sources and targets of justificatory relations—simply because the set of nodes that act as candidate relata for these relations might then differ from interpretation to interpretation. Additionally, linguistic cues will typically explicitly mark text segments as objections or justifications but are often not that explicit about what they object to or justify. Consequently, the analyst can expect that the sources of attack or support relations can often be determined uniquely, while the co- and context might underdetermine their targets.

Second, while the categorisation of justificatory relations in terms of support and attack relations did not allow any degree of interpretation, it is, at this point, not clear whether there are subcategories of these relations that can be determined in a similarly unique way. Consider, for instance, the attempt to introduce subcategories in terms of what can be called probative or modal force: The support and attack relation are grounded in the notion of support. Attacking p is explicated in terms of the support for p being false. But support comes surely in degrees. Arguments and reasons can support what they intend to support more or less; they can make it more or less likely to be true—that’s why the term probative force. Our natural language offers lots of possibilities to express these nuances. A reason can provide some, weak or conclusive evidence; a reason can vindicate or prove something, or show and establish its truth; a consideration can dispute or challenge a point, or it can rebut and debunk it or even refute and disprove it; two different points can be in tension or conflict to each other, or they can contradict each other.

Although our natural language allows us to express different degrees of support, it is hard to determine the degree associated with a specific expression. It will depend on the context. What is more, in most cases, we cannot further qualify this degree by numerically precise

numbers because we hardly think of numbers when we use these termini. The differences between these expressions are, to some extent, vague; their meanings will overlap. Without any further clarification on how to understand the notion of graded support and to translate our ordinary-language vocabulary into these degrees of support, it is, to my mind, pointless to introduce any subcategories of support and attack that are based on these degrees. The resulting degree of interpretation will not be worth the effort.

4.3.3 INDIVIDUATION OF ARGUMENTS AND REASONS

We already saw that there is not necessarily a one-to-one correspondence between arguments and reasons. An argument can comprise more than one reason. But independent of whether the analyst decides to focus on identifying arguments or reasons or both, they have to find out where arguments or reasons start and end in a text. The problem is that there is no general rule of how argumentative components, such as premises and reasons, relate to grammatical units, such as sentences. The task for the analyst is then to individuate arguments and reasons—that is, to decide which text segments make up one particular reason or one particular argument.

The possibility of repeating arguments and reasons issues but one challenge of individuating arguments and reasons. Two different text segments might express the same reason or the same argument. An arguer might, for instance, slightly reformulate a point to enhance clarity or to express its significance. It is, however, not always easy to decide whether the arguer repeated some point or intended to formulate a new reason or argument. A straightforward suggestion is to invoke the notion of semantic equivalence. If two text segments are semantically equivalent—that is, if they have the same meaning—they cannot express different premises or reasons. We simply have to ask ourselves if text segments have the same meaning or, rather, the same intended meaning. But how do we decide that? If two sentences are identical—that is, if they are formulated by using the exact words in the same order—they have the same meaning. This obvious fact is, however, not that helpful since it will rarely happen that an arguer uses the same sentence a second time, even if they want to repeat some point. Our natural language offers ample means to reformulate one point in different ways. Deciding whether two text segments express the same thing is not as straightforward as one might think. I think we are often able to decide such questions uniquely. But we can also expect that there are cases that allow for different interpretations as to whether two text segments are semantically equivalent.

Another paradigmatic challenge is what I will refer to as *horizontal underdetermination of individuation*. If different text segments are expressed as a justification for different claims, they surely belong to different arguments. But if we encounter different text segments that are supposed to justify the same claim, we have to decide whether they belong to the same or different arguments. Similarly, we have to decide whether they express one complex reason or more than one reason.

Often, we are competent in identifying different reasons and arguments and in distinguishing them from each other. From the perspective of non-reconstructive argument analysis, the analyst will rely on their intuitive contextual understanding and on linguistic cues they find. Again, the arguer may provide the audience with indicators that uniquely pin down the individuation of reason and arguments. They might formulate linguistic cues that I call

direct indicators, for instance, direct argument indicators such as ‘*My first argument is ...*’ and ‘*Another argument for that claim is ...*’, or they might formulate *indirect indicators*. The latter piggyback on what we understand as arguments and reasons. For instance, an argument is a set of premises that are intended to provide a sufficient degree of support for some conclusion. The arguer can thereby indirectly signpost text segments as arguments by explicitly expressing the intended degree of support—for instance, by formulating something like: ‘... *What I’ve said sufficiently supports the claim that ...*’ (Henkemans, 1992, as discussed in Freeman 2001, 403). Suppose that an arguer presents a set of reasons or premises that are, according to their intention, in combination sufficient to justify a claim. Suppose further that they present additional reasons for the same claim. In this case, the additional reasons will belong to other arguments.

Often, however, we will not find such linguistic cues and have to rely on the cotext and possibly the context. The question is whether the cotext is sufficient to individuate arguments and reasons uniquely. Let us consider the following example:

There is no evidence that capital punishment for first degree murder constitutes an effective deterrent for these crimes. It cannot restore life to the murder victim. If applied to the wrong person, there is no way that wrong can be redressed. It signals that brutality is an option for the state. Hence, the death penalty for premeditated murder should not be a judicial option. (Freeman 2001, 403)

In this example, four different points are formulated against capital punishment. The author provided an explicit linguistic cue for the conclusion of the argumentation (the word ‘*hence*’). However, there are no other linguistic cues, particularly no linguistic cues, which help us individuate arguments and reasons. But speaking of four different points suggests we have at least four different reasons here.

The individuation of arguments is less obvious. Do the four mentioned points formulate one complex argument or many? Freeman (2001, 404) sees no ambiguity here. For him, the existence of only one conclusion indicator and the fact that the author formulated the conclusion only once is evidence enough that the passage formulates just one argument. I do not see why the number of linguistic cues or how often the conclusion is mentioned matters for the individuation of arguments. In the absence of linguistic cues, many textbooks suggest using a rule of thumb by inserting linguistic cues: If we can supplement a text by adding linguistic cues without distorting what is meant, the resulting interpretation as indicated by the supplemented text is a valid interpretation.¹¹² In the given example, it is easy to change the passage so that it explicitly formulates four arguments.¹¹³ Admittedly, this technique does not lead itself to definite answers; it might be indeterminate which alterations of the text distort the intended meaning, or even begging the question. It is a simple rule of thumb used as a pedagogical tool. Nevertheless, it casts doubts on Freeman’s straightforward interpretation of the individuation of arguments in this example.

Another example illustrates that the individuation of reasons encounters similar difficulties

¹¹²This is recommended by argumentation theorists if argument indicators are missing (see, e.g., Feldman 2014, 144, 146). The same rule is used for annotating the discourse structure in the Penn Discourse TreeBank (Prasad et al. 2008; Prasad, Webber, and Joshi 2014).

¹¹³“To summarise, there are four important arguments against capital punishment. First, there is no evidence [...], Second, it cannot [...] Hence the death penalty [...]”

if we confine ourselves to a non-reconstructive argumentation analysis.

We should not use CRISPR/Cas9 technologies as a therapeutic means to cure monogenetic diseases since they can have off-target and epigenetic effects that we are unaware of.

Both epigenetic side effects and off-target effects pose medical risks.¹¹⁴ It seems, therefore, natural to interpret this passage as one argument. But do both mentioned risks constitute two different reasons or just one complex reason?

On the one hand, both risks are different and even independent. We might observe off-target effects without epigenetic effects or the other way around. This causal independence might suggest interpreting both points as two different reasons. On the other hand, the argument alludes to medical risks. Only by interpreting both effects as medical risks can we see their justificatory relevance. But if both points constitute an objection only because they are medical risks, it seems natural to understand them as one complex reason. On this latter interpretation, CRISPR/Cas9 should not be used as a therapeutic means because it involves severe medical risks. This one reason is complex since it alludes to two causally independent sources of medical risks.

In my mind, this example shows that our ordinary understanding of what a reason amounts to is too ambiguous to sufficiently narrow down the degree of interpretation when it comes to the individuation of reasons. What we need is a conceptual clarification of this notion. We will come back to this question in a subsequent section since informal logic provides such clarifications.

4.3.4 SUMMARY

Let me briefly summarise the suggested findings of this section. I discussed the prospects and limits of non-reconstructive analysis to decrease the degree of interpretation in analysing the structure of ordinary-language argumentation. Non-reconstructive analysis draws on our ordinary-language concepts and adds some preliminary conceptual clarifications of these concepts.

I argued that although arguers may make their argumentation structure maximally explicit by extensive signposting, they will usually not do so. As a consequence, a non-reconstructive analysis of argumentation structure will—depending on the text at hand—allow for different interpretations with regard to all types of underdetermination, namely, node ambiguity, underdetermination of granularisation and relation ambiguity.

Node ambiguity prevails if the analyst faces challenges identifying argumentative units and discerning their justificatory role. I argued that it can be difficult to distinguish justificatory roles from others, such as providing causal explanations or elaborative illustrations. Additionally, linguistic cues can be ambiguous themselves. What they signpost will often depend on the cotext.

¹¹⁴CRISPR/Cas9 is a biotechnology used to alter individual gene sequences. This technology can alter genes it was not supposed to target. These so-called off-target effects can include adverse side effects. The risk of unforeseen epigenetic effects refers to the fact that the impact of gene alterations depends on further factors external to the DNA, which can differ from person to person.

The individuation of arguments and reasons faces two paradigmatic problems. First, an arguer might repeat arguments and reasons by rephrasing them. Accordingly, it can be difficult to understand whether different text segments express different arguments or reasons or whether they are intended to express the same argument or reason. Second, the analyst might be confronted with what I call *horizontal underdetermination*: If the arguer presents different text segments as justifying the same claim, the analyst has to understand whether these text segments make up one or several arguments for that claim.

The identification and categorisation of justificatory relations can also allow for different interpretations. While it is usually possible to uniquely classify justificatory relations as either attack or support relations, the target or source of a relation may be underdetermined. Additionally, it is difficult to introduce subcategories of these relations without a proper conceptual clarification—at least if the analyst is interested in narrowing down the degree of interpretation.

4.4 UNDERDETERMINATION IN RECONSTRUCTIVE ANALYSIS

In reconstructive accounts, the analysis of argumentation focuses on reconstructing arguments and is grounded in its basic understanding of what an argument is—a set of statements that are presented as justifying another statement. A person who puts forward an argument to convince others of the truth of its conclusion will do so by two independent claims. First, they claim that the argument's premises are true or at least plausible or acceptable; second, they claim that the premises lend the conclusion a sufficient degree of support. Applied formal logic explicates this notion of sufficient degree of support in terms of deductive validity.

However, deductive validity cannot be used straightforwardly to analyse ordinary-language argumentation. First, the notion of validity is explicated by its formal counterpart of *f*-validity. Accordingly, natural-language argumentation has to be formalised, which is often a nontrivial task (see, 4.2.3). This translation demands considering the cotext and allows, in some cases, for different possibilities to translate sentences. Often, the translation demands to reformulate the original sentences extensively, for our natural language is generally much more expressive than any formal language. Apart from that, ordinary-language arguments are not valid in the following sense: The sentences we *explicitly* formulate to justify some claim seldom imply the claim, even if we translate them into formal languages.

In light of the last mentioned difficulty, informal logicians question the usefulness of applying the concept of deductive validity to ordinary-language arguments. Arguers may simply not aim at this strong inferential link between their premises and the conclusion but something weaker. Applied logicians, however, favour another solution. In their view, all or at least a significant amount of these invalid arguments must be reinterpreted as valid. Arguers are understood as implicitly committed to deductive validity. Consequently, applied formal logic adheres to what is called *reconstructive deductivism*: Arguments should be reconstructed in their premise-conclusion form in such a way that the argument reconstruction is deductively valid. Reconstructive deductivism must be distinguished from the stronger claim that ordinary-language arguments are, in fact, deductive—whatever that means. In the first instance, reconstructive deductivism holds that the deductive

reconstruction of ordinary-language arguments is helpful in some way—for instance, in that it makes everything explicit that can be the subject of further scrutiny.¹¹⁵

On this view, most ordinary-language arguments are to be interpreted as what is called *enthymemes*—arguments that are incomplete with regard to their premises.¹¹⁶ Consequently, reconstructive deductivism does not only demand reformulating explicitly stated sentences but also adding further premises that the arguer did not explicitly formulate. These added premises are accordingly called *implicit premises*.¹¹⁷ One guiding criterion to identify these implicit premises is the requirement of validity: The resulting reconstruction should be deductively valid. However, as I will argue in subsequent sections, this criterion underdetermines the identification of implicit premises to a large extent.

If we reconstruct arguments as deductively valid, the reconstruction is maximally charitable to the inferential link between premises and conclusion. That has some ramifications for the mentioned third task of argument analysis—the evaluation of the argument’s rational strength. Since validity is the strongest possible support between the premises and the conclusion, an objection to the argument must aim its criticism at one or more of the premises. It cannot criticise the inferential link anymore. The same applies to the treatment of fallacies. Logic and informal logic often describe logical fallacies as flaws stemming from a lack of validity. Accordingly, the deductivist cannot criticise an argument on the grounds of being a logical fallacy. Instead, they must target their criticism to the added premises that are needed to transform a logical fallacy into a valid argument. The deductivist embraces these consequences since the guiding idea of deductivism is transparency and explicitness.¹¹⁸ Everything that might be criticizable has to be made explicit as premises to understand better possible objections raised against the argument and to judge its soundness.

Although deductive validity and the associated idea of reconstructive deductivism play a central role in applied formal logic, applied logicians are not committed to the view that *all* ordinary-language arguments are implicitly valid (*universal deductivism*). They can adhere to a weaker form of reconstructive deductivism, which allows some ordinary-language arguments to be considered non-deductive (*restrictive deductivism*). Proponents of universal deductivism will think of deductive validity as a necessary condition for inferential adequacy. In contrast, advocates of restrictive deductivism will allow additional explications of inferential adequacy, for instance, by explicating a notion of inductive strength.

¹¹⁵This view is, for instance, endorsed by Ennis (1982). The question is, of course, how much should be made explicit and where we draw the line between reconstructing an argument and constructing a new one. Burke (1985), for example, argues that the deductivism of Ennis (1982) leads to the requirement of making something explicit, which justifies the inferential link between the premises and the conclusion. But justifying the inferential link is, in most cases, the construction of a new argument in favour of the original argument’s inferential adequacy.

¹¹⁶As Hitchcock (2019) reminds us, Aristotle did not use the terminus *enthymeme* to denote arguments with unstated or missing premises. For him, an enthymeme is a syllogism from likelihood and sign. The current understanding as an incomplete syllogism dates back to the Stoics and has been repeated ever since. Walton (1996a, chap. 7) provides further etymological remarks.

¹¹⁷Cf. Footnote 86 on page 95 for some terminological remarks.

¹¹⁸For informal logicians these consequences speak firmly against deductivism. Grennan (1994) regards them as “unpalatable positions” (191). Further objections can be found in Govier (1992b), Govier ([1987] 2018) and D. M. Godden (2005). Proponents of reconstructive deductivism include Groarke (1992), Groarke (1999) and Botting (2016).

In sum, applied formal logic is consistent with different views towards deductivism. Its distinguishing feature is the inclination to prefer deductive reconstructions of arguments and to prefer formal methods in analysing non-formal arguments.¹¹⁹

The requirement to reconstruct arguments as deductively valid by reformulating explicitly mentioned statements and adding implicit premises or conclusions cannot be the only principle to guide reconstruction. The problem is that there are often numerous possibilities to reconstruct an argument as deductively valid. The applied logician uses additional hermeneutical concepts and techniques to constrain a possibly exploding interpretative margin. Two important hermeneutical principles guide the reconstruction of arguments.

The *principle of accuracy*, or *faithfulness* demands that the interpretation reflects the arguer's intended meaning of the argument. The formulations used in the reconstruction should be equivalent or at least consistent with what the arguer meant. Identifying the intentions without being able to ask the arguer about them—which is the default case in content analysis—is often tricky. The rough recommendation is to consider the whole cotext and context, if possible and feasible, and use this information to identify intentions that best fit the whole picture. The identified intentions should explain what is known about the arguer as best as possible.

Accuracy alone will not sufficiently narrow down the degree of interpretation. An additional principle can help to reduce the degree of interpretation further. The *principle of charity* demands that the reconstruction performs well with respect to rational strength. Reformulations of explicit statements should not become less plausible than the original statements. The analyst should also refrain from adding implicit premises that are obviously false or implausible. From the perspective of applied formal logic, the principle of charity is motivated by a concern for fairness. After reconstructing the argument, the final step is a critical evaluation of the argument in terms of its rational strength. A critique that is the mere result of an adverse interpretation can be judged to be unfair if a more favourable interpretation exists.

Both principles are often in a trade-off to each other (Feldman 2014, 157; Betz and Brun 2016, 44; Brun 2016, 261). It might be necessary to reformulate premises or conclusions in a way that deviates from the author's formulations to arrive at a favourable reconstruction. But which principle should receive priority? The answer will ultimately depend on the particular goals of argument reconstruction. If we are interested in finding the best arguments and the text's formulations merely serve as a starting point, we should prefer charity. However, if we want to understand and assess the originally formulated argument, we have to consider the actual wording of the text more carefully. Clearly, CAAS aims to analyse the arguer's point of view. Hence, it seems that problems of weighing accuracy against charity do not concern the argument-reconstructing content analyst since they should always prefer accuracy. However, real-world arguments do not allow for such a neat solution, as we will see shortly in the case of identifying implicit premises. The reason is that charity is often instrumentally relevant for accuracy since we can often assume that the arguer intends to present arguments with plausible premises that sufficiently support

¹¹⁹There are also formal accounts of non-deductive arguments in the form of non-monotonic logic (see, Prakken and Vreeswijk 2002 for an overview) and probability logic (see, Demey, Kooi, and Sack 2019 for an overview).

their conclusion.

Like the applied logician, the informal logician will use the hermeneutical principles of accuracy and charity to reconstruct arguments. The most crucial difference between them concerns the diverging importance they give to deductive validity as a helpful concept in argument reconstruction. The informal logician does not reject the use of deductive inference rules. The difference is that they take the phenomenon of deductive arguments as marginal in ordinary-language argumentation. Therefore, deductive inference rules are only of limited use for reconstructing arguments. For them, the reconstruction is predominantly guided by argument schemes and non-strict inference rules.

The informal analogue of deductive validity can be called informal validity (Groarke 2019) and is often spelt out by using two concepts instead of just one. To satisfy the desideratum of informal validity, the premises of an argument need to be relevant for the conclusion—thas is, they have to provide some evidential or probative support for their conclusion—and they have to provide sufficient support in their combined form for their conclusion.¹²⁰

The notion of premise relevance is of importance for the granularisation of argumentation. Applied formal logic focuses on arguments and their constituent elements—premises and conclusions. However, applied formal logic does not give the analyst any means to count reasons. Counting premises as reasons would not do since the analyst is free to rearrange any set of premises into one single premise by formulating their conjunction. In contrast, the informal logician's additional concept of premise relevance offers a means to individuate and count reasons.

In sum, while applied formal logic only provides means to individuate arguments, informal logic can help the analyst to individuate arguments and, additionally, to individuate reasons. What is more, the individuation of reasons does not presuppose already individuated arguments. Therefore, the extent of underdetermination can differ between arguments and reasons. For instance, the individuation of arguments might be underdetermined, while the individuation of reasons is not or at least less underdetermined.

I will analyse the extent of hermeneutical underdetermination in reconstructive analysis along the following steps: In Section 4.4.1, I will describe how a reconstructive analysis deals with node ambiguity, which is not different between both paradigms. I will then move on to discuss the different possibilities to deal with underdetermination in the granularisation of argumentation in applied formal logic (4.4.2) and informal logic (4.4.3–4.4.4). In the context of non-reconstructive analysis, I conceptualised two justificatory relations (support and attack) between text segments in a general way and discussed one specific example to illustrate the accompanying indeterminacies of identifying these relations and their relata (4.3.2). The question is whether applied formal logic or informal

¹²⁰Though the applied logician has also a notion of premise relevance, it is conceptually reducible to deductive validity. For them, a premise is said to be relevant for the conclusion if it is needed for the argument to be deductively valid. Here, *needed* is meant in a weak sense. A premise is needed in this weak sense if the argument turns out to be invalid if we remove the premise from the argument. However, a premise can often be substituted by other premises without jeopardising validity, which motivates the qualifier 'weak'. The corresponding relevance requirement of applied formal logic demands reconstructing arguments with the logical minimum: An argument should not contain a premise that is not needed for the argument's validity.

logic can help narrow down this degree of interpretation. After introducing for each paradigm the basic ideas and one particular account of analysing the macrostructure of argumentation (4.4.5–4.4.6), I will analyse one specific example from the viewpoint of both paradigms (4.4.7). Finally, I will close this section by summarising the most important results (4.4.8).

4.4.1 NODE AMBIGUITY

Argumentation theoretic accounts offer the analyst complementary, not contradictory means to those of non-reconstructive analysis to identify argumentative units. In particular, the described non-reconstructive approach can be considered the starting point to identify argumentative units, which precede a more thorough analysis by reconstructing these units as arguments with an internal premise-conclusion structure.

The reconstruction of arguments can shed light on ambiguous cases of identifying argumentative units by applying the principle of charity. The principle of accuracy advises the analyst to categorise text segments as justificatory relevant according to their intended function. Here, reconstructive accounts do not offer something in addition to considering linguistic cues and cotext. The principle of charity, however, demands avoiding unfavourable interpretations of a text. In particular, the premises of a reconstructed argument should be as plausible as possible. A helpful rule of thumb is the following: If linguistic cues and cotext do not uniquely determine whether a text segment is justificatory relevant, the analyst should reconstruct the apparent argumentative unit(s) in their premise-conclusion form. If the resulting reconstruction performs (very) poorly by evaluative standards of argument strength, the text segment(s) can be considered argumentatively irrelevant. In other words, text segments that are ambiguous as to their justificatory role should only be regarded as argumentative units if they can be considered part of a charitable argument reconstruction.

This rule of thumb will sometimes help decide whether a text segment expresses an argumentative component. However, the described charity assessment might differ from analyst to analyst. The plausibility of premises in an argument reconstruction is, first, a matter of degree and, more importantly, can depend on the subjective judgements of the analyst. There will be cases where such plausibility checks come down to verifying whether premises correspond to or violate factual knowledge. In these cases, analysts should come to the same conclusions. However, evaluating how plausible a statement is will often be more nuanced and complicated. The analyst will then have to rely on their background knowledge, which can vary from analyst to analyst.

4.4.2 INDIVIDUATION OF ARGUMENTS IN APPLIED FORMAL LOGIC

Applied formal logic focuses on identifying arguments and their constituent elements—that is, premises and conclusions. The problem of individuating arguments is to decide which justificatory relevant text segments are part of one argument and which belong to different arguments. The focal concern of this section is to assess whether the reconstructive strategies of applied formal logic help to narrow down the degree of interpretation of individuating arguments.

We will first consider one paradigmatic challenging type of how argument individuation can allow for different interpretations, which I will call *horizontal underdetermination of argument individuation*, or in short *horizontal underdetermination*. Suppose we encounter a larger set of sentences presented as a justification for a specific claim. Following Eemeren and Grootendorst (1984), I will refer to such an argumentation as multiple if these sentences have to be regarded as more than one argument for the claim. In a multiple argumentation, we have different premise sets that are each intended as justificatory sufficient for the claim.¹²¹ In contrast, the argumentation is co-ordinative compound if these sentences represent just one single argument.¹²² Such a situation represents a case of horizontal underdetermination if the argumentation can be interpreted in both ways—as a multiple and a co-ordinative compound argumentation.

Suppose the cotext and linguistic cues are insufficient to decide whether these sentences represent one or many arguments and how to group them into single arguments. How can applied formal logic, particularly the validity requirement—that is, the requirement to reconstruct arguments as deductively valid—help to individuate arguments? According to applied formal logic, we should find a partition of these statements into groups such that each group can be adequately reconstructed as one argument.

Suppose now that one or some individual statements already logically entail the claim independently. In this case, we should regard each of these statements as a one-premise argument and can regard the other statements as belonging to other arguments for the claim (Freeman 2001, 402). In the next step, we should check whether the set of these remaining statements contains pairs that logically entail the claim. Again, we can consider such pairs as individual arguments. We will then search for larger and larger subsets in the set of remaining statements that entail the conclusion until we accounted for all sentences.¹²³

However, such simple cases rarely occur. Usually, the explicitly formulated premises will not logically entail what is supposed to be justified. Therefore, no proper subset of them will entail the conclusion. In these cases, the task of individuating arguments will be more complicated. We would have to choose a partition of subsets and reconstruct each subset of statements as a valid argument by adding implicit premises. However, there are different ways to select partitions. To choose between them, we have to compare the different

¹²¹ Although Eemeren and Grootendorst (1984) do not couch their definition in terms of individuating arguments, their formulation is almost equivalent if we consider the suggested concept of argument as a set of premises intended to be justificatory sufficient for a claim. According to them, a “multiple argumentation, [is an argumentation] in which the premisses are all individually sufficient but in which none of the premisses is individually necessary to justify or refute the expressed opinion” (Footnote 30 on p. 197). The only difference is that each premise has to be individually justificatory sufficient on its own, which would amount to one-premise arguments only. I see no reason for this confinement. What is more, it is a mere stylistic difference since every premise-set of an argument can be rewritten as one single premise by using the conjunction of its members.

¹²² Again, Eemeren and Grootendorst (1984) formulate their definition slightly differently: A “co-ordinative compound argumentation, [is an argumentation] where each of the premisses is individually necessary but where the premisses are only sufficient together” for the claim (Footnote 30 on p. 197).

¹²³ There is a simple reason why the analyst should begin to search for one-premise arguments and incrementally enlarge the search area. Arguments should be reconstructed with the logical minimum—that is, with as few premises as possible. Every additional premise not needed for the inferential link is an additional point of a possible attack. According to the principle of charity, we should, therefore, refrain from adding these unnecessary premises.

partitions according to how well their reconstructed arguments perform with respect to the different hermeneutical principles.

For practical reasons, reconstructing all arguments of all combinatorically possible partitions will usually not be possible since the amount of subsets increases relatively fast. There are 2^n possible subsets for a set of n statements, which becomes 16 for only four statements. Fortunately, there is often no need to consider all subsets. The non-reconstructive analysis will help us group premises into single semantical units—premises that belong together by expressing one point. But even if we do not have to reconstruct all combinatorically possible subsets, we are often left with different possibilities of partitioning sentences into single arguments. The guiding idea is to reconstruct these argument candidates and judge them by the different hermeneutical principles. Since the different subsets will not satisfy the validity requirement on their own, identifying implicit premises will be the deciding factor of argument individuation. Let us, therefore, more carefully consider different suggestions of how to deal with enthymemes.¹²⁴

Argumentation theorists often distinguish between two types of implicit premises: needed and used premises.¹²⁵ A used implicit premise is a sentence that is part of the argument as it is intended by the arguer and, for some reason, not mentioned explicitly. This understanding of used implicit premises invites us to search for used premises by thinking about the intentions of the arguer and demands to invoke the above-mentioned principle of accuracy.¹²⁶ There are different possible explanations for why an arguer decides not to mention a premise. They might, for instance, omit premises that they regard as part of the shared background knowledge or as easily deducible from the context. By omitting these

¹²⁴Two precautionary remarks are important to keep in mind. First, strategies to identify implicit premises are not specifically devised to guide argument individuation. Rather, they are discussed and formulated for cases that already presuppose an individuation of arguments. The question is whether they can be fruitfully used to individuate arguments. Second, although I ponder about enthymemes from the viewpoint of applied formal logic, enthymemes are widely discussed in informal logic, applied logic and Pragma-Dialectics alike. The following considerations are largely drawn from the informal logic literature.

¹²⁵Ennis (1982) introduced both concepts and further distinguished between back-ups and gap-fillers. Back-ups are implicit premises that provide further justification for premises that are part of the argument; gap-fillers are premises used or needed to ensure an adequate inferential link between the premises and the conclusion—in the case of applied logic to satisfy the validity requirement.

¹²⁶Walton prefers to think of used implicit premises (or even implicit premises in general) less in terms of the arguer's intentions but in terms of their commitments (see Walton 1996a, 248–52; 2009; 2006, 157–58). Similarly, proponents of Pragma-Dialectics adhere to a commitment view of implicit premises (see, for instance, Eemeren and Grootendorst 1992, chap. 5 and 6). A commitment can be understood as something a subject is bound to do. In the context of implicit premises, proponents of this view think first and foremost of Gricean conversational implicatures, which are commitments resulting from performative acts—that is, actions done by uttering something. The paradigmatic case is the expression of a promise, which results in the corresponding commitment to keep that promise. Gerritsen (2001, 72) advances important reasons to preferring to concentrate on commitments instead of intentions. She explains that the concept of commitment is theoretically captured and well understood by Gricean speech act theory. Additionally, it is easier to reliably attribute commitments than intentions since commitments are more directly linked to the conversational context and do not depend on the arguer's mental state. I tend to side here with Burke (1985, 117–18), who maintains that analysing the commitments of the arguer is instrumentally valuable in finding the arguer's intentions but does not consider commitments as the defining feature of (used) implicit premises. However, nothing important hinges on that for underdetermination. The granularisation of argumentation is often underdetermined independent of whether you think of (used) implicit premises in terms of intentions or commitments only.

premises, communication can advance more efficiently, and interlocutors can focus on controversial parts of the argumentation.¹²⁷ However, the concept of a used premise should not be confined to propositions entertained consciously or unconsciously by the arguer. Instead, it is about being faithful to the intended idea of the argument. And being faithful to the idea of the argument might demand to consider premises as part of the argument of which the arguer did not think so far.¹²⁸

A needed implicit premise, on the other hand, is a premise that is added to strengthen the inferential link between the premises and the conclusion.¹²⁹ It remedies the lack of inferential adequacy and is also called an inferential gap filler. In the deductivist paradigm, needed premises are those that will make the argument valid.

The distinction between needed and used implicit premises is a distinction between two different types of premises. These types are, however, not exclusive to each other. A premise can be both needed and used (Ennis 1982). The distinction is important because both types call for different strategies to identify implicit premises.¹³⁰ Although these strategies can lead to a convergent identification of an implicit premise, they will often pull in different directions; these strategies might compete with each other or even represent diverging goals of argument reconstruction. Being aware of these trade-offs is necessary to balance the different reconstructive goals consciously.¹³¹

So, how do we identify implicit premises that are needed? The suggested criteria comprise gap-filling ability, plausibility and preservation.¹³² The gap-filling ability is the potential of

¹²⁷Hitchcock (1985, 86) opposes this understanding and thinks it as a marginal phenomenon. According to him, what argumentation theorists usually identify as implicit premises is seldomly consciously entertained by the arguers.

¹²⁸This understanding of used premises is due to Burke (1985) and is broader than the one of Ennis (1982), who characterised used premises as propositions “that a person actually used consciously (or subconsciously, if you believe in subconscious reasons) as a basis of argument” (p.63).

¹²⁹The term *needed* is a little bit unfortunate, as Hitchcock (1985) observes. First, there are always different possibilities for doing this job. Hence, one specific chosen candidate is not needed in a strict sense, since there are always alternatives. Second, and as we will see shortly, what is often identified as an implicit premise is always logically stronger than what is needed to satisfy inferential adequacy.

¹³⁰Of course, some argumentation theorists disagree. For instance, Gerritsen (2001) is sceptical of the distinction “because it implies that ‘used’ premises are not the premises that make the argument logically valid. [...] This would lead to a definitive split between logic and argumentation theory on the one hand, and the study of ordinary argument on the other” (68). This is, however, a misinterpretation since it labours under the assumption that both categories are exclusive to each other.

¹³¹Burke (1985) writes that “we should be clear whether our purpose is (1) to identify the argument intended by the arguer or (2) to determine whether the stated premises provide the makings of an acceptable argument for the conclusion” (107). I would add that these purposes are not exclusive and occasionally have to be balanced against each other.

¹³²The first two are discussed by Ennis (1982) and preservation is mentioned by Burke (1985). Ennis (1982) adds a third one—fidelity to the argument—which requires adding only those premises that do not distort the argument. The addition of premises should be faithful to the argument and not lead to an argument that has little to do with the original argument. Ennis (1982) depiction is slightly obscure. According to him, “the fidelity criterion assures that we are evaluating what we want to evaluate” (72). However, he does not further explain what exactly we should be faithful to and what we want to evaluate. A natural suggestion is to consider the arguer’s intentions or commitments and, in particular, the arguer’s intended idea of the argument. But that is what he elaborates on in the context of identifying used premises. Since I consider the arguer’s intentions and commitments as relevant for used implicit premises, I prefer to think of fidelity as primarily relevant for identifying used premises—though little hinges on that.

premise candidates to ensure inferential adequacy in a non-redundant way. Non-redundancy demands that a premise should be needed in the sense that removing it from the premise set would hamper the inferential link between the remaining premises and the conclusion. In the case of strong deductivism, the gap-filling ability amounts to the already mentioned validity requirement: Implicit premises should, together with the explicit premises, logically entail the conclusion. Weak deductivism and informal logic allow for other types of inferential links, which must be explicated and result in additional possibilities to satisfy inferential adequacy. The desideratum of plausibility is the already mentioned principle of charity with regard to premises. Implicit premises should be as plausible as possible. Lastly, preservation demands that adding premises should not change the argumentative role of other explicit parts of the arguments. In particular, an implicit premise should not render an explicit premise redundant.

The principle of accuracy guides the identification of used premises. According to the suggested understanding, used implicit premises should best fit the intended idea of the argument. Hence, we need criteria that help us identify the arguer's intentions. Sometimes, they are obvious. If, however, we have different competing hypotheses about these intentions, these hypotheses should be judged by their ability to provide explanations of facts about the arguer and the context in which the argument was being put forward. In other words, we should attribute intentions to the arguer inasmuch they best explain what the arguer explicitly expressed and what we know about them.¹³³

I started by considering the case of horizontal underdetermination. When analysts are presented with a set of sentences, all of which are intended as a justification for one claim, they have to decide how to partition them into single arguments. Generally, these sentences do not logically entail the claim. The analyst has, therefore, to use criteria for identifying implicit premises to individuate arguments. The question is how far the discussed strategies to identify implicit premises will take us. Can they pin down implicit premises and, thereby, a particular granularisation of the argumentation in a unique way?

The mentioned possible tension between strategies to identify used and needed premises might constitute one problem for a unique individuation of arguments since different ways of prioritising these strategies can result in different implicit premises, thus allowing for different interpretations. One might think that such trade-offs are irrelevant in content analysis since the purpose of content analysis is to pin down the arguments in line with the arguer's intentions. This intuition suggests a clear rank order of the hermeneutical principles—one that always prioritises criteria to identify used implicit premises over those to identify needed premises. The content analyst of argumentation structure should, therefore, at first search for used implicit premises. If they are left with different competing premise candidates, they should choose among them according to their gap-filling ability, plausibility, and the criterion of preservation.¹³⁴

¹³³See Ennis (1982, 67). Walton (1996a, 251) discusses additional clues for identifying used implicit premises.

¹³⁴In the context of argumentation theory, Walton (1996a) suggests the opposite rank order. He advises to first "see what is required for that type of argument to meet its burden of proof, and to see what is missing to meet these requirements" and then "to ask which proposition the arguer was most likely to have used in getting the conclusion, given what we know from the text and context about the arguer's commitments" (249).

Even if this idea to rank-order these criteria will appropriately answer the question of possible trade-offs,¹³⁵ the individuation of arguments will still often be underdetermined in the context of content analysis. The reason is that the cotext is often too sparse to narrow down the intentions or commitments of the arguer sufficiently. Additionally, the strategies to identify needed premises face some problems not mentioned so far. Let us, therefore, reconsider the capital-punishment example in more detail to see how applied formal logic individuates arguments.

There is no evidence that capital punishment for first degree murder constitutes an effective deterrent for these crimes. It cannot restore life to the murder victim. If applied to the wrong person, there is no way that wrong can be redressed. It signals that brutality is an option for the state. Hence the death penalty for premeditated murder should not be a judicial option. (Freeman 2001, 403)

The example is possibly a paradigmatic case of horizontal underdetermination: It expresses four different reasons against capital punishment that do not by themselves logically entail the conclusion; there are no linguistic cues that would help us to decide whether these points are intended as one single argument or as different arguments against the claim.¹³⁶ Additionally, no further cotext is provided to employ strategies for identifying used implicit premises. From the deductivist's point of view, we should search for implicit premises that render the deduction from these reasons as either one valid argument or different valid arguments.

There is one very simplistic suggestion to reconstruct all points as one argument.¹³⁷ We just use the first four sentences as premises and the last as the conclusion. Deductive validity can be achieved by adding one other sentence as a needed premise:

<Capital Punishment - First Version>:

- (1) There is no evidence that capital punishment for first-degree murder constitutes an effective deterrent for these crimes.
- (2) It cannot restore life to the murder victim.
- (3) If applied to the wrong person, there is no way that wrong can be redressed.
- (4) It signals that brutality is an option for the state.

¹³⁵I think the idea to rank-order these hermeneutical strategies will face some problems in exegetical reality since the distinction is not as clear cut as one might hope. Often, we will have to employ the principle of charity when we are after the intentions of the arguer. The reason is quite simple: In contexts where we can assume that the arguer adheres to veracity, we can assume that they intend to rely on or make use of true or at least plausible sentences. But often our sole guide to judge whether the arguer thinks something to be true or plausible will be our judgement as analysts about what is true or plausible.

¹³⁶The lack of validity is easily explained: The explicitly formulated reasons are descriptive statements and the conclusion is a normative statement. However, to be valid, arguments with a normative conclusion do need a normative premise or a bridge principle expressing the transition from the descriptive realm to the normative. Otherwise, the argument would constitute what is often called a naturalistic fallacy.

¹³⁷In this section, I will, for brevity and simplicity of exposition, neither formalise the presented argument reconstructions nor specify the used inference rules of formal logic to demonstrate deductive validity. I merely formulate the sentences in a way that makes their logical form more or less clear. You have to trust me that the argument reconstructions do not fall short of validity.

- (5) If (i) there is no evidence that capital punishment for first-degree murder constitutes an effective deterrent for these crimes and (ii) if it cannot restore life to the murder victim and (iii) if it is true that if applied to the wrong person, there is no way that wrong can be redressed and (iv) if it signals that brutality is an option for the state, then the death penalty for premeditated murder should not be a judicial option.
 -- from 1-5 --
 (6) Hence, the death penalty for premeditated murder should not be a judicial option.

The added premise (5) guarantees validity. It is even possible to reconstruct all invalid arguments as valid using the same strategy. Suppose that someone formulates several sentences p_1, \dots, p_n as reasons for some other sentence c and that c is not already logically implied by these sentences. If we simply add the sentence “If p_1, \dots, p_n , then c .” as a further premise, the resulting argument will be logically valid. This sentence is called the *associated conditional*.¹³⁸ It is a conditional statement expressing that the conclusion is true if the premises p_1 to p_n are true.

Adding the associated conditional seems like a cheap trick to satisfy the validity requirement; Feldman (2014) refers to it as *cheap validity*. The problem is that this move does not seem very informative. From the perspective of argumentation theory, reconstruction aims at a deeper understanding and systematic evaluation of arguments. Simply adding the associated conditional does not help reach either of these aims. This strategy does not contribute to our understanding of an argument since the associated conditional merely reformulates that there is a justificatory connection between the sentences p_1, \dots, p_n and c but does not further elaborate on what this connection is based. For the same reason, the strategy does not reveal a new perspective to evaluate the argument. We can simply assess the justificatory link between premises and conclusion without adding the associated conditional as an implicit premise.¹³⁹

These problems are particularly relevant to the aims of the argumentation theorist. Above that, the strategy to use the associated conditional will face similarly severe problems for the individuation of arguments—at least if principles to identify used premises cannot decide between different associated conditionals. Suppose, for simplicity, there are only two sentences p and q that are brought forward as reasons for c and which do not logically imply c . The strategy to add an associated conditional as an implicit premise allows to individuate these reasons as one argument

<A1>

¹³⁸The term is coined by Hitchcock (1985). Burke (1985) calls it the *reiterative candidate* and Grennan (1994) the *inference claim*.

¹³⁹These problems are mentioned by informal logicians (e.g., D. M. Godden 2005; Govier [1987] 2018, 6:132), proponents of applied formal logic (e.g., Feldman 2014, 385) and advocates of Pragma-Dialectics (e.g., Eemeren and Grootendorst 1992, chap. 5) alike.

```

(1) p
(2) q
(3) If p and q, then c.
-- justified by 1-3 --
(4) Hence, c.

```

or as two arguments

```

<A2>

(1) p
(2) If p, then c.
-- justified by 1-2 --
(3) Hence, c.

<A3>

(1) q
(2) If q, then c.
-- justified by 1-2 --
(3) Hence, c.

```

Both suggestions satisfy the validity requirement. The next step is to compare these conditionals with regard to their plausibility. A sentence of the form ‘*If s, then r.*’ logically implies another sentence ‘*If s and t, then r.*’ for any other sentence *t*. For if *s* is a condition sufficient for *r*—as expressed by the first conditional—then *r* will be true if *s* is true, whatever else may be the case. On that reasoning, both conditionals <A2.2> and <A3.2> logically imply the conditional <A1.2>. Now, we can compare the plausibility of these different conditionals given our knowledge about their logical relations. It is reasonable to assume that a sentence *w*’ is at least as plausible as a sentence *w* if *w*’ is logically implied by *w*.¹⁴⁰ As a consequence, the conditional <A1.2> is at least as plausible as the conditionals <A2.2> and <A3.2>. But then there is no reason to choose the finer granularisation <A2> and <A3> from the perspective of charity. Only if the associated conditionals of <A2> and <A3> are as plausible as the associated conditional of <A1>, charity allows choosing the finer granularisation of argument individuation. And even in this case, the individuation of arguments would be underdetermined. The analyst could still decide to interpret the reasoning as co-ordinative compound. The problem gets more pressing if we have more than two sentences advanced in favour of some claim.

These considerations show that possible cases of horizontal underdetermination pose a

¹⁴⁰If *w* logically implies *w*’, every logical consequence of *w*’ is also a logical consequence of *w*, since the consequence relation is transitive. If we regard the logical consequences of a sentence as a measure of its expressive power, we can say that *w* is at least as expressive as *w*’—that is, *w* states everything that *w*’ states and possibly more. Every objection to *w*’ will be, therefore, also an objection to *w* but not necessarily the other way around. Hence, *w* can only be less plausible as *w*’ or as plausible as *w*’. An explication of *plausibility* should, therefore, satisfy the constraint that each consequence *w*’ of a sentence *w* should be at least as plausible as *w* itself. All prominent formal accounts of plausibility, such as Bayesianism and Dempster-Shafer theory, satisfy this constraint (Halpern 2005).

problem for the individuation of arguments if we reconstruct them using the associated conditional and if criteria of identifying implicit used premises cannot decide between them. The problem is that charity will either suggest opting for an interpretation as a coordinative compound argumentation or will underdetermine the multiplicity of arguments. Either way, charity will never give the analyst conclusive grounds for an interpretation as multiple argumentation. Since we often have good reasons to interpret an argumentation as multiple, using the associated conditional as an implicit premise and relying on charity alone is not helpful as a strategy to individuate arguments.

Although most argumentation theorists agree on the futility of picking the associated conditional as a needed implicit premise,¹⁴¹ it is not enough to simply advise not picking the associated conditional for that role. The problem is that the associated conditional is the perfect fit with regard to the criteria I described so far (Burke 1985). It is accurate since we can assume that the arguer is committed to there being an inferential link between the premises and the conclusion. The associated conditional simply expresses this claim (Hitchcock 1985). It is charitable because it is the logical minimum to satisfy validity. Any other sentence that guarantees validity will logically imply the associated conditional or will be logically equivalent to it.¹⁴² As a consequence, the associated conditional is at least as plausible as every one of its alternatives.¹⁴³ Using another sentence than the associated conditional can make things only worse in terms of plausibility. Since the associated conditional already guarantees validity, it does not seem to be charitable to use a sentence that is possibly less plausible and, in the best case, as plausible as the associated conditional.

These considerations show that the discussed criteria to identify needed premises favour the associated conditional. However, using the associated conditional is both from an argumentation theoretic point of view and the perspective of individuating arguments an unsatisfactory option.

The requirement of validity drove the consideration of adding the associated conditional. Consequently, there are two possible reactions to the mentioned problems. Either we weaken our requirements for the inferential link between the premises and the conclusion or provide other strategies to identify implicit premises. Informal logicians will favour the former response and deductivists the latter. One such deductivist suggestion is using a generalisation of the associated conditional as an implicit premise. Consider, for instance, the philosopher's favourite illustration of a syllogism: "*Socrates is mortal because Socrates is a man.*" The associated conditional "*If Socrates is a man, Socrates is mortal.*" has one expression—the proper name '*Socrates*'—which is shared by the antecedent and the consequent of the conditional. The idea is to find a generalisation over this shared expression. The sentence "*All men are mortal.*" is such a generalisation and can be regarded

¹⁴¹I am only aware of minimal concessions to the strategy of adding the associated conditional. For instance, Ennis (1982), being aware of the strategy's lack of informativeness, remarks that "in argument reconstruction contexts, it might help to be reminded that an arguer is committed to at least this [the associated conditional] minimal claim." (83).

¹⁴²Let p a sentence that is brought forward as a reason for r , but that does not logically entail r , and let q any sentence that together with p logically entails r . In other words, the conjunction ' p and q ' logically entails r . That can be formalised as '*Necessarily, if p and q , then r .*' But that implies '*Necessarily, if q then also: if p , then r .*' That, in turn, can be interpreted as q entailing the associated conditional '*If p , then r .*'

¹⁴³See Footnote 140 on page 124.

as the intuitively plausible candidate for the implicit premise.

The question is whether this strategy to generalise over repeated expressions succeeds in every case and whether it can help us to solve the problem of individuating arguments. The most elaborate account of the idea was proposed by Hitchcock (1985).¹⁴⁴ He shows that implicit premises identified by this strategy coincide in many cases with intuitions about what the implicit premise is,¹⁴⁵ but also acknowledges that in some cases the intuitively correct implicit premise is logically weaker and in others logically stronger.

Let's grant that these cases are rather exceptions. The more pressing issue concerning the challenge of horizontal underdetermination is the problem of indeterminacy addressed by Hitchcock. In many cases, there are different possibilities for generalising the associated conditional. Arguments might contain more than one repeated expression. Hitchcock (1985) considers the following case:

Marijuana should be legalized, because it is no more dangerous than alcohol, which is already legal (92).

The example contains the repeated expressions '*marijuana*', '*alcohol*' and '*legal(ized)*'. The question is over which of these expressions we should generalise. Over all of them or only some? Hitchcock maintains that it is not adequate to generalise only over the expression '*marijuana*'—that is, to use the sentence "*If x is no more dangerous than the already legal substance alcohol, then x should be legalized.*" as an implicit premise. It would not do justice to the intuitive relevant objection "that by the same reasoning one would have argued in the nineteenth century that heroin should be legalized since it is no more dangerous than opium, which is already legal" (93). This point could only count as an objection if we generalise over both expressions, '*marijuana*' and '*alcohol*', by using the following sentence as an implicit premise: "*If a substance x is no more dangerous than another substance y which is already legalized, then x should be legalized.*" On the other hand, it would not be charitable to additionally generalise over the expression '*legal(ized)*' since "we would attribute to the argument the assumption that any substance which is no more dangerous than another substance should be given all the properties which the other

¹⁴⁴He is, however, not the first one to suggest this strategy. A predecessor is, for instance, Schwartz (1981), who introduced the criterion of generality, which demands to "add premises that are as general as possible, consistent with Fidelity and Generosity" (as cited in Burke 1985, 111). Also, Hitchcock does not consider the generalised associated conditional as an implicit premise but as an inference rule. He even goes as far as saying that "the doctrine of implicit premises is largely a myth" (Hitchcock 2002, 160). However, as he acknowledges, this standpoint is independent of his considerations about the generalisation strategy (Hitchcock 1985, 89). What is more, the question of whether the generalisation strategy can be successfully invoked to individuate arguments uniquely does not hinge on whether we regard generalised conditionals as premises or inference rules. Here, I labour from the perspective of applied formal logic and will therefore consider them as premises.

¹⁴⁵Using some elementary logic, he even shows that the strategy works with the apparent difficult case of switching implicit and explicit premise in the Socrates case. The associated conditional of the argument "*Socrates is mortal because all men are mortal.*" is "*If all men are mortal, then Socrates is mortal.*" The property of being mortal is the only expression shared by the consequent and antecedent. To generalise over this expression means to generalise over a property, which results in the sentence "*All properties that are shared by all men, apply to Socrates.*" As it turns out, this sentence is logically equivalent to the intuitively correct implicit premise "*Socrates is a man.*" If Socrates is a man, then it is true that he has every property that every man has. And since it is tautologically true that the property of being a man is shared by all men, the sentence "*Socrates is a man.*" follows from the generalised associated conditional.

substance has” (93), which is obviously false. Independent of how you judge Hitchcock’s interpretation of which generalisation should be regarded as the implicit premise, the small example establishes one important point: If we adhere to the generalising strategy, there can be cases in which there is not only one but different candidates of generalised conditionals from which we have to choose. According to Hitchcock, this has to be done by considering our background knowledge and applying the principle of charity: An “enthymeme implicitly assumes a universal generalisation of its associated conditional over its repeated content expressions, in fact, the maximal generalisation consistent with *plausibility*” (87, emphasis added). That is, the analyst should apply the principle of charity to choose among different possible generalisations.¹⁴⁶

Let’s grant that the strategy to generalise the associated conditional is appropriate to reconstruct arguments. How can we use this strategy to guide argument individuation? The problem with using the associated conditional was that every partition of supporting reasons or premises for a claim could be equipped with its own associated conditional. We saw that using charity to choose among them could not provide any grounds to prefer an interpretation as a multiple argumentation over an interpretation as a co-ordinative compound argumentation. The question is now whether generalising the different associated conditionals provides any reasons to prefer one granularisation over the other. If we grant that generalising is, in principle, the correct strategy, we are at least not confronted with the above-discussed problem of using the associated conditional; we would not compare the plausibility of different associated conditionals but different suggestions of how to generalise them. Depending on the criteria to pick an adequate generalisation of a particular associated conditional, the generalised conditionals will differ with respect to their plausibility. Thus, comparing their plausibility might help us to individuate arguments.¹⁴⁷

Let us now apply the generalising strategy to the death penalty example. It formulates four points against capital punishment. The first reconstruction simply used the associated conditional that combined all four points into one single argument. Instead of generalising this particular conditional, we will consider reconstructing the individual points as one

¹⁴⁶Here, I only provide a shortened version of Hitchcock’s account, which distinguishes between three different kinds of indeterminacies. However, in all three cases, the principle of charity has to be used to pick an adequate generalisation of the associated conditional. While I’m quite sympathetic with Hitchcock’s account, it has a pressing hurdle. As Burke (1985) points out, every sentence which is implying the associated conditional can only be equally or less plausible than the associated conditional itself (cf. Footnote 140 on page 124). Since a generalisation of a conditional logically implies the conditional, the same applies to further generalisations. Hence, the more general the generalisation is, the less plausible it can get. The question is then how to understand his qualifications to choose the most general generalised conditional that is “consistent with plausibility” (Hitchcock 1985, 87) or “unless it would be implausible to do so” (92). The discussed example about marijuana suggests that Hitchcock not only uses the principle of charity but also invokes additional considerations—in this case, about what counts as an adequate objection to the analysed argument. This manoeuvre can, however, give rise to further criticism. Why should we choose a generalisation that enables what the analyst regards as a relevant objection? That can even conflict with charity. As (Govier [1987] 2018, 6:133–35, 6:158) points out, this kind of criticism applies to the generalisation strategy in general. It courts the danger of not doing justice to accuracy and charity by insinuating implausible premises the arguer had not intended or to which they were not committed.

¹⁴⁷Here, we do not compare different generalised associated conditionals of *one* premise set, but generalised associated conditionals of *different* premise sets—possibly with differing degrees of generalisation. Therefore, charity will not necessarily prefer an individuation that puts every premise into one complex argument (cf. Footnote 146 and the considerations of not using the associated conditional to individuate arguments).

single argument, or we will at least begin with proper subsets of the four points as a single argument.¹⁴⁸

Let's start with the third point since it can be dealt with relatively easily. It raises the impossibility of redressing a falsely accused and convicted person. Once dead, forever dead. This point seems severe enough to be interpreted as a single argument. Using the associated conditional, we would end up with the following reconstruction:

<Incorrect Verdicts>:

- (1) It is not possible to compensate persons that were subject to the death penalty as a consequence of an incorrect verdict in a murder trail.
- (2) If it is not possible to compensate persons that were subject to the death penalty as a consequence of an incorrect verdict in a murder trail, then the death penalty is not a permissible juridical option of punishment.
- justified by 1-3 --
- (3) Hence, the death penalty is not a permissible juridical option of punishment.

According to the generalisation strategy, we have to generalise this conditional over at least one repeated expression. So, how should we generalise this conditional? The whole point seems to be that the death penalty does not allow for compensation. We could repeat the same reasoning for other punishments that do not allow for compensation. Hence, it is reasonable and charitable to generalise over all types of punishments that conflict with the possibility to compensation:

<Incorrect Verdicts>:

- (1) It is not possible to compensate persons that were subject to the death penalty as a consequence of an incorrect verdict in a murder trail.
- (2) If it is not possible to compensate persons that were subject to a punishment as a consequence of an incorrect verdict in a trail, then the punishment is not a permissible juridical option.
- justified by 1-3 --
- (3) Hence, the death penalty is not a permissible juridical option of punishment.

¹⁴⁸The underlying rationale is the same as the one used by Hitchcock (1985). We want to identify general principles as implicit premises that are as general as possible without violating charity—that is, we search for sufficiently plausible and maximally general principles. To that end, starting with a finer granularisation of arguments is a good heuristic since it will tend to provide more general generalised conditionals. If these generalised conditionals are not sufficiently plausible, the analyst can go on to consider a more coarse-grained granularisation.

Now, we have to ask whether it is possible to further generalise this conditional without violating charity. Instead of talking about actions of punishment as juridical options, we could think about further generalising to all types of juridical options, be they punishments or something else. Since the point of compensation is closely linked to juridical options that harm those affected by these options, I don't think a generalisation to any juridical option would lead to a very plausible conditional. A second option is to generalise the expression 'persons'. Although it is not an explicitly repeated expression, we can think of it as being implicitly part of the consequent of the conditional. I don't think this generalisation would lead to a reasonable generalised conditional since only persons are the matter of punishment in our juridical system. In sum, the last reconstruction formulates a sufficiently general conditional, which renders the argument deductively valid and exemplifies an adequate interpretation of the third point as an individual argument.

Let's try to repeat the strategy of generalising the associated conditional for the first point. Again, we want to ask ourselves whether it is possible to reconstruct this point as a single argument. The first point stresses that capital punishment should be banned, for it does not constitute an effective deterrent for first-degree murder. Taking simply the associated conditional yields the following argument reconstruction:

<Not an effective deterrent (1st version)>:

- (1) Capital punishment for first-degree murder does not constitute an effective deterrent for these crimes.
- (2) If capital punishment for first-degree murder does not constitute an effective deterrent for these crimes, then this punishment is not a permissible juridical option.
- justified by 1-3 --
- (3) Hence, the death penalty is not a permissible juridical option of punishment.

Is there a charitable way to generalise this conditional? This reasoning claims that capital punishment is not a suitable means to reach a desirable end—namely, a decrease in the homicide rate. But how is this an argument against capital punishment? Such an argument must at least assume that the decrease in the homicide rate is an intended aim of capital punishment. A suitable generalised conditional of such an argument might suggest that a mean to an intended aim is to be forbidden if it is not effective in reaching the intended aim:

<Not an effective deterrent (2st version)>:

- (1) Capital punishment for first-degree murder does not constitute an effective deterrent for these crimes.
- (2) The intended aim of capital punishment is to be an effective deterrent for first-degree murder.
- (3) If a mean M is not effective in reaching its intended aim, then M is not a permissible measure.

-- justified by 1-4 --

- (4) Hence, the death penalty is not a permissible juridical option of punishment.

Now we have generalised the associated conditional,¹⁴⁹ but it is questionable whether this addition is charitable since the general principle expressed in premise three is not very plausible. For even if a measure is ineffective in reaching a particular intended aim, it might be effective in reaching other worthwhile aims. That, in turn, might give us a reason for pursuing the measure anyway. This line of reasoning shows that the third premise does not hold in its general form. How can we fix this defect? One possibility is to seek further implicit premises and to adjust the general principle accordingly. The problem with capital punishment is not only that it is not effective as a deterrent but that it is in itself a measure with severe consequences—namely, ending the life of a human being. One could argue that measures with such severe repercussions are only permissible if they reach their intended aims and if, additionally, the aims in question are significant enough to justify the associated downsides. The corresponding generalisation is logically weaker than the one in the last reconstruction and leads to the addition of another implicit premise.

There is another possibility to reach a charitable interpretation of the argument. Instead of interpreting the point as an objection to the death penalty, we can interpret it as an objection to an argument that seeks to justify it. To some extent, that will be a reinterpretation of the last reconstruction's second premise. This implicit premise expresses the assumption that the death penalty is intended as a deterrent for first-degree murder. Rather than using this assumption about an intended goal as a premise, we can reinterpret it by imagining someone sharing this intention and formulating an argument in favour of the death penalty—that is, a proponent of the death penalty who argues that the death penalty is an effective means to decrease the rate of first-degree murder. In other words, the former implicit premise of the reconstruction will turn out to be an implicit argument in favour of the death penalty; the former objection to the death penalty will turn out to be an objection to this implicit argument.

Figure 4.5 illustrates both possible interpretations of the first point. Both interpretations turn the first point into one individual argument. However, the second interpretation reveals an argument that was not explicitly mentioned in the text. Admittedly, the second interpretation might seem to be far-fetched. Instead of providing reconstructions of both versions here, I will confine the following considerations to the second interpretation—thus showing that it is a charitable interpretation.

At this point, we can only hypothesise to what this implicit argument amounts. In this case, charity boils down to coming up with an argument as plausible as it can get for an advocate of the death penalty—not any argument, but an argument about the death penalty being an effective deterrent. Since the text provides nothing to work with, we have to be creative and use our imagination and background knowledge. So what does it mean that some measure is effective? Surely, a measure is only effective if it helps to reach the intended

¹⁴⁹Actually, I proposed something more complicated. The added general principle is not a generalisation of the former associated conditional but logically weaker than the generalised conditional because I added another condition in the antecedent—the measure being an intended aim—and a premise that affirmed this condition. Hitchcock (1985) seems to be aware of such cases.

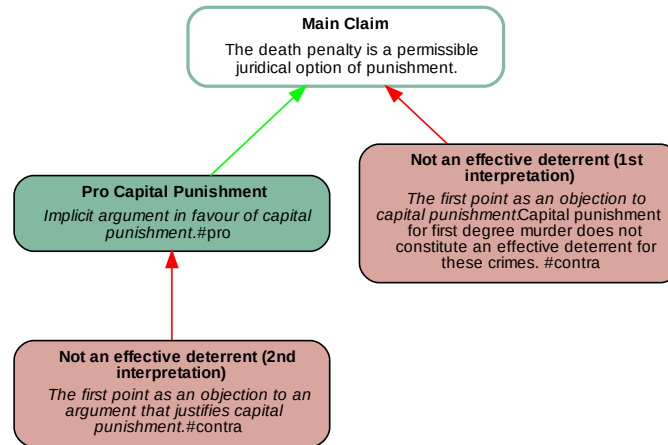


ABBILDUNG 4.5
Two interpretations of the first point.

aim. However, being effective in this weak sense is not a sufficient reason on its own. We usually demand that the measure in question be better in some way than alternative measures that also help reach the aim. When it comes to moral reasoning, comparing measures to reach some morally desirable goal should be based on a moral evaluation of the different measures. One simple suggestion is to compare the effects and side effects of the different measures. We would then, first, compare the side effects of each measure with the side effects of every other measure and should, additionally, compare the side effects of each measure with the importance of the goal. This kind of reasoning boils down to the following argument reconstruction:¹⁵⁰

<Pro Capital Punishment>:

- (1) It **prima facie** ought to be the case the rate of first-degree murder does not increase.
- (2) If death penalty is a juridical option, the rate of first-degree murder is lower as compared to when the death penalty is not a juridical option.
- (3) Death penalty as a juridical option compares in the moral evaluation of its side effects better than all its alternative to not increase first degree murder.
- (4) The moral importance of not increasing the rate of first-degree murder is not outweighed by the moral evaluation of the death penalty's side effects.
- (5) If, first, it **prima facie** ought to be the case that S and, second, A will bring about S and, third, A compares in the moral evaluation of its side effects better than all other measures that bring about S and, fourth, the moral importance of S is not outweighed by the moral evaluation of A's side effects, then A ought to be done.

¹⁵⁰This reconstruction is based on the optimal choice principle—here, as premise five. The chosen formulations are based on the argument scheme used by Betz and Brun (2016).

-- justified by 1-5 --

- (6) Hence, the death penalty ought to be a juridical option of punishment.

There are at least two possibilities to interpret the first point that there is no evidence that capital punishment is an effective means to reduce the rate of first-degree murder. The interpretation depends on what is meant by being non-effective. It could mean that capital punishment does not help to reduce the rate of homicides. In this case, the first point can be interpreted as an objection to the <Pro Capital Punishment> argument that justifies that the second premise is false. Alternatively, it could mean that other options are more effective in terms of having less severe side effects—in particular, less severe than the “side effect” of ending a human life.¹⁵¹ In this case, the first point is again an argument attacking the <Pro Capital Punishment> argument; but now it justifies the falsity of the third premise.

Let’s now consider whether it is also possible to interpret the second point as a single argument. This point stresses that the death penalty will not result in restoring the life of the murder victim. At first glance, this point is structurally similar to the first point. In the same way, as the death penalty is not effective as a deterrent, it is not effective in restoring the life of the victim. It seems that both points demonstrate that two different possible aims will not be reached or not effectively reached by using the death penalty as a juridical option. Therefore, we could reconstruct this point similarly to the first—that is, either as an argument that assumes rescuing the victim’s life is an intended aim or as an objection to another argument that justifies the death penalty by saying that the death penalty is an effective means to rescue the victim’s life.

These suggestions are, however, obviously absurd. Nobody thinks that the death penalty will change anything about the already deceased. The important point here is to understand that the absurdity can be explained by invoking the principle of charity. Both suggestions to reconstruct the argument lead to obviously weak arguments. According to the first suggestion, the second point is an objection to the death penalty; one premise of this objection would maintain that rescuing the victim’s life is an intended aim of this measure (compare the reconstruction of <Not an effective deterrent (2nd version)>). Since this is obviously false, the reconstruction would not be charitable. Following the second suggestion, the point will be interpreted as an objection to an argument in favour of the death penalty. On this suggestion, the latter argument will invoke the premise that rescuing the victim’s life is an intended aim of pursuing the death penalty (compare the reconstruction of <Pro Capital Punishment>). Again, nobody thinks that. Consequently, the second point would be an objection against an argument that nobody puts forward; it would fall prey to a so-called strawman fallacy. This second suggestion leads, therefore, to a weak or irrelevant objection. Hence, neither suggestion to interpret the second point would represent a very charitable reading of the second point against the death penalty.

We could stop here and conclude that the point is moot because nobody suggests that the death penalty is supposed to save the victim. However, I think there are other, more

¹⁵¹The term ‘side effect’ used in the reconstruction is meant to be non-evaluative and refers to all consequences of the action besides the intended aim.

charitable interpretations. As with the first point, the text does not provide very much information, and we can only be creative about it. One suggestion is to interpret this point simply as stressing the severity of ending a human life, even if it is a murderer's one. One could argue that the only thing that could outweigh the intended termination of the murderer's life would be to save the victim. The point is to set a high threshold for justifying the intended termination of the murderer's life—a threshold that obviously cannot be reached.

There are at least two different possibilities to reconstruct this idea. First, it could be understood as a necessary condition for the permissibility of the death penalty. The corresponding argument would be a direct objection to the death penalty:

<Impossibility to save the victim's life>

- (1) Only if the intended termination of a murderer's life would restore the life of the murder victim, death penalty is permissible as a juridical option of punishment.
- (2) Death penalty cannot restore life to the murder victim.
- justified by 1,2 --
- (3) Hence, the death penalty is not a permissible juridical option of punishment.

A second reconstruction takes advantage of the already reconstructed <Pro Capital Punishment> argument in favour of the death penalty. One crucial premise of this argument states that the side effects of the death penalty are negligible. Here, *side effects* is a technical term encompassing the consequences of the death penalty different from the intended aim of decreasing the rate of homicides. In this sense, the argument presupposed the death of the murderer as an acceptable side effect. The suggested charitable interpretation of the second point might, therefore, be reconstructed as an objection against this premise by stating that intentionally ending a human life is not a negligible side effect—even if it is the murderer. Only the counterbalancing side effect of saving the victim's life could outweigh ending the murderer's life. Given the reconstruction of the <Pro Capital Punishment> argument, the following reconstruction summarises the interpretation:

<Impossibility to outweigh intended ending of lifes>

- (1) The death penalty's side effects are only outweighed by the moral importance of not increasing the rate of first degree murder, if the termination of the murderer's life would restore the life of the victim.
- (2) Death penalty cannot restore the life of the victim.
- justified by 1,2 --
- (3) The moral importance of not increasing the rate of first degree murder is outweighed by the moral evaluation of the death penalty's side effects.

On this interpretation, the argument formulates an objection to the <Pro Capital Punishment> argument by justifying that the fourth premise of the argument is false.

But it also seems adequate not to interpret the second point as an argument. We could simply read it as an explanatory remark to emphasise the severity of ending a human's life, even if they are a murderer. It might be as severe as what the murderer did to their victim. This severity must be taken seriously because it is hard to find anything comparably severe that could way in the balance. The interpretation could stop at this point, which would, however, render the second point as argumentatively irrelevant—in the sense of it not being a reason or argument—or we could go further and speculate how this severity might be used to justify something else. Perhaps it is about the burden of justification. Since intentionally ending a human life is so severe in its moral weight, there must be a compelling and profound justification. On this interpretation, the second point shifts the burden of justification to proponents of the death penalty.

What about the last point that the death penalty signals that brutality is an option for the state? The reconstruction of the implicit argument in favour of capital punishment suggests different options to interpret the last point as an objection to the <Pro Capital Punishment> argument.

First, it could be interpreted as an additional single argument, which argues along the lines of the second point. Whatever the consequences are of signalling that ending the life of a murderer is a permissible option, they do not outweigh the moral importance of decreasing the rate of first-degree murder. According to this line of reasoning, the point would be that the side effects of the state expressing such brutality are not negligible in comparison to the intended aim of decreasing the rate of first-degree murder. With this interpretation, we would have an argument that objects to the <Pro Capital Punishment> argument by justifying that its fourth premise is false.

Second, we could interpret the point as justifying that the second premise of the <Pro Capital Punishment> argument is false. This premise formulates that the death penalty is an effective means to reach the intended aim. The implicit reasoning might be interpreted as follows: Signalling that ending a human life is an option for the state might result in individuals thinking that such measures are an option for them too. As a consequence, capital punishment would not lead to a decrease in the rate of murder but perhaps even to an increase. Hence, the death penalty would not be an effective means to reach the desired end.

Finally, it is possible to aggregate the last point with the first point as one complex argument. Instead of interpreting the fourth point as a separate argument, we could align its justificatory force with the first point into one argument. This interpretation is possible because the aforementioned first interpretation of the last point aims its critique to the same premise of the <Pro Capital Punishment> argument as the first interpretation of the second point. Both argue that the adverse “side effects” of the death penalty do not outweigh the moral importance of decreasing the rate of first-degree murder. Both of these points muster their justificatory pressure from the perspective of adverse side effects. We can easily imagine a proponent who does not think that either of these points is on its own a decisive objection to the death penalty but who adheres to the stance that both points are decisive in their combination. Considering only one of these side effects is not enough to

outweigh the moral importance of decreasing homicide rates, but taken together, both side effects weigh heavily in the balance of disfavouring the death penalty as a juridical option.

<Combining point 1 & 4>: Death penalty's negative side effects (ending human life & signalling that brutality is an option) outweigh the moral importance of not increasing homicide rate.

- (1) The side effects of the state signalling that brutality is an option together with the side effect of ending the murderer's life outweigh in their combined moral evaluation the moral importance of not increasing the rate of first degree murder.
- (2) There are no other positive side effects of the death penalty that could way in the balance in favour of the death penalty.
- from 1 & 2 --
- (3) The moral importance of not increasing the rate of first degree murder is outweighed by the moral evaluation of the death penalty's side effects.

We have now arrived at two different interpretations of the first and fourth point. As illustrated in Figure 4.6, we can interpret each point as a separate objection to the implicit <Pro Capital Punishment> argument or combine both points into one single argument. Are there any reasons to prefer one interpretive option over the other? Since the underlying considerations are grounded in an identified implicit argument, it is hard to answer this question by invoking the principle of accuracy. There are no linguistic cues in the text nor contextual information that could provide the needed information about the arguer's commitments or intentions. It is similarly difficult to invoke the principle of charity to decide on this issue. For one thing, it is unclear what exactly are the adverse side effects of the state signalling that brutality is an option. That makes it difficult, at least for me, to judge whether these side effects alone would already outweigh the intended goal of not increasing the homicide rate. If not, it might be more charitable to consider the side effects in their combination as an objection to the death penalty. Using the principle of charity demands using one's background knowledge. In this case, analysts might have disparate opinions as to what the more charitable granularisation is.

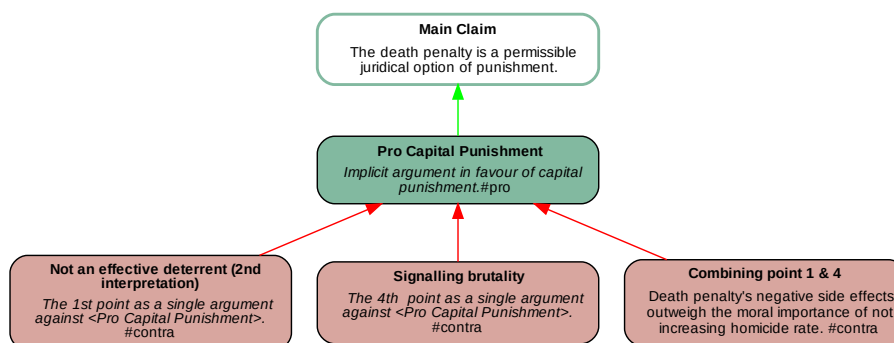


ABBILDUNG 4.6
Two interpretations of the first and fourth point.

I take it that the last identified underdetermination is paradigmatic for moral and practical argumentation since it is connected to the question of balancing. Consequentialist arguments that ponder about different means to a desired end have to commit themselves to premises about balancing the possible side effects. Typically, the exact balancing remains implicit; as a consequence, the granularisation of the argumentation will be underdetermined.

A simplified schematic example might illustrate the difficulties. Suppose an opponent of an action *A* argues against it by pointing to three negative consequences of doing *A*. The question is how to interpret the evaluation of these different consequences in detail. Does the arguer think that each is so severe as to outweigh the positive consequences of *A*, or does they think that only their combined evaluation outweighs the positive consequences of *A*? If the arguer is not explicit about this, the principle of charity generally does not help to dissolve this issue.

Presuming that the arguer meant that only the whole evaluation of all three points outweighs the benefits of *A* has one advantage in terms of charity: It makes a more modest claim with regard to the downsides of each consequence. For if each of the three consequences is taken to outweigh *A*'s positive consequences in itself, the downsides of each consequence must be considered more severe than what is needed if we rely on their combined weight against *A*. However, combining all three consequences leads to formulating one argument against *A*. And one decisive criticism against the relevance of one of these consequences might jeopardise the whole argumentation. On the other hand, making these points more independent by formulating three objections to *A*, each arguing by pointing to one of these three negative consequences, avoids this difficulty but is, in some sense, prejudicial. Now, every one of the three negative consequences must be claimed to be so severe as to outweigh the positive effect of *A* alone.

Thus, the principle of charity can be used to suggest different possibilities of granularisation. The question is how to decide between them. Admittedly, the analyst can invoke their background knowledge to dissolve this underdetermination. But then the individuation of arguments will highly depend on their assessment of the consequences' severity. The point is not so much that such an approach might lead to an interpretation that diverges from the intentions or commitments of the arguer. The more pressing issue in the context of this work is that we can expect that the evaluation of the severity of the consequences will differ from analyst to analyst; the chosen individuation of arguments will differ as well.

Without having reconstructed all mentioned arguments so far, these considerations clearly illustrate that there are different possibilities to interpret the rather short paragraph that served as an initial starting point. I argued that using the associated conditional as an implicit premise will not help the analyst to individuate arguments and opted for Hitchcock's strategy to use a generalisation of it as an implicit premise. However, using this strategy, the individuation of arguments depends on judging the plausibility of generalised associated conditionals. The example allowed for different interpretations of individuating the points as arguments. According to one specific interpretation, the argumentation contains even an implicit argument.

If analysts have to judge the plausibility of statements to decide between different interpretations, they have to involve their background knowledge for these judgements. What are the consequences of this dependence on the background knowledge for the reliability of

argument individuation? We already saw that the invocation of background knowledge can lead to a non-vanishing degree of interpretation. What I regard as a charitable interpretation of an argumentation might not be a charitable interpretation to someone else; analysts will differ in their interpretations if they do not share the same background knowledge. The key question regarding reliability is whether it is possible to guarantee that analysts have the same background knowledge. In principle, we could try to fix the necessary background knowledge for all analysts—for instance, by specifying a fixed set of sentences that serves as shared background knowledge. Whenever analysts have to reconstruct arguments to individuate them, they would have to confine the insertion of implicit premises to these sentences and their logical implications. While such or similar approaches should be pursued to decrease the degree of interpretation, I doubt they can guarantee a unique individuation of arguments. First, the necessary background knowledge might need to involve statements pertaining to the debate's specific topic. Therefore, the background knowledge would have to be specified for each analysis anew. That is, of course, not impossible, but a practical hindrance. Second, even if we try to include topic-specific knowledge, the diversity of ordinary-language arguments would demand a virtually infinite set of sentences. Both points suggest that such an approach becomes less feasible the more successfully we want to narrow down the degree of interpretation by enlarging the fixed background knowledge more and more. Apart from that, this suggestion does not tell us how to solve cases where we are confronted with a trade-off between an accurate and a charitable interpretation.

An alternative suggestion is to allow analysts to share their background knowledge and let them discuss what they regard as plausible. This strategy suggests that analysts should for each interpretational decision agree as to what the best interpretation is. This type of categorisation is known as *consensus coding* in content analysis. I already argued that this strategy is not available in the context of reliability-orientated content analysis (see 3.2.3).

After having discussed at length the difficulties of dealing with what I called horizontal underdetermination, I want to briefly discuss a slightly different case of how argument individuation is underdetermined. Consider again the following example:

We should not use CRISPR/Cas9 technologies as a therapeutic means to cure monogenetic diseases since they can have off-target and epigenetic effects that we are unaware of.

Again, we have different points being put forward for one conclusion. One is about off-target effects, and the other is about epigenetic effects. Since both points do not logically entail the conclusion, the example represents another instance of horizontal underdetermination. Therefore, the analyst must decide whether both points represent a multiple or co-ordinative compound argumentation. Here, I will sidestep this issue and interpret the argumentation as compound. Again, the task would be to reconstruct the argumentation as one deductively valid argument. To do so, we should know that epigenetic and off-target effects are possible side effects of CRISPR/Cas9 with severe consequences to the patient if used as a therapeutic means. That leads to the following first reconstruction:

- (1) CRISPR/Cas9 could lead to off-target effects.
- (2) CRISPR/Cas9 could lead to epigenetic effects.
- (3) Taken combined the risks of epigenetic side effects


```

    of off-target effects are very high.
-- from 1, 2, 3 --
(4) The risks of using CRISPR/Cas9 are very heigh.
(5) If the risks of using a bio-technology as a therapeutic means
    are very heigh, the technology should not be used as a
    therapeutic means.
-- from 4, 5 --
(6) CRISPR/Cas9 should not be used as a therapeutic means.

```

This reconstruction exhibits a more complex structure than the preceding reconstructions. It is not only a co-ordinative compound but, additionally, what pragma-dialecticians call a subordinative compound argumentation (Eemeren and Grootendorst 1984, 92). A subordinative compound argumentation moves along different inferential steps—in our case, two steps. First, the preliminary conclusion is justified by the first three premises of the argument. Second, the preliminary conclusion is used with premise five to justify the argument’s conclusion. Strictly speaking, the preliminary conclusion is not necessary for validity. If the step from the first three premises to the preliminary conclusion is deductively valid and if, additionally, the step from the preliminary conclusion and the subsequent premise to the conclusion is valid, then the conclusion logically follows from all premises combined (in that case the sentences 1,2,3 and 5). However, using preliminary conclusions fosters clarity and arguments are often explicitly formulated to involve several inferential steps.

Subordinative compound argumentation confronts us with another challenge, which I will call *vertical underdetermination of argument individuation*, or, in short, *vertical underdetermination*. The problem is that the given reconstruction could be easily rewritten as two single arguments. The preliminary conclusion would serve as the conclusion of the first argument consisting of the first four sentences and, additionally, as a premise in the second argument consisting of the last three sentences. In this case, the first argument would support the second argument.

Vertical underdetermination is not a problem for the evaluation of arguments. For the argumentation theorist, the decision between both interpretations amounts to a mere stylistic difference. However, in CAAS, the analyst is interested in counting arguments, among other things. Obviously, the different interpretations make a difference in the number of arguments.

There are some strategies to deal with this underdetermination, which, however, do not succeed in any case to disambiguate the individuation of arguments. It seems plausible to demand that separating inferential steps into different arguments is only permissible if the different inferential steps are explicitly formulated. Consequently, the argumentation above should be individuated as one argument. But how should we proceed if we find explicitly formulated inferential steps? Compare, for instance, a mere stylistic variant of the above argumentation:

We should not use CRISPR/Cas9 technologies as a therapeutic means to cure monogenetic diseases since the associated risks are too high; these include the possibility of off-target and epigenetic effects we are unaware of.

Here, the inferential step is explicitly formulated, and the provided reconstruction still represents the given argumentation. The question is whether the alternative to count two arguments instead of one is also a permissible interpretation. To my mind, both argumentations do not differ except in their explicitness. If we aim to partially operationalise argumentative complexity by counting arguments, we might, therefore, prefer to count only one argument in both formulations of the argumentation. That might be a reason to always reconstruct a subordinative compound argumentation as one argument with one or more preliminary conclusions except when we find argument indicators suggesting another interpretation. However, while this rule of individuating arguments fits the given example well, it distorts other more complex subordinative compound argumentations. There might be cases in which we find many inferential steps that are so complex on their own that it is justified or at least intuitively plausible to count some of these inferential steps as individual arguments. Therefore, the distinction between cases of a subordinate compound argumentation that should be regarded as one argument and those that should be regarded as many arguments must remain vague. Although we might often be confronted with more or less clear-cut cases, we will sometimes face cases where different possibilities to individuate arguments prevail.

4.4.3 INDIVIDUATION OF ARGUMENTS IN INFORMAL LOGIC

Besides using the tools of applied logic—remember that informal logic does not question the existence of deductive arguments but merely regards them as a marginal phenomenon—the informal logician has at least two additional options to individuate arguments: *argument schemes* and *non-strict inference rules*.¹⁵²

Argument schemes describe general patterns and forms of argumentation.¹⁵³ In Walton's account of argument schemes, each scheme is equipped with critical questions used to evaluate arguments. Argument schemes are used not only by informal logicians but also by proponents of Pragma-Dialectics and, additionally, by applied logicians—though the latter will prefer to employ valid argument schemes and do not need critical questions. That argument schemes are not only used by informal logicians is hardly surprising since inference rules and argument schemes are similar on a surface level. Both can be understood as abstract argument forms that use placeholders to be instantiated to obtain an argument

¹⁵²Here, I don't discuss the additional option to use the ARS criteria to individuate arguments. The prospects of doing so would depend on providing sufficiently precise explications of the notions of *relevance* and *sufficiency*. In discussing the individuation of reasons, I will raise some worries about explicating the relevance concept. I am only aware of two additional attempts to provide explications of sufficiency: D. Godden and Zenker (2018) propose a characterisation of the ARS criteria from a Bayesian point of view and Walton and Gordon (2015) in terms of proof standards based on the Carneades model (T. F. Gordon, Prakken, and Walton 2007). While I do not consider Bayesian accounts of argumentation here, I discuss the Carneades model in the context of analysing the macrostructure of argumentation in Section 4.4.6. For some general considerations about characterising these concepts, see J. Anthony Blair (2012). He argues that sufficiency is best understood as a placeholder and that its characterisation will depend on the specific circumstances of the argumentation.

¹⁵³Comprehensive collections of argument schemes can be found e.g. in Perelman and Olbrechts-Tyteca (1969), Walton (1996b), Walton, Reed, and Macagno (2008), Kienpointner (1992a) and Kienpointner (1992b). A general historical overview of different accounts can be found in Walton, Reed, and Macagno (2008), chp. 8 and from the pragma-dialectical perspective in Garssen (2001). For a critical discussion on the accounts of Walton and Kienpointner, see Lumer (2011).

reconstruction. In this sense, argument schemes are quasi-formal (J. Anthony Blair 2015). Consider the following two examples:

<Modus ponens>

- (1) p
- (2) If p , then q .
- justified by 1,2 --
- (3) q .

<Practical reasoning scheme - Walton, Reed, and Macagno (2008)>

- (1) I have a goal G .
- (2) Carrying out this action A is a means to realize G .
- justified by 1,2 --
- (3) Therefore, I ought (practically speaking) to carry out this action A .

The first argument scheme represents the deductive inference rule of modus ponens. If the placeholders p and q are instantiated by declarative sentences, we will obtain a deductively valid argument. The second argument scheme is an example of practical reasoning, formulated by Walton, Reed, and Macagno (2008, 333). It also uses placeholders that have to be instantiated. Instead of substituting these placeholders with declarative sentences, they must be instantiated with expressions denoting goals and actions.

An instantiation of the informal logician's argument scheme will not necessarily yield a deductively valid argument but an adequate argument reconstruction in the paradigm of informal logic. Its premises will be relevant and, in their combined form, sufficient for the conclusion—at least in a defeasible way. Raising a critical question can be interpreted as alluding to a possible objection to the argument. A critical question can point to an exception that defeats the underlying defeasible inference, it can represent an objection to the argument's premises or its conclusion, or it can question presupposed assumptions of using a scheme appropriately (Verheij 2003).¹⁵⁴ Thus, argument schemes help the informal logician to reconstruct arguments charitably and offer ways to evaluate them.

Additionally, argument schemes enable the analyst to deal with enthymemes by identifying needed implicit premises. If an ordinary-language argument fits but one argument scheme, then those sentences of the argument scheme which are not explicitly formulated can be regarded as needed implicit premises.¹⁵⁵

¹⁵⁴I will come back to the role of critical questions in the section about argumentation macrostructure.

¹⁵⁵This view is endorsed by Walton (1996a), esp. Section 7.8. This picture is, in some sense, incomplete. How helpful argument schemes are in identifying implicit premises will depend on the abstractness of the argument scheme. As Gerritsen (2001) points out in this context, "the problems of identifying unexpressed premises are often about details and peculiarities" (73). Consider, for instance, the described example of an argument scheme. If the arguer is not explicit about their goal, the argument scheme tells us that we should formulate an implicit premise that makes their goal explicit. But what their goal exactly amounts to is beyond the argument scheme itself and must be determined by additional means such as co- and context.

Argument schemes of informal logic can help to individuate arguments similarly to deductive inference rules in the paradigm of applied formal logic. In order to individuate ordinary-language arguments with the help of argument schemes, the analyst has to assess the form and structure of an explicit argument formulation and has to decide if the form and given content fit a given argument scheme and, if necessary, add implicit premises.

This similarity suggests that the role of analysing the form of ordinary-language argumentation is not so different between applied formal logic and the scheme-based account of informal logic. There are, however, important differences. Deductive inference rules are based on systematic and formal theories about deductive validity. With such a theory, the applied logician can rigorously decide on their validity requirement. Argument schemes in informal logic, on the other hand, are usually formulated in our natural language and not backed up with a formal theory that elaborates on the key notions of sufficiency and relevance.¹⁵⁶ This lack of precision gives the informal logician some kind of flexibility to formulate argument schemes. However, it often suffers from rigour in establishing the fulfilment of the informal logician's requirements of adequacy.¹⁵⁷

Are the informal logician's argument schemes somehow superior when it comes to argument individuation? At first glance, it seems that way. Argument schemes contain a much richer set of content expressions than the inference rules of most logical systems. The analyst can put them to use to identify important types of arguments. What is more, by using argument schemes, the analyst is released from thinking about different possibilities of generalising the associated conditional. Argument schemes suggest an appropriate generalisation—often as an inference principle and not as an implicit premise—right away.

However, on further scrutiny, the informal logician does not seem to have an advantage over the applied logician in connection to the underdetermination of argument individuation. At least, as I will argue, using the existing argument schemes of informal logic will not lead to a unique individuation of arguments in every case.

First of all, argument schemes are not only available to the informal logician. Hence, if they are a helpful means for the individuation of argument, the applied logician can rely

That is, the critical step of instantiating the placeholders of an argument scheme can often represent the major hermeneutical workload of identifying implicit premises. The argument scheme itself is of only little help in this regard.

¹⁵⁶However, there are efforts to provide a formal theory of argument schemes based on subjective probabilities. Hahn and Hornikx (2016) suggest using Bayesian conditioning to analyse the strength of the three prominent argument schemes (arguments from sign, argument from expert opinion and arguments from the appeal to popular opinion).

¹⁵⁷More differences abound on a deeper level. In applied logic, the used logical system determines the placeholders and the content terms of inference rules. For instance, classical logic provides declarative sentences, predicates and proper names as placeholders and only very basic logical connectives as content terms. These include formal equivalents of the natural-language conjunction, negation and the existential and universal quantifier. In informal logic, the determination of placeholder types and content terms is not so rigid simply because the informal logician is not bound to systematicity. However, as already mentioned, this can leave the matter of demonstrating the premises' relevance and sufficiency on an intuitive level only. Another distinguishing feature of Walton's argument schemes is the role of critical questions. Each of his argument schemes is equipped with questions that can be used to assess arguments critically. From the perspective of applied formal logic, however, critical questions seem to be superfluous. Instead of viewing them as something additional to the premises and the argument's conclusion, the applied logician will regard them as hints for additional implicit premises (see Lumer 2011).

on them too.¹⁵⁸ The only disadvantage in applied formal logic is the confinement to valid argument schemes. However, many of the argument schemes of the informal logician are either already valid or can easily be turned into valid argument schemes (Lumer 2011). Technically, it is even possible to turn all invalid argument schemes into valid ones by adding additional premises or weakening the conclusion. There are general worries about such deductivist manoeuvres, which are the motivating grounds for the alternative paradigm of informal logic in the first place. However, these worries are beside the point of how helpful argument schemes are for individuating arguments. Let us, therefore, consider the limits of using argument schemes to individuate arguments in more detail. There are at least four problems.

First, we might encounter arguments that cannot be rendered as instantiations of any scheme since the prominent accounts are incomplete (Lumer 2011). There are some empirical observations which seem to vindicate that claim¹⁵⁹ and a systematic reason. Though argument schemes represent “common types of argumentation” and, in particular, arguments that are “used in everyday discourse” (Walton and Reed 2003, 4), they are supposed to represent arguments that are in some sense adequate. If someone accepts the instantiated premises of an argument scheme, they also have to accept the instantiated conclusion unless they advance some critical response, for instance, by raising a critical question (Walton and Reed 2003, 4). Argument schemes are thus grounded in a normative theory that explicates criteria of adequacy for arguments.¹⁶⁰ Accordingly, arguments that do not satisfy these criteria—that is, arguments that are somehow defective—are not captured by argument schemes. However, the analyst will often stumble over such defective arguments. In such cases, they have to use other means of argument individuation, for instance, the non-strict inference rules, which have their own problems when it comes to argument individuation.

Second, a given argumentation might be interpreted as instantiating several different argument schemes. The presented argument scheme of practical reasoning is a good example. Walton, Reed, and Macagno (2008) remark that this argument scheme comes in two forms—a necessary-condition and a sufficient-condition scheme. Additionally, they propose two further related argument schemes, namely a *Value-Based Practical Reasoning* (324) and the *Argument from Goal* (325). All these variants are somewhat similar and might suit the same ordinary-language argumentation equally well. This abundance might lead to different and equally adequate suggestions of how to individuate arguments.¹⁶¹

Third, even if one and only one argument scheme fits a given argumentation, we cannot

¹⁵⁸In fact, we already utilised an argument scheme to reconstruct the implicit argument of the death penalty argumentation above. The used optimal choice principle represents a deductive argument scheme (see also Footnote 150 on page 131).

¹⁵⁹Habernal and Gurevych (2016) cite different empirical studies of applying argument schemes to ordinary-language argumentation, all of which had to create additional argument schemes ad hoc. Similarly, Lawrence and Reed (2019) describe the assignment of argument schemes to ordinary-language arguments as “an exceptionally difficult task” (16), which they back up by some empirical findings. However, they are optimistic that this situation can be improved by further amending annotation guidelines.

¹⁶⁰The underlying normative theory is often understood in terms of proof burdens. See, for instance, Freeman (2011), 193–194, Walton and Godden (2005) and T. F. Gordon and Walton (2009).

¹⁶¹In the context of argument mining, Stede and Schneider (2018) cite several human annotation efforts, which yielded only low inter-annotator agreement in the identification of argument schemes.

exclude the existence of other adequate interpretations that are not based on schemes. The reason is that the prominent theories of argument schemes, such as the accounts proposed by Walton and Kienpointer, are not based on systematic criteria of what counts as a good argument (Lumer 2011). As a consequence, it is plausible that there are often different argument reconstructions—only some of which are based on argument schemes—that perform equally well with regard to the requirements of argument reconstruction. Again, different reconstructions might lead to different suggestions of argument individuation.

While the second and the third problem can be circumvented by providing a complete system of argument schemes that is grounded in a systematic theory of argument reconstruction,¹⁶² the fourth problem applies to all accounts based on argument schemes—including deductive argument schemes. Suppose an analyst classifies a given argumentation as uniquely instantiating a particular argument scheme. Suppose further that there are no compelling reasons not to use this particular argument scheme. Then, it can still happen that the reconstruction of the argumentation is underdetermined and, thereby, the individuation of arguments. Such cases are possible because the scheme's placeholders can allow for different instantiations. Consequentialist arguments of moral and practical reasoning are good or even paradigmatic examples. These include defeasible argument schemes of informal logic and corresponding deductive argument schemes—as, for instance, the one used by employing the optimal choice principle above. These argument schemes will typically include placeholders for (intended) goals, actions, consequences of actions, intended and unintended side effects of actions and the evaluation of consequences and side effects. The specific instantiation of these placeholders might leave some leeway in argument reconstruction and argument individuation. For instance, an argumentation in favour of some action *A* by pointing to two different desired aims G_1 and G_2 that would be reached by doing *A* could be understood as one or two arguments. To put it simply, we could either put forward two arguments by saying that, first, *A* should be done because it will lead to G_1 and, second, *A* should be done because it will lead to G_2 ; or we can formulate one argument by saying that, *A* should be done because it will lead to G_1 and G_2 . In this case, it is possible to regard both aims as one aggregated aim. Since argument schemes will be silent on the issue of whether and how to aggregate different aims into more complex ones, other hermeneutical principles must be invoked to decide on these different possibilities. The same applies to the other mentioned placeholders. It is similarly possible to aggregate consequences and side effects in different ways. The principle of charity and accuracy might be employed to decide between different interpretations. We already saw in the last section that this will still leave some interpretational leeway for the individuation of argumentation.¹⁶³

The employment of argument schemes is but one possibility to individuate arguments in the paradigm of informal logic. Another suggestion is to use non-strict inference rules, which are grounded in non-strict generalisations. The following illustrative examples can be used

¹⁶²Such as the one proposed by Lumer (2011).

¹⁶³In the example of the death penalty, which I discussed there, we had different possibilities of aggregating adverse side effects into objections. If some side effects are severe on their own, it might be charitable to interpret them as a single argument against the act. If, on the other hand, the side effects are only in their combined form severe enough to provide sufficient grounds against that act, it is charitable to interpret them as one complex argument. Since the precise evaluation of side effects is often implicit, the corresponding argument individuation is underdetermined.

to compare strict generalisations in applied formal logic with non-strict generalisations in informal logic.

1. Argument 1: Socrates is mortal because he is a human.
2. Argument 2: The earth is of (roughly) spherical shape because Tom, an expert in astrophysics, asserts that the earth is of (roughly) spherical shape.

The first argument is usually reconstructed by inserting a universal generalisation as an implicit premise that renders the resulting argument valid by primarily using the above-mentioned *modus ponens* as an inference rule:¹⁶⁴

```
(1) All humans are mortal.
(2) Socrates is a human.
-- from 1, 2 --
(3) Hence, Socrates is mortal.
```

The second example is an argument from expert opinion—a broadly discussed argument type.¹⁶⁵ It would not be very charitable to reconstruct it in the same way as the first example. The corresponding universal quantification would be obviously false. Even experts can and do err on various occasions. However, it is still reasonable to trust an expert in their domain if there are no overriding reasons. This suggests the use of a non-strict generalisation—a generalisation that allows for exceptions—as an implicit premise:

```
(1) Generally, it is true what experts assert in their domain of
    expertise.
(2) Tom is an expert in the domain of astrophysics and asserts
    (on this domain) that the earth is of (roughly) spherical shape.
-- from 1, 2 --
(3) Hence, it is true that the earth is of (roughly) spherical shape.
```

This reconstruction is not valid due to the use of a non-strict generalisation. As long as no overriding evidence cast doubts on our trust in Tom's expertise, it is reasonable to believe the conclusion if we believe the premises to be true. However, we might get to know that Tom gains massive personal benefits from asserting just this statement, which might motivate us to be careful in our trust in Tom. In this case, the formulated premises are even in their combined form not a compelling reason for accepting the conclusion. Hence, the premises can be true, while the conclusion is false.

Despite this important difference, both arguments have a very similar form. Because of this similarity, Walton, Reed, and Macagno (2008) suggest describing the inference in the second example as a *defeasible modus ponens*.¹⁶⁶ Additionally, they suggest making the

¹⁶⁴Since the first premise has not precisely the form of "*If p, then q.*" but is a universal quantification of the form "*For all x: If x is F, then x is G.*" the modus ponens cannot be used immediately. Rather, we must first infer the sentence "*If Socrates is human, then Socrates is mortal.*" as a preliminary inferential step.

¹⁶⁵Walton devoted a whole book to this topic (Walton 2010).

¹⁶⁶Walton, Reed, and Macagno (2008) attribute this idea to Verheij (2003), who calls this inference scheme *Modus Non Excipiens*.

difference between both more explicit by inserting a further premise, which expresses that the arguer has no countervailing evidence that would defeat the inference.¹⁶⁷ The following formulation is but one possibility of such an argument scheme:¹⁶⁸

<Defeasible modus ponens>:

1. Generally, **Fs** are **Gs**.
2. **a** is an **F**.
3. **a** is not an exception to the rule that **Fs** are **Gs**.
- from 1,2 --
4. Therefore, **a** is a **G**.

The general strategy is then, in both cases, to insert a generalisation as an implicit premise that allows using the modus ponens or the defeasible modus ponens as inference pattern. Walton, Reed, and Macagno (2008) are optimistic “that all argumentation schemes can be cast in the [defeasible or conventional] modus ponens form” (366). Their suggestion can be used to systematise all argument schemes under a general (formalised) form and to reconstruct any other argument that is grounded in a non-strict generalisation.

In this way, finding non-strict generalisations as implicit premises is of similar importance for the individuation of arguments as the role of finding generalised associated conditionals in the paradigm of applied formal logic. The already discussed applied logician’s strategies to deal with hermeneutical underdetermination of argument individuation can be applied similarly in the paradigm of informal logic if we follow Walton’s suggestions. The only difference is that instead of finding a universal generalisation of the associated conditional, we enlarge the search area by additionally allowing non-strict generalisations as implicit premises.¹⁶⁹

Using the defeasible modus ponens to reconstruct defeasible arguments is an appealing approach to equip the informal logician’s repertoire since there might be arguments that do not fit any existing scheme. However, from the perspective of individuating arguments, this approach confronts the analyst with all the problems we already observed in the paradigm of applied logic. By enlarging the pool of premise candidates, the interpretational leeway will be the same or even increase.

¹⁶⁷This is a rather surprising strategy for the informal logician. As Verheij (1999)—from whom they take the argument scheme—points out, adding such a premise renders the argument deductively valid. The strategy of adding this particular implicit premise is, therefore, to be expected from the deductivist and not from the informal logician. The rationale for them is to formalise argument schemes. There are, however, alternatives to formalise defeasible reasoning, which typically formalise the idea of blocking the inference in the case of defeating evidence and do not suggest adding premises that render the argument valid (for an overview, see Prakken and Vreeswijk (2002), Koons (2017) and Strasser and Antonelli (2019)).

¹⁶⁸Following Verheij (2003), Walton, Reed, and Macagno (2008) prefer to insert the premise that “it is not the case that there is an exception to the rule that if P, then Q” (366). This formulation might, however, lead to misunderstandings since instead of expressing that the case at hand is not an exception, it might suggest that there is no exception at all to the rule. Obviously, the former is meant. For otherwise, there would be no difference to the case of a universal generalisation.

¹⁶⁹Using Hitchcock’s strategy to reconstruct defeasible arguments is discussed by Freeman (2011, 181–82). However, Hitchcock and Freeman would not consider the non-strict generalisation as an implicit premise but as an inference rule. But this has no bearing on using these accounts to individuate arguments.

Without reiterating these considerations by applying them to a particular case, it should be clear by now that the informal logician will encounter similar problems of argument individuation as the applied logician.¹⁷⁰ Let us, therefore, assess an alternative path that is open to the informal logician, namely to individuate reasons.

4.4.4 INDIVIDUATION OF REASONS IN INFORMAL LOGIC

The idea of distinguishing between premises in arguments and reasons is motivated by two intuitions. First, it might happen that the premise of an argument is not on its own a reason for the conclusion of the argument. Instead, it might explain how another premise constitutes a reason for the conclusion. Or, we might have to combine two or more premises to count them as one reason. However, such cases are not so relevant for distinguishing between the individuation of arguments and the individuation of reasons. For it could still be generally the case that one argument corresponds to one reason—be it in the form of one or more premises. It would, then, matter little whether we count arguments or reasons.

The second intuition is more relevant. We might also encounter arguments that combine *different* reasons. Such cases make a difference between the overall amount of reasons and arguments. The question is whether these intuitions can be substantiated with the help of argumentation theory. As we will see shortly, the distinction between convergent and linked arguments, which informal logicians often use, is a natural candidate to guide the individuation of reasons and will help to answer these questions.

The multiple-coordinated distinction is relevant for the individuation of arguments and draws primarily on the concept of sufficiency—by invoking inference rules or argument schemes. The widely discussed distinction between linked and convergent arguments, on the other hand, is relevant for the individuation of reasons and draws on the concept of relevance. However, the linked-convergent distinction was not designed to provide means for the individuation of reasons. But, as we will see shortly, it can be adapted to the need of individuating reasons. I will argue that this suggestion has shortcomings similar to those already discussed. I will first elaborate on the linked-convergent distinction described by Freeman (2011) and then explain its relevance for the individuation of reasons. Finally, I will describe some of its problems for the individuation of reasons.

The linked-convergent distinction is very prominent in the informal logic literature. Unfortunately, different authors seem to labour under different understandings as to what the distinction amounts to. For some authors, the multiple-coordinated distinction seems to be the same as the convergent-linked distinction—that is, a distinction relevant to the individuation of arguments. On this view, an argumentation for a claim is linked if the parts of this argumentation are to be interpreted as one single argument; the argumentation is convergent if its parts represent several independent arguments for the claim.¹⁷¹ According

¹⁷⁰To be fair, the described strategies are tailored to deal with other problems, namely to provide an adequate account of defeasible arguments. The individuation of arguments is not considered to be an important problem by the applied and informal logician alike.

¹⁷¹This understanding can be attributed, for instance, to Ralph H. Johnson and Blair (1994) who understand convergent arguments as consisting of different premise-sets which are independent of each other in supporting the conclusion and suggest to think of each as a *separate* argument (38). Thomas (1986) endorsed a similar understanding. According to him, a “convergent argument is equivalent to *separate arguments* (or

to another understanding, the linked-convergent distinction is introduced to denote two different types of arguments, which are distinguished by their internal structure. On this alternative view, a linked argument is roughly an argument whose premises depend on each other in their support for the conclusion. In contrast, in a convergent argument, the premises are in some sense independent of each other.¹⁷² We will see shortly that only the latter understanding is relevant for the individuation of reasons.¹⁷³ So let us briefly review some characterisations to understand how to adopt this distinction to individuate reason.

For Thomas (1986) “[r]easoning is linked when it involves *several reasons*, each of which needs the others to support the conclusion.”¹⁷⁴ or, in other words, if it “involves the logical combination of *two or more reasons*”.¹⁷⁵ When, on the other hand, “two or more reasons do not support a conclusion in a united or combined way, but rather each reason supports the conclusion completely separately and independently of the other, the reasoning is convergent.”¹⁷⁶ But how do we know whether reasons are independent of each other in their support for a conclusion? According to Thomas, we have to ask whether the support of a reason for a conclusion depends on the truth of the other reasons for that conclusion. If the support of a reason is weakened in case another reason for the same conclusion turns

evidence coming from separate areas) for the same conclusion” (quoted by Henkemans 2000, 465, emphasis added).

¹⁷²This characterisation is in many ways provisional and needs further refinement—some of it will be provided later in this section. Here, I only want to stress that, on this view, an argument can be either linked or convergent, but not both. Later on, I will prefer to speak of linked vs. convergent premises—thereby allowing an argument to have both a subset of linked premises and another subset of convergent premises.

¹⁷³According to Henkemans (2000), this second understanding can be attributed to Govier (1992a), Copi and Cohen (1990) and Groarke, Tindale, and Fisher (1997). The non-uniform use of the distinction can, unfortunately, lead to misunderstandings. For instance, Walton (1996a) discusses several suggestions for explicating the linked-convergent distinction. He explicitly speaks of linked and convergent *arguments*; hence, he seems to understand the linked-convergent divide as a question about the internal structure of individual arguments. But then he is quite explicit that the distinction between linked and convergent arguments is the same as between multiple and co-ordinated argumentation (114–15). Consequently, he seems to conflate criteria tailored to distinguish between linked and convergent arguments with those used to distinguish multiple from co-ordinated argumentation. There are several explanations for the prevailing ambiguities. Different informal logicians have a different understanding of how to individuate arguments. For instance, Copi and Cohen (1990) think that “the number of conclusions determines the number of arguments” (19–20, cited by Henkemans (2000), 458); for them, everything that is put forward as a justification for one claim is part of one argument, in contrast to the terminology suggested here that allows having different arguments in favour of one conclusion. As Freeman (2001) observes, there is also a difference in what is referred to by the terminus ‘argument’ between informal logicians and proponents of Pragma-Dialectics, which might contribute to the confusion. According to him, the advocates of Pragma-Dialectics refer to what is usually called a premise as an argument. On this interpretation, Pragma-Dialectics lacks a proper terminus for the whole premise-conclusion nexus what is here called an argument. Even if this is only a misfortunate interpretation of the Pragma-Dialectic’s understanding of the term ‘argument’, it suggests that contradicting terminologies can explain some of the confusion. Freeman (2011) provides an additional explanation. According to him, informal logicians such as Thomas (1986) and Yanal (1991) fail to distinguish between justificatory sufficiency and justificatory relevance; thereby, they conflate questions of individuating reasons from questions of individuating arguments. For more information on the terminological differences, compare Henkemans (2000) and Hitchcock (2015). I side here with Henkemans (2000), Freeman (2001) and Hitchcock (2015), who emphasise that the multiple-coordinated distinction is about the multiplicity of arguments for a claim and the linked-convergent distinction about the internal structure of arguments.

¹⁷⁴Cited by Freeman (2011), 90.

¹⁷⁵Cited by Freeman (2011), 90 and Henkemans (2000), 456, emphasis added

¹⁷⁶Cited by Freeman (2011), 90 and Henkemans (2000), 456.

out to be false, then the former reason is not independent of the latter. In such a case, the reasoning is considered linked.

Thomas' account is, however, without further adjustments, not helpful to individuate reasons. For one thing, he labours under the assumption that convergent reasoning has to be interpreted as representing different arguments and that it is not a specific type of the internal structure of single arguments.¹⁷⁷ Additionally, convergent and linked reasoning do not differ in their multiplicity of reasons in his characterisation. While he invokes the notion of reason, he describes both linked and convergent reasoning as involving more than one reason. The difference is whether these reasons need to be combined. He obviously presupposes an individuation of reasons.

Walton (2009) provides a very similar explication. For him, a "linked argument is one where the premises work together to support the conclusion, whereas in a convergent argument each premise represents a *separate reason* that supports the conclusion" (2, emphasis added). This characterisation is more promising for our purpose since it suggests that the difference between a linked and a convergent argument is connected to the multiplicity of reasons. Whereas in a convergent argument, we find more than one reason, the premises in a linked argument can be regarded as one reason, combining the premises in their support for the conclusion. However, he then goes on to explain that "[a]rguments fitting the form of an argumentation scheme are linked because all of the premises are needed to adequately support the conclusion" (2). On this view, arguments that instantiate an argument scheme are linked. So, are there any convergent arguments then? It would be at least odd that Walton is sceptical of providing argument schemes for convergent arguments. Since argument schemes can be regarded as means to individuate arguments, this might even suggest that he rather considers a convergent argument as being equivalent to different arguments for the same conclusion.¹⁷⁸

So far, we have identified two relevant points we have to be careful about if we want to connect the question of reason individuation with the linked-convergent distinction: First, the distinction needs to be about the internal structure of arguments and not the multiplicity of arguments. Second, a difference in the internal structure should imply a difference in the multiplicity of reasons.¹⁷⁹

There is one additional point to bear in mind. Both accounts draw on the notion of support without exploiting the distinction between sufficiency and relevance—where *relevance* roughly means providing some support for the conclusion, and *sufficiency* means providing sufficient support. According to Freeman (2011), this is the key to providing a characterisation of the link-convergent distinction that is different from the multiple-coordinated distinction. To decide whether we have one or more arguments for a

¹⁷⁷Cf. Footnote 171.

¹⁷⁸This is a likely interpretation considering that he, additionally, invokes the notion of *adequate* support—what I take to mean *sufficient support*. As we saw in the last section, the notion of sufficient support is usually involved in the question of argument individuation. As I already noted, Walton is at least unclear, if not inconsistent, about his understanding of the linked-convergent distinction (cf. Footnote 173).

¹⁷⁹These desiderata exclude, for instance, using the account of Yanal (1991) without adapting it for the purpose of reason individuation. For him, "dependent reasons form one argument; independent reasons form multiple arguments" (139). That is, he presupposes individuated reasons and ties the linked-convergent distinction to the individuation of arguments.

claim—that is, whether we have a multiple or a co-ordinated argumentation—we have to ask which premises can be combined in such a way that they provide sufficient support for the conclusion. In other words, the concept of *sufficiency* is relevant for the individuation of arguments. The concept of *relevance*, on the other hand, can be used to individuate reasons: We have to ask which premises are relevant on their own for a conclusion to decide whether we have one or more reasons. On his account, linked premises “must be taken together or are intended to be taken together to see why we have *one relevant reason* for the conclusion” (94, emphasis added). In a linked premise set, only the combination of all premises constitutes one reason, or some of the premises simply explain why the other premises of the set constitute one reason. In the latter case, we can consider some of the premises to answer the question of why the remaining premise or premises are relevant. The essential point is that a linked premise set does not represent more than one reason. In contrast, a convergent set of premises contains two or more premises that are justificatory relevant on their own to the conclusion. Each of these premises “gives a separate piece of evidence for the conclusion. We can imagine each one, after the first, being given to answer the question—Can you give me an *additional reason*” (94, emphasis added)? Their independent relevance to the conclusion renders them as separate reasons.

Freeman provides the following illustration of convergent premises:

Cigarette smoking poses a substantial health risk to the smoker. It also poses a risk to those nearby who must breathe the smoke secondarily. Therefore people should not smoke cigarettes. (Freeman 2011, 91)

Both points are brought forward against smoking. According to Freeman, they are independent of each other. Neither the first nor the second is needed to realise the relevance of the other. Although both points have to do with risks, they represent independent points because there is an important difference between these risks. The first risk concerns the smoker—that is, the person who can decide on the action that poses the risk; the second risk concerns third parties—that is, persons who have to live with the smoker’s decision. It is even plausible to regard both points as two different arguments. The risks the smoker poses to other persons weigh heavily in the balance against smoking. Even if the smoker does not care about their health, the second point might be sufficient to ban smoking. So clearly, these points are independent in some sense. This independence suggests interpreting the argumentation as convergent reasoning.

The paradigm case for a linked premise set are arguments that instantiate the *Pure Hypothetical Syllogism*, such as the following:

- (1) If argumentation theory helps to individuate reasons in a unique way, it can reduce the interpretational leeway of identifying argument structure.
- (2) If argumentation theory can reduce the interpretational leeway of identifying argument structure, it should be employed in content analysis.
- from (1) & (2) --
- (3) Hence, if argumentation theory helps to individuate reasons in a unique way, it should be employed in content analysis.

These premises are linked because it is hard to see why each is relevant without being aware of the other. Even if we consider the first premise as relevant on its own, we could explain its relevance with the second premise if someone doubts its relevance. In other words, we can consider the second premise as an explanation of why the first one is relevant; therefore, we should consider both premises as interdependent points—that is, convergent reasoning.

Freeman regards the linked-convergent distinction as vital because it makes a difference in argument evaluation. If the premises of an argument are linked, they stand and fall together. Questioning one premise amounts to questioning the whole reason and, thereby, the whole argument. If, on the other hand, premises are convergent, each represents one reason. Questioning one premise does not necessarily undermine the remaining reasons. In this case, it is possible to re-evaluate the remaining premises to decide whether they are sufficient in their support for the conclusion.¹⁸⁰

Our concern is not argument evaluation but the individuation of reasons. For Freeman's account, the multiplicity of reasons is connected to whether the premises of an argument are linked or convergent. Providing criteria to decide on the internal structure of arguments might, therefore, help to decide on the individuation of reasons. However, since his focus is argument evaluation, he is primarily concerned with reasons within arguments. That is, he presupposes already individuated arguments and then asks whether the premises of an argument are linked or convergent. We, on the other hand, are interested in criteria of reason individuation. What is more, we investigate criteria to individuate reasons as an alternative to individuate arguments since the latter is beset with some problems (4.4.3). Hence, we should not presuppose any individuation of arguments. Fortunately, this will not require any real adaption of Freeman's characterisation since he primarily speaks of linked and convergent premise sets. We simply shouldn't assume that these belong necessarily to the same argument.

The leading idea of individuating reasons is to ask whether premises are independently relevant for their conclusion. If they are, we count them as separate reasons and otherwise as one reason. As seen above, there are examples where we agree on our intuitive judgement as to whether premises are independently relevant. However, we must characterise the notion of independent relevance in more detail to assess its prospect to minimise hermeneutical underdetermination.

Consider now the following variant of the above case about smoking:

Cigarette smoking poses a substantial health risk to the smoker. I do have other means for relaxation. Therefore, I should not smoke cigarettes.

Again, two points are mentioned as an objection to smoking. My intuition tells me that, taken together, both points are relevant to the conclusion. But are they relevant on their own? According to Freeman, both points should be considered as one reason since the first point can explain why the second point is a reason: "I should not smoke cigarettes because I have other means for relaxation!"—"I don't understand! Why? Why is that relevant?"—"Smoking cigarettes is very unhealthy, you know?"

¹⁸⁰See Freeman (2011, 89 and 129).

But in some sense, both points are independent of each other. Even if I am unaware of the health risks of smoking, I might know that people usually smoke as a means of relaxation. In isolation, the second point tells me there is an alternative to the speaker. Given the described background knowledge, I can understand what the second point has to do with the speaker's conclusion. Both sentences share a subject matter ("means for relaxation"), and the relevance notion in question is therefore called *topical relevance* (Walton 2006, 270). The simple example suggests that Freeman invokes another relevance concept—namely, the notion of *positive probative relevance*. Two statements are probatively relevant to each other if one supports the other or can be used to cast doubt on the other (Walton 2006, 270). Accordingly, it is common to distinguish between positive and negative probative relevance. We will say that a "statement A is positively relevant to another statement B if and only if the truth of A counts in favour of the truth of B. This means that A provides some evidence for B or some reason to believe that B is true" (Govier 2013, 148). While the second point seems topically relevant independent of the first point, it lacks positive probative relevance if we consider it in isolation. We couldn't say whether it counts against or in favour of smoking. Only by combining it with the first point do we arrive at a probatively relevant point. On this view, the second point is not probatively relevant in an independent way to the conclusion.¹⁸¹ It, therefore, does not constitute a reason on its own according to the suggested criterion.

Freeman invokes the notion of probative relevance to show the convergent structure of two other argument types, namely arguments that instance the *rule of conjunction* and arguments from *enumerative induction*.¹⁸² The former is the deductively valid inference from two premises to their conjunction:

- ```
(1) p.
(2) q.
-- from 1,2 --
(3) p and q.
```

Both premises are independently relevant because each conjunct "gives us 'half' of the information we need for [the conjunction 'p and q']" (Freeman 2011, 93). Each premise provides, independently of the other, some support for the conclusion. Accordingly, we would have to count two reasons. Indeed, combining both premises gives us conclusive evidence, and that is a good reason to combine both premises into one argument. However, that does not affect whether each gives independent support for the conclusion.

From there it is only coherent to regard inductive generalisations as convergent. Consider the following example:

<sup>181</sup>In the following, I will, for the sake of simplicity, tend to use the term *relevance* for *positive probative relevance*.

<sup>182</sup>These cases are paradigmatically convergent given a shared understanding of what the linked-convergent distinction amounts to. Walton (1996a) agrees with Freeman and regards enumerative induction as convergent. In contrast, Thomas (1986) and Yanal (1991) intuitively consider it to be linked. It is also here that the diverging intuitive judgments suggest that argumentation theorists have different views about what intuitions the linked-convergent distinction is supposed to capture (see also Footnote 173).

```

(1) Crow 1 is black.
(2) Crow 2 is black.
(3) ...
(n) Crow n is black.
-- from 1-n --
(n+1) All crows are black.

```

In contrast to the rule of conjunction, such arguments are not deductively valid since the population referred to in the conclusion does contain more subjects than the argument refers to in the premise set. Otherwise, we would simply have an argument that successively applies the rule of conjunction. But the argument also shares some important features with the rule of conjunction. In particular, we could say that each premise does lend the conclusion some support independent of the other premises. For Freeman's account, the argument is thus convergent, and each premise can be regarded as an individual reason.

The suggested criterion of identifying reason draws on our intuitive judgement of whether a premise is independently probatively relevant for a conclusion. In particular, we analysed the preceding examples without applying systematic criteria to judge whether a statement provides some support for the conclusion. I will, therefore, refer to the suggested criterion as the *intuition-based criterion to individuate reasons*. The preceding examples might suggest that the intuition-based criterion gives the analyst a tool to individuate reasons in an unambiguous way. There are, however, cases that are not so easily categorised. Consider the following example:<sup>183</sup>

```

(1) The child is either in the nursery or playing outside.
(2) The child is not in the nursery.
-- from 1,2 --
(3) Hence, the child is playing outside.

```

Are both premises independently relevant? At least, my intuitions go both ways. Surely, we can say that both premises provide some independent support for the conclusion. We could say that the first premise provides a support of 0.5 to the conclusion on its own. If there are only two possibilities—as the first premise states—and if we do not know anything else, each has a chance of 0.5 for its realisation. A similar consideration would grant the second premise some support of its own to the conclusion—albeit a smaller one since eliminating the nursery alone leaves more than two possibilities as to where the child could be. However, there is also a clear intuition to regard these premises as one reason. If someone offers the second premise as a reason for the conclusion, questioning the relevance of the premise seems to be a usual reaction since there could be many other places the child could be. Providing the first premise as an answer in such a case seems a natural way of explaining the relevance of the second premise.<sup>184</sup>

<sup>183</sup>This is a slight adaptation of an example that Yanal (1988) uses to explain his account of the linked-convergent distinction.

<sup>184</sup>Yanal uses these intuitions to develop his criterion to distinguish linked and convergent arguments in Yanal (1988), Yanal (1991) and Yanal (2003). In contrast to Freeman, the rule of conjunction and inductive generalisation are linked according to this account. For a profound criticism of this theory, see Goddu (2003).

Freeman is aware that the intuition-based criterion is, to some extent, vague. What is needed is a further elaboration of the relevance notion.<sup>185</sup> So far, we have understood relevance in a very intuitive way: A premise counts as (independently) relevant to a conclusion if it lends the conclusion some support (on its own). However, we did not further elaborate on the concept of support.

One way to explicate this concept in more detail is to decide whether this notion is to be understood as a quantitative measure or a mere binary relation. In the former case, we aim to provide numerically precise numbers that determine the support; in the latter case, we want to provide means to decide whether a statement supports another (or not). Either way, we would then need to formulate criteria to determine the support of specific statements for a conclusion. Proceeding in this way, we would arrive at a semantical account of relevance since it grounds relevance in the notion of support by providing suggestions of what support means.

However, Freeman prefers to formulate a syntactical characterisation of independent relevance. In contrast to a semantical account, syntactical criteria of relevance are applied by analysing the grammatical or rather logical form of statements to decide on their relevance rather than analysing their meaning or content. Freeman proceeds in the following way. First, he introduces the notion of relevance not as a binary relation between premises and conclusion but as a ternary relation between premises, conclusions and inference rules. Then, he introduces the notion of dependent relevance by asking whether the premises share content expressions that do not appear in the conclusion. Let us understand these ideas in more detail.

Freeman motivates his characterisation of relevance as a ternary relation by asking about the origin of our relevance judgements. Where do our intuitions come from when we judge that a particular statement provides at least some support for a conclusion? Following Peirce (1955), Freeman grounds such intuitions in our inference habits. If we have a stable disposition to infer “*Socrates is mortal.*” from the statement “*Socrates is a man.*” we will regard the latter statement as relevant to the former. The inference habit of inferring a statement of the form “*x is mortal*” from “*x is a man*” explains our intuition to regard the latter as relevant to the former. Formulating these inference habits as inference rules allows us to explicate the notion of relevance as a ternary relation in the following way:

A set of  $n$  premises is relevant to a conclusion  $C$  with respect to a set of inference rules  $\mathcal{IR}$  if there is an  $n$ -premised inference rule in  $\mathcal{IR}$  such that  $C$  can be deduced from the premises with the inference rule.<sup>186</sup>

---

<sup>185</sup>Freeman (2011, 129) does not offer vague or problematic cases for the intuitive characterisation but responds to the criticism of Walton (1996a) who argues that the relevance concept is to some extent unclear and in need of further explication.

<sup>186</sup>See Freeman (2011, 131). Surely, this criterion needs further clarification. It is hard to judge this criterion without saying more about the properties of inference sets. One question concerns the successive application of more than one inference rule. Although Freeman’s formulation explicitly demands the existence of one such inference rule in  $\mathcal{IR}$ , such a constraint is hard to justify since relevance is surely a transitive relation: If a statement  $p$  lends another statement  $q$  at least some support and a third statement  $r$  is supported by  $q$ , then  $r$  should also be supported by  $p$ . It should, therefore, be sufficient to have some inference rules in  $\mathcal{IR}$ —instead of exactly one—whose successive application warrants to infer the conclusion. But perhaps Freeman thinks that sets of inference rules are always closed so that they include such successive applications of more than one inference rule as a combined rule. A more pressing worry is that the criterion might be too weak by

The inference rules are not necessarily restricted to the rules of formal logic. They can comprise deductive inference rules and rules that describe defeasible reasoning or even fallacious reasoning. For instance, the above example can be described as employing an inference rule corresponding to a mere contingent fact and not to a logical truth: Although immortality is conceivable for humans, it is known that all humans are mortal. In light of this contingent fact, we take it for granted to infer that someone is mortal if they are human.

The next step is to formulate conditions that tell us when premises are linked using this ternary concept of relevance. The simple suggestion to regard  $n$  premises as linked if they instance an  $n$ -premise inference rule will not do. According to this suggestion, a set of  $n$  premises would count as dependently relevant for a conclusion relative to an inference set  $IR$  if there is an inference rule in  $IR$  that allows deducing the conclusion from the premises. On this view, the rule of conjunction would, however, represent convergent reasoning if  $IR$  contains at least the formal inference rules of classical logic. However, according to Freeman, the rule of conjunction should turn out to be convergent and not linked since each of the two statements ' $p$ ' and ' $q$ ' provides some support for their conjunction ' $p$  and  $q$ '. Consequently, we should regard the mere existence of an inference rule that instantiates a premise set not as a sufficient condition for the premises to be linked.

The existence of such an inference is surely necessary for premises to be linked. However, further conditions are needed for premises to be linked. Freeman motivates his criterion by pointing to a syntactical feature of the *Pure Hypothetical Syllogism*, which has the following form:

- ```
(1) If p, then q.
(2) If q, then r.
-- from 1,2 --
(3) Hence, if p, then r.
```

Here, the premises have a shared content expression—the statement q —which does not appear in the conclusion. This pattern can also be observed in another paradigm of a linked premise set, namely the *Syllogism in Barbara*, which has the following form.

- ```
(1) For all x: If x is F, then x is G.
(2) For all x: If x is G, then x is R.
-- from 1,2 --
(3) Hence, for all x: If x is F, then x is R.
```

The shared content expression  $G$  denotes a property in this argument pattern. Freeman calls these shared content expressions *mediating components* because they topically connect premises. However, the fact that premises share a content expression is insufficient to count them as linked. The mediating component must not reappear in the conclusion. The

---

allowing the insertion of irrelevant sentences to an already relevant premise set. If  $I$  is an  $n$ -premise inference rule that allows deducing a conclusion  $C$  from a premise set  $\mathcal{P}$ , we should exclude  $n + 1$ -premiss inference rules that would allow adding more or less arbitrary additional statements to  $\mathcal{P}$  to deduce  $C$ .



reason is that we easily find examples where the premises and the conclusion share a content expression while the premises are convergent. For instance, in the conjunction “*My sweater is green and I like my sweater.*” the content expression *sweater* is shared by both conjuncts. Hence, deducing this conjunction using the rule of conjunction would yield such an example. However, following Freeman, we want the corresponding premise set to be convergent.

Using Freeman’s characterisation of linked and convergent premises, we are now able to formulate a corresponding criterion to individuate reasons. The simple idea is to regard two or more premises as one reason if they are linked and as individual reasons if they are convergent. Since we do not presuppose already individuated arguments at this point, we can formulate the criterion directly without going a roundabout way by using the linked convergent distinction.<sup>187</sup>

*Criterion to individuate reasons:* A set of  $n$  premises represents one individual reason for a conclusion with respect to an inference set  $\mathcal{IR}$ , if (i) there is an  $n$ -premise inference rule in  $\mathcal{IR}$  that licences the inference to the conclusion and if (ii) the premises share a mediating component which does not appear in the conclusion. If, on the other hand, there is an  $n$ -premise inference rule and the premises do not share a mediating component or the shared mediating component does appear in the conclusion, each premise represents an individual reason.

We do not have to take any stance here on whether Freeman succeeds in providing a satisfactory characterisation of the linked-convergent distinction. Nevertheless, we have to ask ourselves whether the adopted criterion successfully narrows down the degree of interpretation regarding the individuation of reasons.

One obvious challenge is the relatedness to the set of inference rules  $\mathcal{IR}$ . The individuation of reasons hinges crucially on how we determine  $\mathcal{IR}$ . If different analysts take different inference sets as reference point, they will individuate reasons differently, and this will hamper reliability. Therefore, we need to determine a canonical set of inference rules or conditions that help us to determine a canonical set of inference rules in a specific context of analysis. This canonical set can, of course, vary from context to context. The critical constraint is that this set is the same for every analyst in a specific context of analysis. In this way, a canonical set of inference rules can help to establish reliability in reason individuation.

So, how should we fix the relevant set of inference rules? From the perspective of understanding and analysing the argumentative structure of a text, the identification of reasons shares an important feature with the identification of arguments. In both cases, we want to understand the reasoning structure from the viewpoint of the arguer. The typical problem we are trying to solve with the suggested criterion is the following: Suppose the argumentative text provides enough linguistic cues and context to categorise text segments as reasons from the perspective of the arguer. However, the granularisation of these reasons

---

<sup>187</sup>Freeman is not interested in the question of individuating reasons. He formulates a criterion for linked and convergent arguments. In his final formulation (Freeman 2011, 139), he does not even use the concept of reason or relevance but defines the notion of linked argument with the help of the syntactical criterion I borrow from him.

is underdetermined by the linguistic cues alone. For some of these text segments, there are insufficient linguistic cues to determine whether they represent one or many reasons. Now, we are asked to use the suggested criterion. Being faithful to the perspective of the arguer, or, in other words, following the principle of accuracy, demands answering the question of reason individuation by considering their inference rules and not ours.<sup>188</sup>

However, the text and the available context will often allow for different interpretations of the arguer's inference habits. An interlocutor will usually not explicitly formulate the underlying inference rule, which explains why they regard something as a reason. Furthermore, even if the context provides enough linguistic cues to determine some of the arguer's inference habits, the suggested criterion demands considering all of their inference habits because we have to check for the existence of an inference rule that satisfies the conditions (i) and (ii). If we only determined some of their inference habits, we might miss relevant inference habits for identifying reasons.

Thus, the principle of accuracy will often leave reason individuation underdetermined. Again, we might think about employing the principle of charity to narrow down the degree of interpretation further.<sup>189</sup> Following charity, we should regard the arguer as reasonable. If there is no evidence to the contrary, they will not employ inference rules that correspond to false statements. Let's suppose that by using the principle of accuracy we determined a subset  $\mathcal{IR}_{acc}$  of the arguer's inference rules  $\mathcal{IR}$  in a unique way. Which additional inference rules  $\mathcal{IR}_{char}$  should we regard as part of  $\mathcal{IR}$  by using the principle of charity?

One straightforward suggestion is to confine  $\mathcal{IR}_{char}$  to the rules of formal logic. This suggestion is charitable because it does not presume any questionable inference rules on the part of the arguer. It depicts the arguer as comprehending and using the laws of formal logic. From the analyst's viewpoint, this suggestion has the advantage of precision. There wouldn't be any ambiguities in determining the elements of  $\mathcal{IR}$ . There would be those that we identified by using the principle of accuracy ( $\mathcal{IR}_{acc}$ ) and then, additionally, the rules of formal logic.<sup>190</sup>

The question is whether this precision translates into a corresponding precision of individuating reasons. Confining  $\mathcal{IR}_{char}$  to the inference rules of logic is, in essence, the deductivist's idea. So, we might worry that it leads to similar problems we identified in the formal logician's picture of individuating arguments. Let us, therefore, follow the steps of how we would individuate reasons according to this picture. As described in the section on argument individuation, we can expect to face the following situation: An arguer explicitly formulates premises in favour of a claim without there being inference rules of formal logic in  $\mathcal{IR}_{char}$  that licence the inference from the premise to the conclusion. If we are lucky, we

<sup>188</sup>See also Freeman (2011, 132).

<sup>189</sup>Freeman (2011, 154–56) argues that the principle of charity shouldn't play any role in analysing the internal structure of arguments. According to him, the invocation of charity conflates questions of argument analysis with questions of argument evaluation. This worry is, however, irrelevant to our purpose since our aim is confined to the analysis of argumentation structure.

<sup>190</sup>Of course, there remain open questions as to the precise nature of determining  $\mathcal{IR}$ . For instance, there is not one system of formal logic but many different. So, we would have to determine a specific system or a set of systems. We should, additionally, check whether the union  $\mathcal{IR} = \mathcal{IR}_{acc} \cup \mathcal{IR}_{char}$  is free of contradictions. We should also introduce operations to combine inference rules into new ones and demand that  $\mathcal{IR}$  is closed under these operations. Here, we can grant that these questions can be answered since I aim to show that even under these assumptions, there remain problems for the individuation of reasons.

find such an inference rule in  $\mathcal{IR}_{acc}$  and could then apply the criterion formulated above to determine the individuation of reasons. However, it is plausible that in many cases  $\mathcal{IR}_{acc}$  will be relatively sparse and will not provide such an inference rule. Remember that  $\mathcal{IR}_{acc}$  contains only those inference rules, which we could uniquely attribute to the arguer.

Following the paradigm of applied formal logic, the solution to this problem is to search for further needed implicit premises that would allow identifying a suitable inference rule in  $\mathcal{IR}$ . From here on, we would proceed in the same way as the applied formal logician would reconstruct arguments. We would search for general conditionals as implicit premises that would allow deducing the conclusion from the whole set of premises—possibly together with additional implicit premises—or a subset thereof. Consequently, much would depend on how we identify the general conditionals and other implicit premises.

Without rehearsing the details from above, let us consider an abstract and simplified example to show that we would run into the same problems of underdetermination. Suppose we identified two explicitly formulated premises  $p$  and  $q$  for some claim  $c$ . Suppose further that these sentences share no content expression and that there is neither an inference rule in  $\mathcal{IR}$  that allows deducing  $c$  from  $p$  and  $q$  together nor an inference rule to deduce  $c$  from  $p$  or  $q$  individually. Our search for further implicit premises could now end in two different results: We might find some suitable generalisation of the sentence “*If  $p$  and  $q$ , then  $c$ .*” or two different suitable generalisations of the sentences “*If  $p$ , then  $c$ .*” and “*If  $q$ , then  $c$ .*” In the former case we would end up regarding  $p$ ,  $q$  and the added premise as one combined reason, because the generalised conditional would necessarily introduce content expressions that its antecedent shares with  $p$  and  $q$ .<sup>191</sup> In the latter case, both conditionals would introduce a mediating component that they share with  $p$  and  $q$ , respectively. One conditional would, together with  $p$ , instantiate one reason, and the other conditional would, together with  $q$ , instantiate another reason. Hence, depending on the result of identifying implicit premises, we will consider  $p$  and  $q$  as one reason or as two independent reasons.

One could think that these problems are not very surprising. If we remain in the picture of applied formal logic, we will stumble over its problems. After all, the idea of informal logic is to properly account for ordinary-language argumentation by broadening the perspective. Freeman is very explicit that our stock of inference rules contains more than the rules of logic. It can comprise inference rules expressed by semantical truths other than the truths of formal logic, for instance, the sentence “*If a person is the mother of a female, the latter is the daughter of the former.*”; strict generalisations that are contingently true, for instance, laws of nature such as “*Nothing can be faster than light.*” and non-strict generalisations—that is, general sentences which allow exceptions—for instance, the sentence “*Humans use language to communicate.*” So the suggestion is to extend  $\mathcal{IR}_{char}$  in this way and to allow the inclusion of such sentences into  $\mathcal{IR}_{char}$  as long as this can be considered accurate and charitable.

However, this extension will render the set  $\mathcal{IR}_{char}$  fuzzy. We can think of inference rules as being expressed by general conditionals. However, if we allow contingent and non-strict general conditionals to be inference rules, we will lack a criterion for a precise delineation

<sup>191</sup>In the case that we just add the conditional “*If  $p$  and  $q$ , then  $c$ .*” the sentences  $p$  and  $q$  would represent mediating components in Freeman’s sense. If we generalise the conditional, some content expressions will be left, which are by construction shared between the conditional and both  $p$  and  $q$ .

of  $\mathcal{IR}_{char}$ . Surely, charity demands that only plausible general statements be attributed to the arguer. However, this criterion is fuzzy in itself, and there is no additional criterion to define a canonical set of inference rules in a precise way.

Two problems concern the indeterminacy of  $\mathcal{IR}_{char}$ . The first problem concerns the question of identifying and choosing a general conditional used to reconstruct one or more premises as an argument or a reason for a conclusion. The second problem concerns the demarcation between premises and inference rules: If we identify a general conditional, we have to decide whether it acts as a premise or an inference rule.

Let us, for now, suppose that we answered the first question and face the second problem. That is, we identified the relevant general conditional but are unsure whether we should regard it as an inference rule or a premise. The question is whether this type of underdetermination will render the individuation of reasons underdetermined. In some cases, it is of no difference whether we regard a general statement as expressing an inference rule or as a premise.<sup>192</sup>

Consider the Socrates example again:

```
(1) Socrates is a human.
(2) All humans are mortal.
-- from 1,2 --
(3) Hence, Socrates is mortal.
```

Intuitively, this argument expresses one reason. Here, it does not matter whether we regard the general conditional as a premise or as an inference rule. In both cases, the suggested criterion will tell us that the argument contains one reason. If we view the general conditional as a premise, we can use an inference rule of formal logic to deduce the conclusion. In this case, both premises share the content expression *human* that does not appear in the conclusion. Hence, both premises represent one reason and not two. If, on the other hand, we view the general conditional as an inference rule, we use this inference rule to deduce the conclusion from one premise. Again, we only have one reason.

However, there are other cases in which the decision to treat the general conditional as a premise or inference rule makes a difference in the individuation of reasons. Consider as an abstract example an argument pattern that is very similar to the Socrates example:

```
(1) a is F.
(2) a is G.
(3) Entities that are F and G, are H entities.
-- from 1,2 --
(4) Hence, a is H.
```

Independent of whether the general conditional (3) is a universal or a non-strict conditional, treating it as a premise introduces the mediating components *F* and *G*, which do not appear in the conclusion. Hence, all three premises count as one single reason. If, however, we

<sup>192</sup>See Freeman (2011, 145).

regard such a conditional as an inference rule, the two remaining premises only share the proper name *a*, which also appears in the conclusion. Hence, both premises would be considered as independent reasons. This abstract example shows that reason individuation demands further criteria to distinguish between general conditionals that serve as premises and general conditionals that serve as inference rules.

A straightforward suggestion is to consider all such conditionals as inference rules.<sup>193</sup> The worry is that this simple criterion oversimplifies matters and doesn't do justice to our intuitions about the individuation of reasons. Consider again the adjusted smoking example:

Cigarette smoking poses a substantial health risk to the smoker. I do have other means for relaxation. Therefore, I should not smoke cigarettes.

I sympathise with considering the general conditional "I should give up unhealthy habits if their personal benefit can be achieved by other means" an inference rule. However, treating it as an inference rule would render the second sentence as an independent reason for the conclusion since the first two sentences do not share a mediating component that does not appear in the conclusion. The only content expression they (implicitly) share is '*cigarette smoking*'.<sup>194</sup> But this expression does appear in the conclusion. While the second sentence is topically relevant to the conclusion, I doubt it supports the conclusion on its own. Instead, the intuition is that both premises represent one single reason.

But even if we had criteria to demarcate premises and inference rules in a precise way, we would still be confronted with the problem we encountered above when we confined  $\mathcal{IR}_{char}$  to the rules of logic. Without any criteria that determine a unique set of general conditionals that we can use as inference rules or premises, the abstract counter-example from above issues a challenge. When different analysts come to different conclusions as to what the relevant general conditionals are, they may count reasons differently.

The problems of individuating reasons discussed so far correspond to what I called earlier horizontal underdetermination of argument individuation. There, we asked how to combine different premises for some claim into one or more arguments; here, we asked how to combine different premises for some claim into one or more reasons. Similar to the case of argument individuation, we can now ask whether the suggested criterion of reason individuation solves the problem of vertical underdetermination. A case of vertical underdetermination can arise because an arguer can provide higher-order reasons—that is, reasons for their reasons. The example we already used above illustrates the connected difficulties:

We should not use CRISPR/Cas9 technologies as a therapeutic means to cure monogenetic diseases since the associated risks are too high; these include the possibility of off-target and epigenetic effects we are unaware of.

How should we individuate or count the formulated text segments as reasons? One possibility is to view the associated risks as a reason to refrain from using CRISPR/Cas9 as a therapeutic means. But then, how should we interpret the last clause? Is it simply

<sup>193</sup>This is in line with Hitchcock, who argues that generalised associated conditionals are inference rules and not implicit premises. See Hitchcock (1985), Hitchcock (1998) and Hitchcock (2002).

<sup>194</sup>The second sentence can be read as '*I do have other means for relaxation than cigarette smoking*'.

an elaboration of these risks by explaining the precise nature of the risks, or should we view this clause as expressing one or two additional higher-order reasons that justify the first reason? Another possibility is to interpret the last clause as expressing one or two reasons for prohibiting CRISPR/Cas9. In this case, we might prefer to view their framing as risks simply as an explanation for the relevance of these reasons and not for an additional independent reason. We surely can consider the explanation that these effects amount to *risks that are too high* as an answer to the question of why off-target effects and epigenetic effects are relevant for deciding on the issue of CRISPR/Cas9.

Therefore, the intuition-based criterion to individuate reasons cannot help pin down one specific interpretation. We can identify one, two or three independent reasons. The more precise criterion that presupposes a defined set of inference rules  $\mathcal{IR}$  will similarly not help to solve these indeterminacies. The structure of the reasoning can be interpreted in two different ways. On the first interpretation, it is an instance of what is called a serial or subordinate argumentation. Schematically:

```
(1) p. (off-target effects)
(2) q. (epigenetic effects)
-- from 1,2 --
(3) r. (high risks)
-- from 3 --
(4) Hence, s. (prohibition of CRISPR/Cas9)
```

The first two premises justify the preliminary conclusion (3), which is then used to justify the conclusion (4). Suppose that there is an inference rule in some canonical set  $\mathcal{IR}$  that allows deducing  $r$  from the first two premises and another inference rule that allows deducing  $s$  from  $r$ . We can further assume that  $p$  and  $q$  neither share a content expression that they do not share with  $r$  nor one that they do not share with  $s$ . The only content expression that  $p$  and  $q$  share with each other is the term ‘*CRISPR/Cas9*’, which appears both in the preliminary conclusion and in the conclusion. Using the criterion for reason individuation, we can regard (1) and (2) as independent reasons for (3) and (3) as a reason for (4). Hence, according to this interpretation, we count three independent reasons. However, that interpretation depends on viewing the preliminary conclusion as a premise for the conclusion (4). Given our assumptions about the inference rules  $\mathcal{IR}$ , it is plausible to assume further that there is an additional inference rule in  $\mathcal{IR}$  that allows deducing the conclusion from the first two premises directly. The corresponding reasoning structure has the following form:

```
(1) p. (off-target effects)
(2) q. (epigenetic effects)
-- from 1,2 --
(3) Hence, s. (prohibition of CRISPR/Cas9)
```

In contrast to the first interpretation, we count only two reasons because we can deduce the conclusion directly without interpreting the clause about the risks as a preliminary conclusion or premise for the conclusion.

The problem is that the formulated criterion presupposes that we already decided what text segments are premises in contrast to, say, explanations or reformulations. Without any additional criteria to individuate premises, the given criterion will, therefore, not eliminate vertical underdetermination in all cases.

This all boils down to this: The individuation of reasons faces similar difficulties as the individuation of arguments. Depending on the argumentative text at hand, the criteria I discussed in this section will allow for different interpretations of individuating reasons. Freeman's intuition-based criterion characterises a reason as a statement or a set of statements that are independently probatively relevant for a conclusion. Statements that only explain the relevance of another statement do not constitute a reason on their own. This criterion is grounded in our understanding of relevance. However, without any further explication of this notion, there will be borderline cases that allow for different interpretations to identify reasons.

In light of this problem, Freeman suggests an additional criterion that aims at more precision. This syntactical criterion demands identifying an inference rule that warrants the inference to the intended conclusion from the given set of premises. Depending on the existence of mediating components, the given premises will turn out to represent one or several reasons. The worry with this approach is that it hinges crucially on our assumptions about the arguer's set of inference rules. The informal logician will tend to allow all sorts of statements to be candidates for inference rules. The principles of charity and accuracy will then come into play in determining the set of inference rules in a specific argumentative context. However, the set will remain blurry and fuzzy, and I see no (practically feasible) possibility of fixating sets of inference rules among different analysts in a unique way.

The problem of finding a canonical set of inference rules is similar to the applied logician's problem of fixing the background knowledge that is used to reconstruct arguments. This similarity is hardly surprising. Both approaches rely on using inference rules and demand to reconstruct arguments. The only difference is that the applied logician uses a precisely delineated set of inference rules, whereas the informal logician has a broader understanding of inference rules. But that only pushes the applied logician's problem to another place. Their challenge is a charitable and accurate determination of general conditionals as implicit premises; the informal logician's challenge is a charitable and accurate determination of general conditionals as inference rules.

But that is not to say that the individuation of reasons is with regard to underdetermination on a par with the individuation of arguments. While the worries I formulated are systematic, it is open to empirical scrutiny of how prevailing these problems are. I provided only some examples with which the discussed approaches could not deal. The empirical question of how paradigmatic these examples are and how often they occur in ordinary-language argumentation is beyond this work.

My tentative assessment can be summarised as follows: The individuation of arguments in the picture of both applied formal logic and informal logic demands applying sophisticated concepts and techniques that require in-depth training and experience. The same applies if we use the suggested syntactical criterion to individuate reasons. I argued that these approaches will still face problems of underdetermination. The intuition-based criterion to individuate reasons is based on probative relevance and asks the analyst to determine

whether a statement or larger text segment provides some independent support for a conclusion. While this characterisation is to some extent vague and will therefore face problems of underdetermination, it is simple and does not involve the introduction of sophisticated technical termini. From the perspective of minimising the degree of interpretation alone, there is, therefore, no good reason to involve advanced concepts and criteria to individuate arguments or reasons as compared to using the intuition-based criterion to individuate reasons.

#### 4.4.5 RELATION AMBIGUITY IN APPLIED FORMAL LOGIC

How do the different accounts perform with respect to relation ambiguity? Let us first understand how attack and support relations can be characterised from the viewpoint of applied formal logic. According to the general characterisation I gave,  $p$  supports  $q$  if the truth, plausibility or acceptability of  $p$  is presented as providing support for the truth, plausibility or acceptability of  $q$ . Similarly,  $p$  attacks  $q$  if  $p$  is presented as providing support for  $q$  being false.

The concepts and methods elaborated so far concern the assessment of individual arguments by focusing on the internal structure of arguments—or, in other words, by focusing on the microstructure of argumentation. The reconstruction of arguments already involves a support relation: To reconstruct an argument, the applied logician identifies text segments that are presented as providing support for the conclusion of the argument. The analysis of the macrostructure, on the other hand, concerns how different arguments are related to each other in terms of attack and support relations. There are different ways to explicate justificatory relations that pertain to the macrostructure. Here, I will briefly present Betz's (2010) theory of dialectical structures since it conforms to reconstructive deductivism. It is, therefore, particularly suitable in the context of applied formal logic to analyse the macrostructure of argumentation.

The theory of dialectical structure introduces two justificatory relations—a support relation and an attack relation. Let's start with the attack relation—that is, the question of how to understand an objection to an argument. Considering the internal structure of an argument, there are three different possibilities to object to an argument. First, one might question a premise's truth, plausibility or acceptability. Such an objection is called an *undermining defeater*. Second, one might question the inferential link between premises and conclusion, and finally, one might question the conclusion itself. Following Pollock (1995), the second type of objection is called an *undercutting defeater* and the third a *rebutting defeater*.

The applied logician tends to reconstruct arguments as deductively valid. If there is some flaw in the inferential link between the *explicitly* formulated premises and the conclusion, it will turn out to be a flaw in the *implicit* premises added to the reconstructed argument. Hence, if we reconstruct arguments as deductively valid, an objection intended to question the inferential adequacy will turn out to be an objection that questions a premise in the reconstructed argument. The theory of dialectical structure adheres to a strong form of deductivism and demands to reconstruct all arguments as deductively valid. Accordingly, there is no need to introduce undercutting defeaters in the theory of dialectical structures since apparent undercutting defeaters will turn out to be undermining defeaters after



reconstructing an argument (Betz 2010, 57).<sup>195</sup>

What about rebutting defeaters of an argument—that is, objections that contradict the conclusion of an argument? In the theory of dialectical structures, this symmetrical relation between arguments will not be directly represented as a relation between arguments but between arguments and statements. So-called argument maps visualise the macrostructure of argumentation. An argument map is a graph that encompasses nodes of two types and directed attack and support relations between these nodes (visualised by red and green arrows). Argument nodes represent arguments (coloured boxes) and statements nodes represent statements (grey boxes). A support relation from an argument node to a statement node visualises that the statement is logically equivalent to the argument's conclusion. Similarly, an attack relation from an argument node to a statement node visualises that the conclusion contradicts the statement. Hence, if the conclusion of an argument will be represented by a statement node, it is easy to visualise a rebutting defeater: A rebutting defeater  $A_2$  to an argument  $A_1$  is represented by an attack relation to the conclusion of  $A_1$ , as illustrated in Figure 4.7.

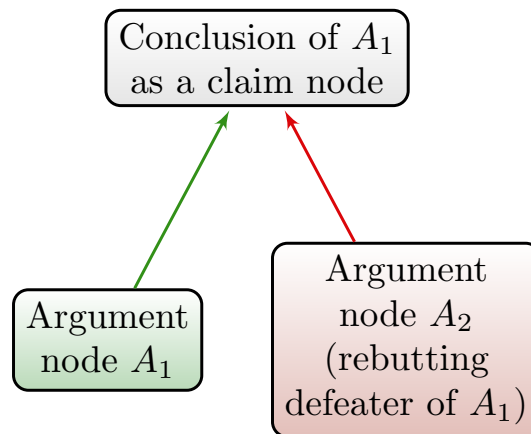


ABBILDUNG 4.7  
Visualisation of a rebutting defeater.

The same line of reasoning applies to how arguments can support each other. Given the internal structure of an argument, there are three possible targets of support relations, but only one is presented as a relation between arguments. An arguer can justify one of the premises, the inferential link or the conclusion of an argument. Only the justification of premises is represented by a relation between argument nodes. A justification of an argument's conclusion is represented by a support relation to the statement node that represents the conclusion. In analogy to the consideration of why there is no need to represent undercutting defeaters, justifications for the inferential link will, in some cases, turn out to be justifications for implicit premises and are, in other cases, practically irrelevant.<sup>196</sup>

<sup>195</sup>It is, of course, also possible that an arguer tries to question that a set of premises *logically* entails the conclusion. Such an attack relation is not represented in the theory of dialectical structures. This account labours under the assumption that such objections are irrelevant for the analysis of ordinary-language argumentation or that they are non-existent in real argumentation for all practical purposes. There is no room here, however, to defend this presupposition.

<sup>196</sup>This is motivated as follows: The applied logician will add as many premises as long as the argument

So far, I only characterised the attack and support relation in terms of what exactly they are attacking and supporting in an argument—namely, its premises. Another question concerns the precise nature of the attack and support relation. What does it mean that an argument supports another by being presented as a justification for one of its premises? Fortunately, the answer is quite straightforward. An argument is presented to justify its conclusion. Suppose now that an argument's conclusion is just another argument's premise. Since the premises of the first argument are being presented as a justification for its conclusion, they are also being presented as a justification for the corresponding premise of the second argument. The support relation can, therefore, be characterised as a semantic relation between statements: *An argument supports another argument* if the conclusion of the former is logically equivalent to a premise of the latter (Betz 2010, 56). It is similarly straightforward to introduce that attack relation by a semantic relation: We will say that an argument attacks another if the conclusion of the former contradicts a premise of the latter.

As we already saw, it is helpful to introduce these relations between argument and statement nodes (see Figure 4.8). We will say that an argument node supports a statement node if its conclusion is logically equivalent to the statement; a statement node supports an argument node if the statement is logically equivalent to one of the argument's premises. Similarly, an argument node attacks a statement node if its conclusion contradicts the statement; a statement node attacks an argument node if the statement contradicts one of its premises.

These introduced relations are expressive enough to visualise a large number of different structures (Betz 2010, 53–61). Consider, for instance, the following simple illustrations. Figure 4.9 depicts the case of two arguments having the same conclusion ( $A_1$  and  $A_2$ ) and two contra arguments ( $A_3$  and  $A_4$ ) that are rebutting defeaters—that is, arguments whose conclusions contradict the conclusion of  $A_1$  and  $A_2$ . Figure 4.10 visualises three arguments. Two of them share one premise ( $A_1$  and  $A_2$ ), which is visualised by the statement node, while a third one ( $A_3$ ) has a premise that contradicts the shared premise of  $A_1$  and  $A_2$ .

Figure 4.11 visualises a multi-level argumentation. One argument ( $A_1$ ) supports the main claim. The supporting argument  $A_1$  is attacked by another argument ( $A_2$ ), which is attacked by a further argument ( $A_3$ ). In addition, there is another argument ( $A_4$ ), which supports the argument  $A_1$ .

The theory of dialectical structures is theoretically parsimonious and conceptually precise. It invokes just the two justificatory relations of support and attack instead of introducing subcategories or other relations; it defines these relations in terms of elementary semantic relations between statements—logical equivalence and inconsistency—instead of employing more complex concepts. The question is whether these features are sufficient to eliminate relation ambiguity.

---

is not *formally* deductively valid—that is, valid according to some logical system. Consequently, many inferential connections that are perceived as sufficient support will be explicitly expressed by additional premises, which other arguments can support. For instance, inference rules that correspond to some form of contingent validity (also called *material validity*) will be reconstructed as general premises. The same holds for conceptual truths that are not captured by formal validity. Hence, justifications for the inferential link of *reconstructed* arguments are confined to justifications for the inference rules of formal logic. These justifications are considered practically irrelevant in ordinary-language argumentation since inference rules of formal logic are presupposed by arguers. Again, it is, of course, conceivable that an arguer intends to justify the formal validity of an argument. See also Footnote 195.







| Support relations                      |                                                                                  |                                                                                                                                                                                               |
|----------------------------------------|----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| from argument nodes to argument nodes  | The conclusion of A is equivalent to one of the premises of B.                   | <div>Argument A:<br/>(1) ...<br/>(2) p</div>  <div>Argument B:<br/>(1) p<br/>(2) ...<br/>(3) ...</div>     |
| from argument nodes to statement nodes | The conclusion of A is equivalent with the statement <i>p</i> .                  | <div>Argument A:<br/>(1) ...<br/>(2) p</div>  <div>Claim C:<br/>p</div>                                    |
| from statement nodes to argument nodes | The statement <i>p</i> is equivalent with one of the premises of B.              | <div>Claim C:<br/>p</div>  <div>Argument A:<br/>(1) p<br/>(2) ...<br/>(3) ...</div>                        |
| Attack relations                       |                                                                                  |                                                                                                                                                                                               |
| from argument nodes to argument nodes  | The conclusion of A is contradictory or contrary to one of the premises of B.    | <div>Argument A:<br/>(1) ...<br/>(2) p</div>  <div>Argument B:<br/>(1) Not-p<br/>(2) ...<br/>(3) ...</div> |
| from argument nodes to statement nodes | The conclusion of A is contradictory or contrary to the statement <i>p</i> .     | <div>Argument A:<br/>(1) ...<br/>(2) p</div>  <div>Claim C:<br/>Not-p</div>                                |
| from statement nodes to argument nodes | The statement <i>p</i> is contradictory or contrary to one of the premises of B. | <div>Claim C:<br/>p</div>  <div>Argument A:<br/>(1) Not-p<br/>(2) ...<br/>(3) ...</div>                    |

ABBILDUNG 4.8

Overview of attack and support relations.

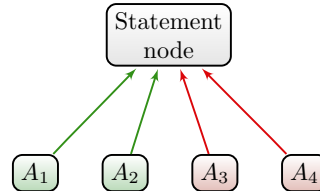


ABBILDUNG 4.9

Arguments supporting and attacking a claim.

Similarly to the analysis of single arguments, the analysis of macrostructure begins with identifying intended justificatory relations between arguments by searching for linguistic cues and considering the whole cotext and, then, moves on to analyse these justificatory relations by reconstructing identified arguments. In the section about non-reconstructive analysis, I already presented one example that can serve as a starting point for further reconstructive analysis of how arguments are related to each other (see Figure 4.4 on page 107). To determine attack and support relations between arguments, the analyst has to reconstruct the arguments in their premise-conclusion form. Once arguments are reconstructed, the attack and support relations between arguments are completely fixed. The analyst has only to identify conclusions and premises that are logically equivalent to each other or inconsistent with each other.

The question is how these introduced relations are connected to the intended justificatory relations, which are identified before reconstructing the arguments during the non-reconstructive analysis. Surely, they can fall apart. A non-reconstructive analysis might give

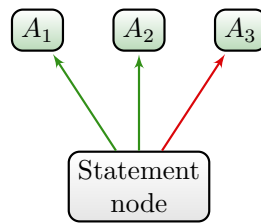


ABBILDUNG 4.10

Arguments sharing a premise or having conflicting premises respectively.

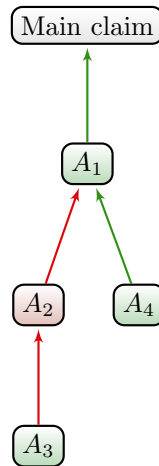


ABBILDUNG 4.11

Illustration of multi-level argumentation.

the analyst reason to speculate that the author presented an argument *B* as a justification for another argument *A*. However, after reconstructing both arguments, the analyst might find that the conclusion of *B* is not logically equivalent to a premise of *A*. Consequently, there wouldn't be a support relation between both arguments as they are defined in the theory of dialectical structures.

These possibilities might provoke concerns. But despite this possible divergence between intended justificatory relations and their defined counterparts in the theory of dialectical structures, there is nothing wrong with this theory. Instead, the introduced relations will provide additional constraints for the reconstruction of arguments and can thereby guide argument reconstruction (Betz 2010, 180–83). In the same way as the justificatory relation of deductive validity gives the analyst a constraint on identifying implicit premises, justificatory relations of support and attack give the analyst an additional constraint on how to reconstruct arguments—at least if linguistic cues uniquely determine the type, the source and the target of the intended relation. In the former case, the requirement of validity demands adding implicit premises such that the conclusion follows logically from the premise. For the latter case, let's suppose that a non-reconstructive analysis uniquely identified two arguments *A* and *B*, identified what is supposed to be justified by *B*—that is, its conclusion—and that *B* is presented as a support for *A*. In this case, the reconstruction of argument *A* should incorporate the conclusion of *B* as one of its premises. In other words, relations between arguments that are already uniquely determined by considering linguistic cues during the preparatory non-reconstructive analysis should result in corre-

sponding support and attack relations between reconstructed arguments. Arguments should be reconstructed in such a way as to reproduce these uniquely identified relations.

More important for the central question of this section is that the reconstruction of arguments can even help to eliminate relation ambiguities that appear during a non-reconstructive analysis of justificatory relations. Consider the following abstract case:<sup>197</sup> An argument  $A_1$  is presented as a justification for some claim  $C$ . Suppose there is an additional argument  $A_2$  and linguistic cues uniquely determine the role of  $A_2$  as a counterargument without revealing whether  $A_2$  justifies that  $C$  is false or whether  $A_2$  is supposed to justify that some premise of  $A_1$  is false. Thus, the hypothetical non-reconstructive analysis uniquely identifies  $A_1$  as supporting  $C$  and  $A_2$  as attacking either  $C$  or  $A_1$  (see Figure 4.12). In such a case, the reconstruction of arguments can help to disambiguate the apparent sink ambiguity of the attacking relation.

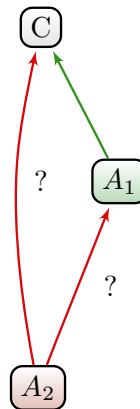


ABBILDUNG 4.12  
Example of disambiguating sink ambiguity.

To that end, the analyst should pursue two different lines of reconstructing both arguments. First, they have to reconstruct  $A_1$  and  $A_2$  most charitably such that  $A_2$  attacks  $A_1$ —that is, in such a way that the conclusion of  $A_2$  contradicts some premise of  $A_1$ . Then, they have to reconstruct  $A_1$  and  $A_2$  most charitably such that  $A_2$  attacks the claim  $C$ —that is, in such a way that both arguments have contradicting conclusions. Now, the principle of charity can be used again to choose between both suggestions. If the first line of reconstruction resulted in a more charitable reconstruction,  $A_2$  should be interpreted as attacking  $A_1$ ; if the second line of reconstruction resulted in a more charitable reconstruction,  $A_2$  should be interpreted as attacking  $C$ .

This illustrative consideration shows that the reconstruction of arguments and the invocation of attack and support relations, as the theory of dialectical structures defines them, can help to narrow down relation ambiguity. However, as in the case of argument individuation, these disambiguations depend on applying hermeneutical principles such as the principle of charity—that is, on judging the plausibility of those statements that are used in reconstructions. As I argued in the last sections, these plausibility judgements might differ from analyst to analyst. Consequently, interpretations of how arguments are related to each other

<sup>197</sup>I will provide a concrete example in Section 4.4.7.

might vary between different analysts. The reconstruction of arguments will, therefore, not always narrow down the degree of interpretation when it comes to relation ambiguity.

#### 4.4.6 RELATION AMBIGUITY IN INFORMAL LOGIC

The analysis of argumentation structure has a long history in informal logic and is intricately linked to argument diagramming—that is, the visual representation of argumentation structures (Reed, Walton, and Macagno 2007). Argument diagramming methods even precede the rise of informal logic. Reed, Walton, and Macagno (2007) trace the first use of argument diagrams back to Richard Whately in 1859, whose diagrams visualise the reasons for a conclusion as a tree structure. Wigmore (1913) used argument diagrams in the legal domain to lay out the inferential structure leading from factual evidence to hypotheses of legal matters. By now, argument diagramming techniques are used in diverse scientific contexts such as informal logic, legal reasoning and computer science. In educational contexts, they are used to teach critical thinking, informal logic and applied logic. Numerous software applications help people to construct argument diagrams. Some of these tools are designed as single-user applications, while others support the collaborative creation of argument diagrams.<sup>198</sup>

The question is to what extent these different accounts foster the analysis of macrostructure. Most of the techniques used in informal logic focus on analysing the internal structure of arguments. The corresponding argument diagrams represent the individual inferential steps by arrows that connect nodes, representing the arguments' premises and conclusions. The analytical aim is often to distinguish different types of how the premises are inferentially linked to the conclusion and to visualise these differences in the argument diagram.

A focus on the microstructure of arguments does, of course, not preclude the analysis and visualisation of macrostructure at the same time. Many accounts at least partially support the analysis of how different arguments are related. Let's take Toulmin's argument model as an illustration since it is probably the most prominent account among informal logicians.

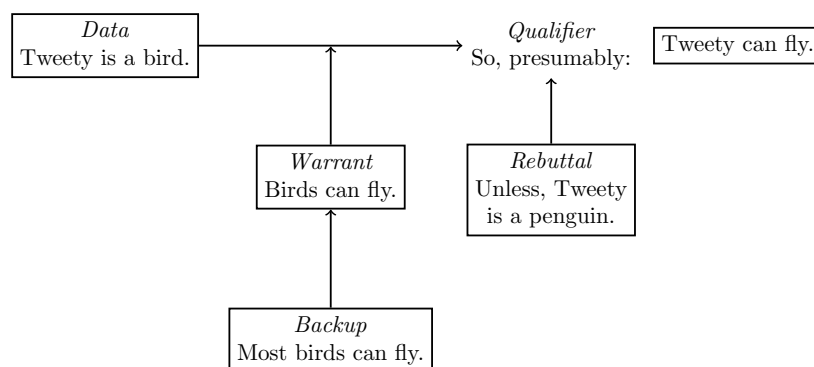


ABBILDUNG 4.13

Graphical representation of the Toulmin argument scheme.

Toulmin (1958) famously introduced the distinction between data and warrant as two different types of support for a conclusion. Data represent premises that are based on

<sup>198</sup>For an overview, see Scheuer et al. (2010).

empirical facts; warrants represent rules that licence the inference from the data to the conclusion. The importance of Toulmin's account is its explicit consideration of defeasible reasoning: The inferential link between the data and the conclusion is not confined to deductive relationships. Different modal or probative qualifiers, for instance, 'probably', 'certainly', 'necessarily' and 'presumably', indicate the premises' degree of support for the conclusion. Additionally, an inference rule might be subject to an exception, which defeats the application of the inference rule—so-called rebuttals in Toulmin's terminology.<sup>199</sup>

Consider the often-used Tweety example (see Figure 4.13). The sentence "*Tweety is a bird.*" represents the data that supports the conclusion that Tweety can fly. The general sentence "*Birds can fly.*" expresses the warrant, which is in Toulmin's account not considered a premise but rather an inference rule.<sup>200</sup> In Toulmin's terminology, a backup justifies the inference rule. For instance, the empirical fact that most birds can fly can be taken to justify inferring that something can fly if it is known to be a bird. The information that Tweety is a penguin is a rebuttal since it represents an exception to the used inference.

The Toulmin scheme is laying out the internal structure of a single argument. It visualises how a premise is inferentially related to a conclusion by an inference rule. But the scheme also visualises macrostructure to some extent. Backup and rebuttal can be considered as representing elements of additional arguments that stand in justificatory relations to the main argument. The backup can be interpreted as a supporting argument for the main argument. While it is similarly possible to regard the backup as part of the main argument, Toulmin's rebuttal cannot be interpreted as belonging to the main argument. It is not used to justify the conclusion but to attack the justification of the conclusion. However, at the same time, the Toulmin scheme is limited in terms of analysing macrostructure. While we might think of iterating the justificatory path from the data to the conclusion—that is, chaining supporting arguments—the Toulmin scheme does not incorporate any other attacks besides Toulmin's rebuttal. Arguments that attack either the data (undermining defeaters) or the conclusion (rebutting defeaters) are not designed in the Toulmin scheme.

Another important strand of analysing the macrostructure began with the pioneering work of Dung (1995). He introduced abstract argumentation frameworks as a formal theory of defeasible reasoning, which is applied in computer science and was extended in numerous ways.<sup>201</sup> In contrast to Toulmin's scheme, Dung's argumentation framework is thoroughly directed towards analysing the macrostructure of debates. In his model, arguments are atomic elements—that is, their internal structure is not considered—which can attack each other. The resulting macrostructure is represented as a graph, a so-called argumentation framework. Nodes correspond to arguments; directed edges correspond to attack relations between arguments.

Since abstract argumentation frameworks are oblivious to the internal structure of arguments, the defeasibility of arguments cannot be modelled by introducing non-deductive

<sup>199</sup>Note that Toulmin's work precedes the work of Pollock (1995) who introduced the distinction between rebutting defeaters and undercutting defeaters. Toulmin's rebuttals are ambiguous in this connection (Freeman 2011, 20). Toulmin's examples suggest that his rebuttals are to be understood as undercutting defeaters in Pollock's sense.

<sup>200</sup>In light of this, the inference is better expressed differently to make its rule character more explicit—for instance: "*Given that something is a bird, one may take it that it can fly.*"

<sup>201</sup>For an overview, see Baroni, Caminada, and Giacomin (2011).

inference rules or generics as premises. Instead, the defeasibility of arguments is captured by considering the whole macrostructure of a debate (Baroni, Caminada, and Giacomin 2011, 1). Whether an argument should be accepted depends on all arguments of the debate. In particular, the acceptance of an argument depends on whether all objections to this argument have been satisfactorily answered. Consequently, adding new arguments to a debate can change whether an argument should be accepted. This idea is given a precise form in terms of so-called acceptability or argumentation semantics. An acceptability semantics is a set of rules determining which arguments can be accepted given a specific macrostructure represented by a specific argumentation framework. The basic idea is that an argument *A* can be accepted if other arguments refute every counterargument against *A*.

Dung's account has some desirable features from the viewpoint of argumentation theory. The theory is formal and provides a precise conceptual framework to analyse and evaluate the macrostructure of a debate (Betz 2010, 43). Additionally, the macrostructure is not confined to tree structures but allows directed and non-directed cycles. However, Dung's theory alone is unsuitable for analysing the macrostructure of ordinary-language argumentation due to its abstractness (Prakken 2010, 94) and its lack of additional justificatory relations besides the attack relation (Bentahar, Moulin, and Bélanger 2010, 233).<sup>202</sup> Dung's theory abstracts away the internal structure of arguments and does not provide a semantic interpretation of the attack relation. Without further clarifications, it is doubtful whether the theory alone can guide the reconstruction of ordinary-language argumentation as an argumentation framework. In other words, an argumentation framework does not provide an adequate representation of an ordinary-language debate.<sup>203</sup> Although there are several promising extensions of Dung's original theory to overcome these restrictions,<sup>204</sup> I will use another account, which has a different origin than argumentation frameworks.

The Carneades model of argument (T. F. Gordon, Prakken, and Walton 2007) is a formal model of defeasible reasoning that can be used to reconstruct, evaluate and visualise argumentation. Similar to argumentation frameworks, the evaluative aim is to determine claims that can or should be accepted given the macrostructure of a debate. However, in Carneades the evaluation does not depend on the mere existence of supporting arguments and counterarguments, but on the internal structure of arguments, the dialectical status of statements and how allocated burdens of proof have been satisfied.

---

<sup>202</sup> Additionally, the evaluation of debates with the defined acceptability semantics will lead to implausible results if applied to reconstructions of real debates (Betz 2010, 45–47). This problem is, however, irrelevant in the context of CAAS, which is not concerned with the evaluation of debates.

<sup>203</sup> Admittedly, abstract argumentation frameworks are not designed to serve this purpose. Instead, the idea is to construct an argumentation framework based on a knowledge base, which is then evaluated in terms of an acceptability semantics to draw acceptable conclusions from the knowledge base (Baroni, Caminada, and Giacomin 2011, 1). The main rationale behind the theory is to solve conflicts in the knowledge base, which are modelled via attack relations. This focus also explains the confinement to this one relation.

<sup>204</sup> Amgoud et al. (2008) extend Dung's theory of abstract frameworks to a theory of so-called bipolar argumentation frameworks by introducing a generic support relation. Boella et al. (2010) suggest a theory that can represent a defeasible support relation that can be attacked, and Cayrol and Lagasque-Schiex (2013) suggest three different interpretations of the support relation in bipolar argumentation frameworks (deductive support, necessary support and evidential support) and discuss corresponding acceptability semantics. However, all these accounts are still abstracting away from the internal structure of arguments. Notably, Prakken (2010) developed a theory of argumentation frameworks that considers arguments' internal structure. This account can model the three main types of attacks (undermining defeaters, rebutting defeaters and undercutting defeaters). However, this account does not have a support relation.



Carneades has several features that make it particularly adept at representing the macrostructure of argumentation in line with the paradigm of informal logic. First of all, Carneades is grounded in a reconstructive analysis of argumentation structure: Arguments are represented by their internal premise-conclusion form. The reconstruction of arguments is based on Walton's theory of argument schemes and can model critical questions. Consequently, Carneades is well-equipped to capture defeasible reasoning as envisaged by informal logicians. Additionally, it provides both attack and support relations, is not confined to tree structures and can model undermining defeaters, rebutting defeaters and undercutting defeaters. Finally, though it is a formal model, it can be applied without any knowledge of its formal apparatus (T. F. Gordon 2007). I suppose there is only one missing feature from the informal logician's point of view: In contrast to, for instance, Toulmin's model, Carneades does not explicitly represent modifiers that qualify the degree of support between premises and the conclusion. However, the model is not committed to any specific interpretation of the inferential link between premises and conclusion and can, thus, be expanded to include qualifiers for inferential links.

Let us more closely understand these features to assess the model's potential to deal with hermeneutical underdetermination. In Carneades, the macrostructure of debates is represented by so-called argument graphs, which are similar to the argument maps of the theory of dialectical structures. An argument graph can have two different types of nodes: statement nodes, which represent the premises and conclusions of arguments (visualised as boxes), and argument nodes, which represent arguments (visualised as rounded boxes or circles). The model distinguishes two different types of arguments by virtue of what they justify. A con argument against a claim  $p$  provides reasons to reject  $p$ ; a pro argument for a claim  $p$  provides reasons to accept  $p$ . These different roles are visualised by using different edges from argument nodes to the respective statement nodes—visualised as green and red arrows (see Figure 4.14 for an example).<sup>205</sup> In Carneades, the internal structure of arguments is fully explicit in the graph: all premises and conclusions of arguments will be represented by statement nodes. A statement node with an outgoing green edge to an argument node corresponds to a premise, an outgoing red edge to the negation of a premise, and a statement node with an ingoing edge from an argument node represents a conclusion (or in the case of a con argument the logical complement of the argument's conclusion).<sup>206</sup>

207

The Carneades model further distinguishes different premise types, which have a crucial role in the evaluative step of the model (T. F. Gordon, Prakken, and Walton 2007, 879).

<sup>205</sup>T. F. Gordon, Prakken, and Walton (2007) use different arrowheads to visualise this difference. In the following, I will continue to deviate from their visualisation of different edge types: To enhance the readability of argument graphs, I prefer to use textual edge labels instead of disparate arrowheads.

<sup>206</sup>Therefore Carneades' argument graphs have similarities with Pollock's inference graphs (Prakken 2010, 115). Every argument node represents one inferential step. In the theory of dialectical structures, it is similarly possible to visualise the internal structure of arguments in this way. However, it is not obligatory and left to the analyst which statements they want to visualise in an argument map as statement nodes.

<sup>207</sup>To stay consistent with how I introduced the concept of argument, I use a slightly different terminology than the one adopted by T. F. Gordon, Prakken, and Walton (2007). For them, con arguments against  $p$  have the statement  $p$  as their conclusion (compare their third definition on p. 881). In this work, the conclusion of an argument is always the statement that is supposed to be justified by the premises. The conclusion of an argument against some statement  $p$  should, therefore, be a statement contrary to  $p$  and not  $p$  itself. But nothing hinges on these terminological differences.

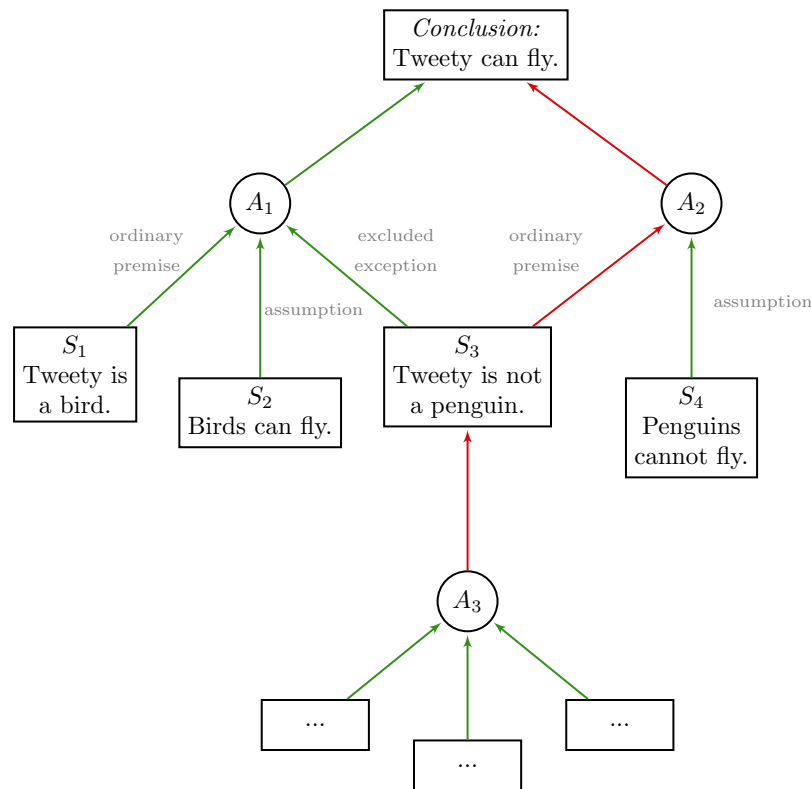


ABBILDUNG 4.14

A Carneades argument graph of the Tweety example.

Premise types determine the precise role a statement plays in an argument. Thus, they do not qualify statements per se, but the relation between statements and the arguments in which they appear as premises. The differences between these types are defined in how premises are subject to different proof standards—that is, in how they are subject to different justificatory standards for their acceptability.<sup>208</sup> In other words, to say that a premise is of a specific type is to say that the premise carries a specific burden of proof. Carneades introduces three premise types, which are represented by corresponding subcategories of edges in argument graphs: Ordinary premises are premises “that must always be supported with further grounds” (879); assumptions are premises “that can be assumed until they have been questioned” (879); and excluded exceptions are premises that hold until there is evidence to the contrary (see Figure 4.14 for some examples).<sup>209</sup> These different types are essential not only for evaluating an argument graph—which is not relevant in the context of CAAS—but also for how Carneades treats the use of argument schemes.

<sup>208</sup>The used concept of proof standard is weaker than the alternative conceptualisation in the deductive paradigm. It does not demand that what is to be proved can be logically implied by some axioms. It is simply an umbrella terminus for standards that explicate the acceptability criteria of claims given their justifications and objections in an argumentation.

<sup>209</sup>For the last type I deviate again from the author’s terminology to stay consistent with my use of the term *premise*. They introduce exceptions as premises “that do not hold in the absence of evidence to the contrary” (879). However, in my terminology, a premise is always a statement whose truth is invoked to justify the argument’s conclusion. Again, nothing hinges on these terminological decisions.

Instantiating the placeholders of an argument scheme results in a proper argument reconstruction in its premise-conclusion form. However, each of Walton's argument schemes is equipped with critical questions. The question is, therefore, how to properly formalise critical questions. Carneades simply suggests interpreting answers to critical questions that defend the argument as additional premises of the argument (T. F. Gordon, Prakken, and Walton 2007, 887). The introduced different premise types are then used to distinguish the different roles of premises and critical questions. Carneades' ordinary premises represent the premises of the argument scheme, and assumptions and excluded exceptions represent answers to critical questions. Whether a critical question is to be understood as an assumption or an excluded exception depends (in virtue of their definitions) on the allocation of proof burdens, which is a domain-specific question (T. F. Gordon, Prakken, and Walton 2007, 887).

Let us finally understand how Carneades models the basic attack relations of undermining defeat, rebutting defeat and undercutting defeat. An undermining defeater is an attack against one of the premises of an argument, which is, in Carneades, simply a con argument against the premise in question. A rebutting defeater of an argument provides reasons for rejecting its conclusion. Hence, similar to the theory of dialectical structures, a rebutting defeater of an argument can be understood as a con argument against the conclusion of the former argument. In the example of Figure 4.14, argument  $A_1$  is a rebutting defeater of argument  $A_2$  (and the other way around). Finally, an undercutting defeater is modelled via Carneades' excluded exceptions. An undercutting defeater attacks an argument by justifying that the case at hand is subject to an exception, which defeats the application of the used non-strict inference rule. Thus, an undercutting defeater is an argument that attacks a premise which excludes an exception to the inference rule ( $A_3$  in the example of Figure 4.14).

How can Carneades be used to deal with relation ambiguity? Similar to the theory of dialectical structures, Carneades relies on the reconstruction of arguments. Ambiguities concerning relations between arguments that occur during a preliminary non-reconstructive analysis of macrostructure can be dissolved by providing reconstructions that confirm one or the other preliminary interpretation. It stands to reason that this strategy inherits some of the already discussed difficulties of using argument schemes to reconstruct arguments.

I will discuss some general worries before illustrating the connected difficulties with the help of an illustrative example in more detail (4.4.7).

There are two challenges: First, the informal logician regards the distinction between undermining and undercutting defeaters as important. The category of an undercutting defeater is motivated by explaining the defeasibility of arguments in terms of non-strict inference rules. As a consequence, an argument can be defeated by showing that the case at hand is an exception to the rule—thus attacking the applicability of the inference rule rather than a premise of the argument. The challenge is to provide criteria that uniquely determine which attacks are undermining defeaters and which are undercutting defeaters. Second, models drawing on Walton's argument schemes might want to distinguish attacks that target a premise of the argument scheme from attacks that correspond to raising critical questions. Again, the challenge is to provide criteria that uniquely distinguish one from the other.

T. F. Gordon, Prakken, and Walton (2007) offer promising answers to these questions. To be more precise, they offer the same answer to both challenges. In Carneades, the distinction between “normal” premises and critical questions and the distinction between undermining defeaters and undercutting defeaters amount to categorising premises according to different premise types—namely ordinary premises, assumptions and excluded exceptions. Determining the different premise types boils down to categorising premises according to their associated proof standards.

Regarding the distinction between the premises of an argument scheme and its critical questions, Carneades is in line with the paradigm of informal logic. In informal logic, the nature of critical questions is often explicated by their dialogical role: If a proponent puts forward an argument by instantiating all premises of an argumentation scheme the “respondent is obliged to accept the conclusion. But this acceptance [...] is provisional in the dialogue. If the respondent asks one of the critical questions matching the scheme, the argument defaults and the burden shifts back to the proponent” (Walton and Reed 2003, 5). In other words, the proponent has fulfilled some initial burden of proof for the conclusion by advancing the premises of the argument scheme. They are not obliged to answer any of the critical questions until some opponent raises them.<sup>210</sup> Hence, the distinction between critical questions and premises of an argument scheme is explicated in terms of how they relate to proof standards in a dialogue. This is the same approach Carneades adopts to distinguish between “normal” premises and critical questions—in their terminology, between ordinary premises, on the one hand, and assumptions and excluded exceptions, on the other hand.<sup>211</sup>

Carneades’ treatment of undercutting defeaters is grounded in the same approach. Undercutting defeaters are distinguished from undermining defeaters by their relation to proof standards. Though the informal logician might take some issues with Carneades’ modelling of undercutting defeaters,<sup>212</sup> it answers some problems of the distinction between undermining and undercutting defeaters. The received view of undercutting defeaters is this: They attack the inferential link between premises and the conclusion and neither some premise nor the conclusion (Pollock 1995, 41). The inferential link is grounded in an inference rule, which licenses the inference from the premises to the conclusion. Consequently, an undercutting defeater can be understood as blocking the application of the inference rule. This is explained by the fact that there are non-strict inference rules; there are exceptions in which the application of the inference rule is not allowed. Tweety being a penguin is an exception to the rule that one might take that Tweety can fly if Tweety is a bird. It is possible to express such an inference rule as a non-strict general sentence—in the example: “*Birds can fly.*” The question is how the analyst can tell inference rules apart from general sentences that do not correspond to an inference rule. Surely, there are cases in which general sentences function as a premise and do not express an inference rule.

---

<sup>210</sup>Compare also Freeman (2011, 193–94).

<sup>211</sup>This is not to say that Carneades’ modelling of critical questions in terms of how they relate to proof burdens is equivalent with the informal logician’s understanding of different proof burdens. I think there are important differences. But these matters are beyond the scope of this work. Here, I want to stress that the general approach of distinguishing between “normal” premises and critical questions in informal logic is the same in Carneades: It is explicated in terms of proof burdens.

<sup>212</sup>In informal logic, the undercutting defeat is supposed to attack the justificatory relation between premises and the conclusion and not some premise of the argument (Prakken 2010, 115).

But even if we can formulate criteria that demarcate inference rules from general sentences that are premises in a more or less unique way, there is the further problem of determining the exact formulation of inference rules. Each inference rule formulates antecedent conditions, which determine whether an attack amounts to an undermining or undercutting defeater since only the affirmation of the antecedent conditions is expressed by premises. The question is, for instance, whether the inference rule is to be formulated by the sentence “*Birds can fly.*” or by “*Birds that are not penguins can fly.*” In the former formulation, there is only one antecedent condition. Something being a bird is sufficient to (defeasibly) infer that it is a bird. In this case, a counterargument grounded in Tweety being a penguin is an undercutting defeater. If the latter formulation is the correct formulation of the inference rule, Tweety not being a penguin is regarded as an additional antecedent condition, which is to be expressed as a premise. Accordingly, the discussed attack turns out to be an undermining defeater.

This illustrative example might not pose a real problem. Our, or at least my intuitive grasp of language, tells me quite clearly that penguins are exceptional birds. Therefore, the first formulation corresponds already to an appropriate inference rule. However, this simple example should not mislead us: often, these questions are beyond our intuitive grasp or at least different people will have different intuitions. Accordingly, we need precise criteria to determine the appropriate formulation of inference rules. Carneades provides at least a partial answer to that question. What distinguishes antecedent conditions from exceptions to the inference rules is determined by differences in terms of proof standards.

However, the Carneades model alone will not avoid hermeneutical underdetermination, particularly edge-type ambiguity. The model translates the distinction between critical questions and “normal” premises on the one hand and the distinction between undercutting and undermining defeaters on the other hand into a distinction between different relation types (ordinary premises, assumptions and exceptions). To determine these types, the analyst categorises statements of an argument according to their attached proof standards. However, Carneades is not equipped with any systematic account to do so. The question is whether there is such a theory that can be used in the variety of natural-language contexts the analyst intends to investigate. I will briefly come back to this question in the subsequent section.

#### 4.4.7 RELATION AMBIGUITY – A COMPARATIVE EXAMPLE

The considerations I shared so far point to problems of hermeneutical underdetermination on a theoretical level without showing how these difficulties play out when analysing the macrostructure of real debates. Let us, therefore, consider an illustrative example to understand the severity of these and possibly other problems. We have already begun to analyse the macrostructure of an example about affirmation actions (see Section 4.3.2). Given our preliminary non-reconstructive analysis, I will present a reconstructive analysis of the example in the paradigm of applied formal logic using the theory of dialectical structures and compare the results with a reconstructive analysis in the paradigm of informal logic by using the Carneades model. Instead of reconstructing the whole example, I will only consider the following small part of the example:

[1] Governments should employ measures of affirmative action (AA) to help minorities that are under-represented in higher education, business and politics (=  $C$  and  $R_1$ ). [2] There are, of course, objections [...] [5] Others argue that affirmative action can even damage the development of a country; [...] [7] This helps neither majorities nor minorities (=  $R_3$ ).

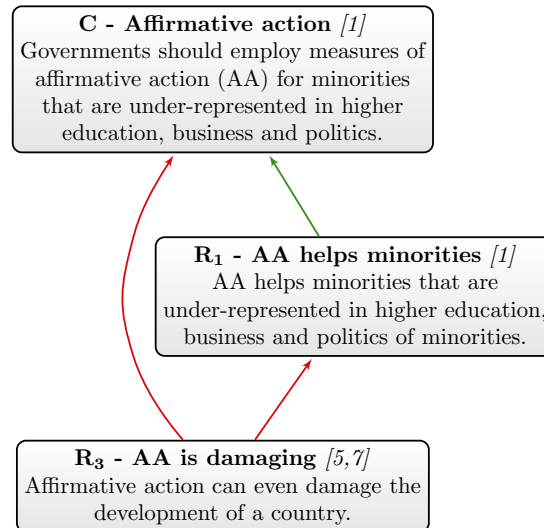


ABBILDUNG 4.15

Part of the preliminary argument map of the affirmative action example.

Our preliminary analysis of the macrostructure identified one main claim ( $C$ ) and two arguments ( $R_1$  and  $R_3$ , see Figure 4.15). I interpreted the first sentence as the main claim and argued that the first sentence could even be interpreted as formulating a reason or argument in favour of the main claim ( $R_1$ ). I went on to suggest that the seventh sentence must be interpreted as a con argument. There are, however, two possibilities to interpret the target of the argument. It can be interpreted as attacking the main claim directly or as an attack targeting the supporting argument  $R_1$ . In other words,  $R_3$  can be interpreted as an undermining defeater of  $R_1$  or as a rebutting defeater of  $R_1$ .

How can a reconstructive analysis help to dissolve this ambiguity? A simple suggestion is to use the discussed hermeneutical principles and, in particular, the principle of charity to compare different reconstructions of  $R_3$ . We would have first to reconstruct argument  $R_1$  and then reconstruct the other argument in two versions: One reconstruction that interprets  $R_3$  as an undermining defeater to  $R_1$  and one interpretation as a rebutting defeater to  $R_1$ —that is, one reconstruction of  $R_3$  with a conclusion that is contrary to one of the premises of  $R_1$  and one reconstruction of  $R_3$  that has a conclusion contrary to the main claim  $C$ .

This strategy to deal with this relation ambiguity can be generalised to all relation ambiguities, and we will follow it to understand its prospects and downsides. However, in this particular case, there is already one limitation that a reconstructive analysis cannot overcome: The identified sink ambiguity depends on what I refer to as node ambiguity. There is the question of whether the first sentence expresses a reason. Suppose both are permissible: to interpret the first sentence as already expressing a reason in favour of  $C$  and as not expressing a reason—and I don't see any conclusive reasons to exclude one of these interpretations. Suppose further that a reconstructive analysis would favour  $R_3$  as an

undermining defeater to  $R_1$ . In that case, the target of  $R_3$  would depend on our interpretive decision regarding the first sentence. Interpreting the first sentence as formulating an argument results in interpreting  $R_3$  as attacking that argument; interpreting the first sentence as not formulating an argument results in interpreting  $R_3$  as attacking the main claim. In this sense, the sink ambiguity would depend on the node ambiguity and could not be dissolved even if charity would favour one reconstruction of  $R_3$  over the other.<sup>213</sup>

Let us now proceed by using the sketched reconstructive strategy. We will begin with reconstructing the argument  $R_1$  in both paradigms. The argument is formulated in the first sentence of the text and is relatively short:

[1] Governments should employ measures of affirmative action (AA) to help minorities that are under-represented in higher education, business and politics.

The argument's conclusion is the identified main claim  $C$  (the sentence's main clause). The main point of the first sentence is to express a means-ends relationship: Affirmative action is a means to help minorities by eliminating their under-representation in higher education, business and politics. To justify the conclusion, it is not enough to simply state this relationship. The arguer must add a normative premise that attaches moral value to the aim. The elimination of underrepresentation should be regarded as a morally valuable goal. Taken combined, it is pretty clear that argument is a moral variant of a practical syllogism. Since some goal  $G$  has a high moral value and  $M$  is a means to achieve  $G$ , we are morally obliged to do  $M$ . Though that much is clear, this consideration alone is not enough to pin down a reconstruction of the argument since different argument schemes fit such a means-end reasoning. The question we need to answer concerns the precise nature of the means-ends relation: Is the means considered necessary or sufficient for the intended goal? These options are not exclusive. The important difference is that there might be alternative means to reach the goal in the latter case. The sentence itself is silent on the issue. Charity suggests, in this case, to opt for a sufficient-means interpretation. At least according to my background knowledge, it is more plausible to assume that affirmative action is one of many possible means to eliminate under-representation.

The case of the sufficient-means syllogism is more complex than the necessary-means syllogism since it demands a comparison between the different alternatives to reach the aim. The means that the argument justifies should be, in some kind of way, the optimal means among its alternatives. Fortunately, both paradigms provide some guidance for the reconstruction in the form of argument schemes. The paradigm of applied formal logic demands reconstructing the argument as deductively valid. We can use the already used argument scheme from optimal choice (Betz and Brun 2016, 72–73). According to this

---

<sup>213</sup>One could, of course, suggest that a more holistic principle of charity must be taken into account: If a reconstruction of  $R_3$  that is attacking  $R_1$  is more charitable than one attacking  $C$ , then this is itself a good reason to prefer interpreting the first sentence as expressing an argument. While I agree with this strategy's correctness, I doubt that it can always lead to the same interpretations among different analysts. The reason is that the accompanying holistic assessments might become very complex and might involve different trade-offs in terms of charity or between charity and other hermeneutic principles, such as doing justice to the discursive aims of the arguer (see Betz 2010, 180–83). Suppose, for example, that it would be more adequate to interpret the first sentence as not expressing a reason and, simultaneously, more charitable to interpret  $R_3$  as attacking  $R_1$ . In this case, these different considerations are locked in a tie with regard to the interpretive choices concerning the first sentence and  $R_3$ . We could begin by considering how much more plausible these interpretations are in comparison, but that would make things even more complicated.

scheme, the optimal means is the best alternative in terms of side effects (expressed in the third premise). Besides the comparison to its alternatives, it is additionally important to compare the side effects of the means with the importance of the aim itself (expressed in the fourth premise). Finally, for the argument to be deductively valid, a strict general principle (the optimal choice principle) is included as a premise.<sup>214</sup> The instantiated argument scheme yields the following reconstruction:

**<R1-AFL Version>:** Reconstruction with a sufficient-means scheme.

- (1) It prima facie (i.e. without considering negative side-effects that are inevitable when bringing about S) ought to be the case that there is no underrepresentation of minorities in higher education, business and politics.
  - (2) Measure of affirmative action by governments will eliminate the underrepresentation of minorities in higher education, business and politics.
  - (3) There is no alternative to measures of affirmative action to eliminate underrepresentation that is more suitable.
  - (4) The certain, likely and possible side-effects of affirmative action are collectively negligible as compared to the elimination of underrepresentation.
  - (5) *\*Optimal choice principle:\** If (i) it prima facie ought to be the case that S, (ii) S will be the case if agent A does X, (iii) there is no alternative to X for agent A that will bring about S and is more suitable than X and (iv) the certain, likely and possible side-effects of agent A doing X are collectively negligible as compared to the realisation of S, then agent A ought to do X.
- form 1-5 --
- (6) Thus, governments should employ measures of affirmative action for minorities that are under-represented in higher education, business and politics.

Walton, Reed, and Macagno (2008) offer on page 96 a similar argument scheme to reconstruct the sufficient-means version of the practical syllogism in the paradigm of informal logic:<sup>215</sup>

**<Sufficient-means scheme of informal logic>**

- (1) My goal is to bring about A (\*goal premise\*).
- (2) I reasonably consider on the given information that each one of [B1, B2, ..., Bn] is sufficient to bring about A (\*alternatives premise\*).

<sup>214</sup>For brevity of exposition, I omitted instantiating the agent A of the principle in some of the premises.

<sup>215</sup>They provide several other argument schemes that seem to fit similarly well. One could use the *value-based practical reasoning scheme* (324), the *argument from goal* (325) or the *argumentation from means and ends* (325). These alternative schemes will result in different reconstructions, which I won't consider here.



- (3) I have selected one member Bi as an acceptable, or as the most acceptable sufficient condition for A (\*selection premise\*).
- (4) Nothing unchangeable prevents me from bringing about B, as far as I know (\*practicality premise\*).
- (5) Bringing about A is more acceptable to me than not bringing about Bi (\*side-effects premise\*).
- from 1-5 --
- (6) Therefore, it is required that I bring about Bi.

This scheme has to be adapted since it is envisaged as one of practical and not of moral reasoning. However, the already used scheme of applied formal logic indicates how to frame such a scheme into moral vocabulary. For instance, it is not enough that the agent in question has some goal. Instead, the first premise should state that something ought to be aimed at. The following reconstruction is but one possible adaption and instantiation of the scheme:

**<R1-IL Version>:** Reconstruction with a sufficient-means scheme.

- (1) The goal of governments should be the elimination of underrepresentation of minorities in higher education, business and politics (\*goal premise\*).
- (2) Measures of affirmative action are sufficient to eliminate this underrepresentation (\*alternatives premise\*).
- (3) Measures of affirmative action are the morally most acceptable sufficient means to eliminate underrepresentation (\*selection premise\*).
- (4) Nothing unchangeable prevents governments from employing measures of affirmative action (\*practicality premise\*).
- (5) To eliminate underrepresentation is morally more acceptable to governments than not employing measures of affirmative action (\*side-effects premise\*).
- from 1-5 --
- (6) Therefore, it is morally required that governments employ measures of affirmative action for minorities that are under-represented in higher education, business and politics.

There are some minor differences between the applied logician's version of the reconstruction (<R1-AFL Version>) and the informal logician's (<R1-IL Version>). The AFL Version does not have the practicality premise of <R1-IL Version> as a separate statement. However, on a charitable reading, it is implicitly expressed in the second premise of <R1-AFL Version>. Additionally, and not surprisingly, the AFL version does not express the optimal choice principle (premise 5 of <R1-AFL Version>) as a premise. For the informal logician, the conclusion is licenced by a non-strict inference rule, which corresponds roughly to the optimal choice principle.

Reconstructing the argument by instantiating the argument scheme is not enough in Carneades. Carneades models critical questions as additional premises and introduces

three different premise types, which categorise premises according to their role for proof burdens. Walton, Reed, and Macagno (2008) provide the following critical questions for the sufficient-means scheme:

- *Other-Means Question*: Are there alternative possible actions besides Bi to bring the goal?
- *Best-Means Question*: Is the means Bi the best (most favourable) of the alternatives?
- *Other-Goals Question*: Do I have goals whose achievement is preferable and that should have priority?
- *Possibility Question*: Is it possible to bring about the goal, in the given circumstances?
- *Side-Effects Question*: Would bringing about the goal have known bad consequences that ought to be taken into account?

Fortunately, all of these critical questions are already represented in the argument scheme in the following way: Asking a critical question is tantamount to objecting to a premise in the argument scheme.<sup>216</sup> The *other-means* and *best-means question* correspond to an attack against the third premise. Questioning whether the means of the conclusion is the morally most acceptable means to reach the intended goal (third premise) is asking whether there is an alternative sufficient means (other-means question), which is more favourable from a moral point of view (best-means question). Similarly, the *possibility question* corresponds to an attack against the second premise: The practical impossibility of bringing about the goal is equivalent to saying there is no sufficient means to achieve the goal. The *other-goals question* can be interpreted as attacking the first premise: If there are other goals whose achievement takes precedence, then the intended goal should not be aimed at. Admittedly, such reasoning presumes that there is some kind of trade-off between reaching both aims, which should be made explicit in the attack. Finally, the *side-effects question* concerns comparing the side effects of achieving the aim with the moral importance of the aim. These side effects depend on the means one takes to reach the aim. Since the best-means question already captures a balancing of side effects among the alternative means, we can interpret the side-effects question as comparing the side effects of the morally most favourable alternative among all the sufficient means with the moral importance of the aim. On this reading, the side-effects question can be interpreted as attacking the fifth premise. Therefore, there is no need to translate the critical questions into additional premises.

The next task of modelling the argument as an argument graph in Carneades is to categorise the different premises according to their burden of proof as ordinary premise, assumption or excluded exception. T. F. Gordon, Prakken, and Walton (2007) suggest categorising premises of the argument scheme as ordinary premises and premises that correspond to critical questions as either assumptions or excluded exceptions. The decision between the latter two is often domain-dependent (T. F. Gordon, Prakken, and Walton 2007, 879). Since premises and critical questions overlap in the case of  $R_1$ , we have to search for other criteria. The different types are defined by how they are related to proof burdens. Walton and Godden (2005) discuss the necessary-means variant of the practical syllogism and seem to suggest that the premises correspond to Carneades excluded exceptions: As long as a respondent or opponent of  $R_1$  does not object to premises of  $R_1$  by advancing evidence

---

<sup>216</sup>Walton and Godden (2005) agree on this for the necessary-means variant of the practical syllogism, which has the same associated critical questions as the sufficient-means variant (Walton, Reed, and Macagno 2008, 96).

or arguments against the premises, the proponent of  $R_1$  has no burden of proof to justify these premises further. At this point, we can just follow this suggestion and categorise all premises as excluded exceptions.

While I am very sympathetic to the idea of explicating the role of critical questions in terms of proof standards, there are some problems of categorising the premises as Carneades' ordinary premises, assumptions and excluded exceptions. I concur with T. F. Gordon, Prakken, and Walton (2007), that proof standards for premises will often depend on the specific domain and its context.<sup>217</sup> For instance, it is plausible that proof standards depend on what is at stake in accepting a claim.

The context-dependence of proof standards might pose the following problems for this specific example and in general. First, we had to translate the original scheme of practical reasoning into one of moral reasoning. So even if there are good reasons to categorise all premises of the original scheme as Carneades' excluded exceptions, the translation might demand categorising the premises differently. A moral context might require other proof standards than a mere prudential one. Second, there is a general context shift: Argument schemes and the Carneades model are conceptualised for dialogical settings—that is, contexts in which an arguer puts forward an argument and a respondent challenges the arguer by advancing counterarguments and critical questions. However, the text snippet we analyse is not a transcribed dialogue but a monologue by an arguer who considers pro and con arguments on their own. The question is whether proof standards differ between real and internalised dialogues. Finally, these worries generalise in the following way: Even if the analyst is supplied with a suggested categorisation of premises for all existing argument schemes, these argument schemes will often have to be adapted to the specific reconstruction context. Additionally, the analyst will encounter arguments that do not fit any of the existing schemes since they do not encompass all possible arguments. What is needed is a systematic theory of proof standards that can categorise premises of ordinary-language arguments and that can deal with the heterogeneity of ordinary-language arguments in their diverse contexts.<sup>218</sup> I do not want to say that these problems cannot be solved. However, these considerations show that the categorisation of premises according to the Carneades model will most probably differ from analyst to analyst without a very complex and ambitious account of proof standards.

Let us now reconstruct the objection  $R_3$  in two versions—one that attacks the conclusion of  $R_1$  and one that attacks  $R_1$  directly. Again, we have to reconstruct the argument based on a concise formulation:

... [5] Others argue that affirmative action can even damage the development of a country; [7] This helps neither majorities nor minorities. ( $R_3$ )

---

<sup>217</sup>See also Walton and Godden (2005), 478.

<sup>218</sup>Walton and Gordon (2011) argue for a specific categorisation of critical questions and premises of the argument scheme from expert opinion; Walton and Godden (2005) do the same for the practical syllogism. However, they do not ground their considerations on some systematic theory of proof standards. While this work is surely helpful, it cannot be generalised in a simple way to all argument schemes, let alone all arguments. There are also systematising theories of proof standards for the legal domain (for instance, T. F. Gordon and Walton 2009). However, while legal argumentation might be subject to very conventionalised proof burdens that allow a more or less precise systematisation, I doubt that such theories can be easily devised for all other argumentation contexts.

What might be the underlying idea of the objection interpreted as an attack against the conclusion *C*? Due to the scarcity of explicitness, there are different interpretations of the objection. The fifth sentence alone might suggest that affirmative action should be avoided since it can damage the development of a country. While that might be an argument on its own, it ignores the role of the seventh sentence. The objection seems to proceed along two steps: Affirmative action can damage the development of a country. That, in turn, prevents aiding minorities. But is that an objection against affirmative action? The underlying idea might be that measures that prevent helping minorities should be avoided.

There is another important qualification in the explicit formulation. The fifth sentence does not claim that affirmative action will damage the development of a country, but formulates something weaker: It *can* damage the development. How should we interpret that in a charitable way? Perhaps a brief look at the target of the objection might help. To attack the conclusion *C* of *R*<sub>1</sub>, the conclusion of *R*<sub>3</sub> should contradict *C*: “*Governments should employ measures of affirmative action for minorities that are under-represented in higher education, business and politics.*” Now, *C* is a general statement about what governments should do. Saying that no government is required to implement measures of affirmative action is clearly in contradiction to *C* and would, therefore, do as the conclusion of *R*<sub>3</sub>. But it is also a very strong claim. Charity to *R*<sub>3</sub> might suggest using a weaker conclusion—perhaps that not all countries should employ measures of affirmative action.<sup>219</sup> In light of this, it seems plausible to interpret the *can* in the fifth sentence as follows: Under certain conditions, affirmative action *will* damage the development of a country, and in some countries do these conditions apply.<sup>220</sup>

These considerations lead to the following reconstruction:

- <R3 as an argument against C>:** R3 as a rebutting defeater to R1.
- (1) There are countries in which the employment of affirmative-action measures would damage the economical development of this country.
  - (2) Damaging the economical development of a country prevents aiding under-represented minorities of this country.
  - from 1,2 --
  - (3) Hence, there are countries in which the employment of affirmative-action measures would prevent aiding under-represented minorities of this country.
  - (4) Governments should help under-represented minorities of their country.
  - (5) Governments should not employ measures that prevent them from what they should accomplish.
  - from 3-5 --
  - (6) Hence, not all governments should employ measures of affirmative action.

<sup>219</sup>This interpretation assumes *C* to be a strict generalisation—a sentence about what *all* countries should do.

<sup>220</sup>This interpretation is also in line with the sixth sentence, which I interpreted as an argument supporting *R*<sub>3</sub> and which uses South Africa as an example of how affirmative action damages a country.

The reconstruction might be in this form already satisfactory for both the informal logician and the applied logician. The main underlying inference rule from the first two premises to the preliminary conclusion can be called *transitivity of cause and effect*: If *a* causes *b* and *b* causes *c*, then *a* causes *c*.<sup>221</sup> This is surely true, even interpreted as a strict generalisation. The applied logician might prefer to add it as an additional premise. The inference from sentences 3–5 to the conclusion is also already deductively valid (minus some minor reformulations). According to sentences 3 and 4, there are countries in which affirmative action would prevent governments from helping minorities (3) even though they ought to help them (4). In other words, not employing measures of affirmative action is necessary for those countries to reach their goals. Accordingly, the general principle formulated in the fifth sentence is a reformulation of the practical syllogism in its necessary-means version: If an aim *A* ought to be reached and *B* is necessary to reach *A*, then *B* ought to be done. This principle can be regarded as a strict generalisation; hence, the inference is even deductively valid.<sup>222</sup> Thus, we arrived at a charitable reconstruction of  $R_3$  as a rebuttal of  $R_1$ .

The question is now whether we can provide a comparably charitable interpretation of  $R_3$  as an undermining defeater of  $R_1$ . To that end, the conclusion of  $R_3$  has to contradict one of the premises of  $R_1$ . So, which premise could be the target of the argument? Fortunately, the answer is easy to find. The objection that affirmative action can damage the economic development of a country does not propose any alternative better means to solve the problem of minority under-representation (premise 3 of <R1-AFL Version> and <R1-IL Version>); it does not question the aim itself (premise 1 of <R1-AFL Version> and <R1-IL Version>) and it does not doubt that affirmative measure will help to reach that goal (premise 2 of <R1-AFL Version> and <R1-IL Version>). But it points to side effects of such measures—namely, the damage to the economic development. As both reconstructions of  $R_1$  exemplify, it is not enough to merely allude to side effects. What is needed is a stronger claim, namely that these side effects are so severe as to override the moral importance of the goal. On this interpretation,  $R_3$  contradicts the fourth premise of the applied logician’s version of the reconstruction (<R1-AFL Version>) and the *side-effects* premise of the informal logician’s version (<R1-IL Version>).

This interpretation motivates the following simple interpretation of the objection as an undermining argument to  $R_1$ :

**<R3-AFL Version>:**  $R_3$  as an undermining defeater to  $R_1$ .

- (1) There are countries in which the employment of affirmative-action measures would have the side effect of damaging the economical development of this country.
- (2) Damaging the economical development of a country is not negligible compared to the elimination of underrepresentation

<sup>221</sup>There is neither an argument scheme in Walton (1996b) nor in Walton, Reed, and Macagno (2008) for this simple pattern. Probably due to its triviality.

<sup>222</sup>Betz and Brun (2016) provide a deductively valid argument scheme for this pattern. Alternatively, the informal logician can use Walton’s necessary-means version of the practical syllogism (Walton, Reed, and Macagno 2008, 95–96), which, however, would demand some further adjustments of the reconstruction.

```

 of minorities.
-- from 1,2 --
(3) Hence, the side-effects of employing affirmative action are
 not negligible in all countries as compared to the elimination
 of underrepresentation.

```

In this reconstruction, the conclusion follows deductively from both premises, and the conclusion contradicts the fourth premise of the applied logician's version of  $R_1$ . A corresponding version for the informal logician demands only a slight fine-tuning of the used formulations:

```

<R3-IL Version>: R3 as an undermining defeater to R1.

(1) There are countries in which the employment of affirmative-action
 measures would have the side effect of damaging the economical
 development of this country.
(2) Avoiding affirmative action's side effect of economical damage is
 more acceptable than the elimination of underrepresentation.
-- from 1,2 --
(3) Hence, there are countries in which avoiding affirmative action's
 side effect of economical damage is more acceptable
 than the elimination of underrepresentation.
-- from 3 --
(4) Hence, it is not the case that the elimination of
 underrepresentation is morally more acceptable to all governments
 than not employing measures of affirmative action.

```

Having reconstructed argument  $R_3$  both as a rebutting and an undermining defeater to  $R_1$ , we can now compare these versions to dissolve this sink ambiguity of the outgoing justificatory relation of  $R_3$ . To that end, we have to compare the premises of the different reconstructions and evaluate how they perform with respect to charity and accuracy. For simplicity, let's assume that the different versions compare similarly well with regard to accuracy and concentrate on a charity assessment—that is, an evaluation of how strong the different argument versions are.

Both versions share the first premise—the statement that there are countries that would suffer some economic damage when employing measures of affirmation action. We can therefore concentrate on the other premises. The current reconstruction of  $R_3$  as a rebutting defeater is, in one aspect, vague: It does not clarify in what aspect minorities suffer from the economic damage and how it relates to the upside of them being less under-represented (the second premise of the reconstruction). That makes it difficult to assess its plausibility. The most crucial difference between the different versions concerns their evaluative statements. The interpretation of  $R_3$  as a rebutting defeater states that countries ought to help minorities without any qualification (premise four), which seems to be a strong claim. It might be more plausible that governments have only a pro tanto obligation to help minorities—that is, an obligation to help them, but one that other moral concerns can outweigh. The interpretation of  $R_3$  as an undermining defeater to  $R_1$  states that the economic damage

caused by affirmative action outweighs the value of eliminating the under-representation of minorities (premise two of both the AFL and the IL version). This statement is, to some extent, vague. Whether it holds depends surely on the extent of the economic damage. Without further qualification, it is, therefore, hard to assess how plausible this premise is.

In light of this, I am hesitant to provide an obvious or straightforward comparison of both versions of  $R_3$ . At this point, there are no conclusive reasons to prefer one reconstruction over the other. What makes things more complicated is the provisional character of these reconstructions and argument reconstructions in general (Betz 2010, 183). I arrived at reconstructions that are adequate to some extent. But there are possibly other more adequate reconstructions. For any reconstruction, there could always be another reconstruction that performs better with respect to hermeneutical principles. Consequently, any disambiguation of a hermeneutical underdetermination by argument reconstructions is provisional. Accordingly, different analyst might come to different interpretations if they base their comparison on a different set of reconstructions, which is a plausible assumption due to the provisional character of any specific set of reconstructions.

The example illustrates another important point: I already argued from a theoretical perspective that the interpretation of argumentation structures by reconstructing arguments depends on the background knowledge the analyst employs in the reconstruction process. The example shows that in the case of moral arguments, the analyst might be forced to compare the plausibility of normative statements to compare different argument reconstructions regarding charity. Fixating the background knowledge among different analysts to ensure intercoder reliability demands in these cases, therefore, not only to fix a body of factual knowledge but also to determine a shared value system.

#### 4.4.8 SUMMARY

Let me close the discussion by briefly summarising the preliminary results. A reconstructive analysis aims to interpret argumentative unit(s) as arguments by reconstructing them in their premise-conclusion form. The basic idea to narrow down the degree of interpretation is to reconstruct different suggested interpretations of arguments and to use hermeneutical principles such as charity and accuracy to choose between them.

In applied formal logic, the reconstruction of arguments is driven by the requirement to reconstruct them as deductively valid, which often requires identifying implicit premises. I discussed how Hitchcock's -Hitchcock (1985) strategy to deal with enthymemes can be used to individuate arguments. The simple examples I analysed show that the tools provided by applied formal logic will not always succeed in pinning down one unique individuation of arguments in a given argumentation. The analyst may face cases of what I called *horizontal underdetermination*, which demands deciding whether different points in favour of the same claim are to be interpreted as one or several arguments for that claim. Additionally, the analyst may face what I called *vertical underdetermination*, which occurs when an argumentation moves along subsequent inferential steps that depend on each other. There seems to be no general recipe to decide how to individuate such argumentations uniquely if the arguer does not provide explicit linguistic cues.

By comparing different suggestions of how to reconstruct arguments in terms of how these



suggestions perform in terms of charity, the applied logician might be able to choose between different interpretations of the argument structure. A reconstructive analysis can thereby narrow down the degree of interpretation with respect to node ambiguity and argument granularisation. However, choosing between different interpretations by using the principle of charity demands judging the plausibility of premises. These judgments depend on the particular background knowledge of the analyst and can, therefore, vary from analyst to analyst. If analysts work independently from each other and do not share their background knowledge, we should expect differences between analysts in how they categorise text segments as justificatory relevant and in how they individuate arguments. Consequently, reconstructive analysis based on applied formal logic will—depending on the text at hand—allow for different interpretations.

In contrast to applied formal logic, the informal logician can think about the granularisation of argumentation in two different ways. They can either use the concept of sufficiency to individuate arguments or the concept of relevance to individuate reasons.

The individuation of arguments is based on argument schemes and inference rules. Argument schemes describe recurring forms of argumentation. By instantiating the placeholders of such a scheme, the analyst arrives at an argument reconstruction, which represents one individuated argument. I argued that argument individuation based on argument schemes might diverge between different analysts for the following reasons: First, the analyst will encounter arguments that do not fit any of the given schemes since the prominent accounts are incomplete. In this case, they have to resort to other reconstruction strategies. Second, an argumentation can be reconstructed with different schemes, resulting in different argument individuations. Finally, argument schemes are not based on a systematic theory of argument reconstruction. As a consequence, there might be good reasons to reconstruct a specific argument in a way that differs from a reconstruction guided by argument schemes.

The alternative of using non-strict inference rules to reconstruct arguments runs into the same difficulties as the reconstruction of arguments in the paradigm of applied formal logic.

The informal logician uses the prominent convergent-linked distinction to distinguish different types of arguments and can be used as a criterion to individuate reasons. I discussed two different criteria, which I borrowed from Freeman (2011). An *intuition-based criterion* that grounds the individuation of reasons in our judgements of independent probative relevance and a *syntactical criterion* that grounds the individuation of reasons in the inference rules of the arguer. I arrived at the following conclusions: The intuition-based criterion is helpful and easy to apply—in the sense of neither depending on sophisticated theoretical baggage nor the reconstruction of arguments. However, it does not exclude cases of horizontal and vertical hermeneutical underdetermination in reason individuation. On the other hand, the syntactical criterion demands that inference rules be attributed to the arguer. Since there is no canonical way to determine these inference rules uniquely, the syntactical criterion cannot exclude cases of horizontal and vertical hermeneutical underdetermination in reason individuation.

To what extent are reconstructive methods able to narrow down relation ambiguity? I introduced two different accounts of macrostructure to answer this question: the theory of dialectical structures (Betz 2010) as a representative of the applied-logic paradigm and



Carneades model of argument (T. F. Gordon, Prakken, and Walton 2007) as a representative of the informal-logic paradigm.

The theory of dialectical structures introduces justificatory relations between arguments based on deductively valid argument reconstructions. It is parsimonious—in that it defines only two relations between arguments that are reduced to logical relations between statements—but still expressive enough to analyse the macrostructure of complex debates. The theory can even help to disambiguate apparent cases of relation ambiguity. However, the analyst has to invoke the principle of charity to choose between different interpretations of the macrostructure. What counts as charitable will depend on the specific background knowledge of the analyst. Hence, interpretations between analysts might vary to the extent that their background knowledge diverges. Therefore, this theory cannot guarantee the elimination of the variability between analysts in the identification of argumentation macrostructure.

The Carneades model of argument is primarily designed for the context of legal reasoning but can be considered a paradigmatic account of argumentation macrostructure from the informal-logic viewpoint since it is based on Walton's argument schemes and their associated critical questions. I argued that while Carneades can circumvent some of the problems to distinguish between “normal” premises and critical questions on the one hand and between undercutting and undermining defeaters, on the other hand, its reliance on proof standards to distinguish between different types of justificatory relations is a drawback when it comes to relation ambiguity. Without a systematic and complex theory of proof standards that can account for the heterogeneity and extensiveness of ordinary-language argumentation, different analysts might come to different interpretations in the analysis of macrostructure.

After having discussed the prospects of both accounts on a theoretical level, I illustrated their application on a concrete example to disambiguate a sink ambiguity; to be more precise, I considered an argument *A* that could be interpreted as a rebutting defeater but also as an undermining defeater to another argument *B*. On the former interpretation, the conclusion of *A* contradicts the conclusion of *B*; on the latter interpretation, the conclusion of *A* contradicts a premise of *B*. The basic strategy of disambiguation was to reconstruct *B* and two different versions of *A*—one as a rebutting and one as an undermining defeater to *B*—and then to compare the different versions of *A* according to hermeneutical principles such as the principle of charity. The version that performs better with regard to this comparison determines the more adequate interpretation of the justificatory relation to *B*.

It turned out that the differences between both paradigms do not matter much for the disambiguation of this relation ambiguity. If anything, the applied logician might be faced with additional degrees of interpretational freedom since they introduce subcategories of justificatory relations that are irrelevant to the applied logician (the distinction between undercutting and undermining defeaters on the one hand and the distinction between premises of an argument scheme and its critical question on the other hand).

In sum, and not surprisingly, dissolving relation ambiguity faces the same problems as a unique individuation of arguments and reasons: The comparison of different reconstructions according to the hermeneutical principles might involve trade-offs and depends on the background knowledge of the analyst. Consequently, different analysts might come to

different conclusions about the interpretation of justificatory relations between arguments.

## 4.5 PRELIMINARY CONSEQUENCES FOR A MINIMAL ANNOTATION SCHEME OF ARGUMENTATION STRUCTURE

I want to close this chapter by using the results we arrived at so far to motivate guidelines for a minimal annotation scheme of argumentation structure, which you will find in Appendix A.1 of this work. The annotation scheme is designed to narrow down the degree of variability between different annotators. It will be minimal in the sense that it is parsimonious in the number of introduced categories. In addition, it is motivated by considering the problems discussed in this chapter to be paradigmatic. I will, therefore, refrain from using elaborate concepts from argumentation theory if they do not seem to contribute to narrowing down the interpretational distance between different annotators.

These annotation guidelines are, of course, just one possible category system to analyse argumentation structure. It can be used in different ways for future empirical work to assess the argumentation structure in argumentative texts or to assess the phenomenon of hermeneutical underdetermination in the analysis of argumentation structure. It might, for instance, serve as a starting point to design more ambitious category systems by introducing further subcategories, or it might serve as a reference point for assessing whether more elaborate concepts borrowed from argumentation theory can narrow down the degree of interpretation—in contrast to what the considerations of this chapter suggest.

Let me start by providing some clarifying remarks to assess the scope and limitations of this chapter's findings. I started by introducing three broad types of hermeneutical underdetermination (node ambiguity, underdetermination of granularisation and relation ambiguity) and argued that a reconstructive analysis cannot disambiguate all problematic cases of these indeterminacies. I provided some general worries and concrete examples to show that different annotators can come to different conclusions about the structure of a given argumentation. The main problem is that independent of the preferred reconstructive paradigm—informal logic or applied formal logic—the annotator will have to invoke their background knowledge to reconstruct arguments, which may differ from analyst to analyst.

However, the extent of hermeneutical underdetermination in a specific corpus is a contingent property of this corpus. Accordingly, only an empirical analysis can gauge the degree of interpretation. In particular, the examples I used to illustrate and justify the methodological considerations are anecdotal and cannot be simply generalised to any corpus whatsoever. While I maintain that the examples I used are, to some extent, paradigmatic and not only constructed to support my points, it is possible that these cases present a negligible phenomenon in ordinary-language argumentation. In sum, it is an empirical question to what extent different annotators vary in their assessment of argumentation structure, and I did not provide any general empirical findings on that matter. Instead, this work points to possible problems and aims to provide a framework to investigate the phenomenon of hermeneutical underdetermination empirically.

The minimal category system that I suggest has three main features: It is designed to analyse argumentation structure with a focus on macrostructure; it is grounded on relational

categories only and does not depend on any reconstructive analysis of arguments. Let me elaborate on these features in more detail.

The category system is confined to the analysis of argumentation structure.<sup>223</sup> It will, for instance, not address rhetorical power, style, persuasiveness or literary merit. Informal and applied formal logic are normative in that they aim to assess arguments and reasons with respect to their rational strength. The annotation of argumentation structure will be non-evaluative and can, therefore, refrain from invoking concepts solely designed to evaluate argumentation. Additionally, the category system will abstract away from dialogical aspects of argumentation. In particular, it is not concerned with who maintains which stance, who puts forward which argument or who tries to persuade whom. While these aspects can be relevant to the analysis of argumentation structure, they are not categorised. The aim is to distinguish mere structural properties of argumentation from other aspects strictly. It is, of course, possible to introduce additional categories that account for these features later.

One reason for this level of abstraction is the danger of neglecting argumentative texts that do not fit into dialectical categories. The annotation guidelines have a broad scope of application. The category system should be able to deal with any argumentative text. These can include transcripts of ordinary-language debates but may also encompass texts in which an author presents a wide range of pro and con reasons on a particular topic without attributing the different considerations to someone in particular. It is often the case that an author presents their point of view by providing their reasons for it; it is also often the case that an author intends to persuade their audience. But presenting arguments and reasons is only contingently connected to arguing for a point of view and the aim of convincing someone. An author might present arguments while at the same time upholding a neutral position. As a consequence, it would be a mistake to conceptually tie the category system to dialectical concepts.<sup>224</sup>

The broad scope of application is also linked with the focus on analysing the macrostructure. Longer texts will often include not only one argument and one main claim but many. The author might consider objections to their central claims and refute them with further arguments. Such argumentative considerations can increase in their depth and represent high-level hierarchies of reasons and counter-reasons that stand in different justificatory relations to each other. This suggests representing such structures at least by trees with an arbitrary depth. However, the confinement to tree structures will not be sufficient to represent the argumentative complexity of any text. Reasons and their justificatory relations to each other and to statements can form circles.<sup>225</sup> Consequently, the model of macrostructure should allow graph structures and not only tree structures. The nodes of these graphs will represent text segments, and the edges between these nodes will represent

---

<sup>223</sup>See also Section 2.5.1.

<sup>224</sup>The annotation guidelines Habernal, Eckle-Kohler, and Gurevych (2014), which are used in Habernal and Gurevych (2016), are a good illustration of this point. The authors conceptually tie the concept of argumentation to the aim of persuasion (Habernal and Gurevych 2016, 129). Accordingly, they asked annotators to categorise texts that are formulated to convince others (Habernal, Eckle-Kohler, and Gurevych 2014, 4). However, later, they correctly observe that an author can provide arguments while upholding a neutral stance and redefine the category ‘persuasive text’—thereby becoming a misnomer—to include such texts as well (Habernal, Eckle-Kohler, and Gurevych 2014, 13).

<sup>225</sup>See, for instance, Stab (2017, 47). Examples of such complex macrostructures can be found in Betz and Cacean (2012) and Cacean (2012).

justificatory relations.

The second feature of the proposed category system is its confinement to relational categories between text segments. All categories will qualify a text segment as standing in relation to another text segment. You can think of it as the constraint that there will be no unconnected nodes in the graph representing the argumentation structure. The motivating reason for this design decision is that it is not a significant restriction for the analyst of argumentation structure and avoids the introduction of additional categories that might hamper the reliability of the annotation. It is not a severe restriction since the most important argumentative categories (*reason*, *objection*, *premise*, *conclusion*, *assumption*, ...) are relational. They categorise a text segment as having a justificatory role for another text segment. For instance, a text segment presented as a reason is a justification for another text segment; a text segment presented as an objection is a justification that some other text segment is wrong or implausible. The same applies to the other main argumentative categories.

However, this is a simplification that needs qualification: Often, one of the relata is not another text segment but an implicit statement. Implicit claims are the paradigm case. An author might present reasons for some claim without explicitly expressing this claim.<sup>226</sup> That does, however, not change the relational character of the categories. It is still a relation between a text segment and something that could have been a text segment. The important point here is that the relata are statements, or rather they can be formulated as statements in contrast to other types of relations.

What I want to exclude with the confinement to such relational categories are non-relational categories or relational categories between text segments and other types of entities. For instance, a statement can be regarded as the main claim of the author if it summarises their standpoint and the author's aim is to convince their audience of its truth. This can be regarded as a relational category between a statement and a person. I will not introduce this and similar categories because they are either outside the scope of analysing argumentation structure or they can be characterised by using the described relational categories. Let's take the example of the category 'main claim' again. It can happen that an author explicitly expresses the main claim as I characterised it ("*My claim is that ... and I want to convince you of its truth.*") without providing any reason for that claim. This case is simply irrelevant in terms of argumentation structure since the argumentation structure is trivial. Usually, however, the author will provide reasons for their main claim. But then, we can employ an alternative characterisation and define the main claim of a text as a text segment that is justified by other text segments and not used to justify other text segments. In consequence, there is no need to define this concept as a separate category that needs annotation. Whether a text segment is, in this sense, the main claim is determined by the given relational categories. At the same time, we are avoiding the introduction of additional categories that can hamper the reliability of the annotation process.

Finally, the annotation scheme will not be based on a reconstructive analysis. The annotator is not asked to transform text segments into the premise-conclusion form of arguments by using either the described methods of informal or applied formal logic. Instead, the

---

<sup>226</sup>In their annotation study, Habernal and Gurevych (2016) observe that in their corpus, in 48% of the cases, the author's claims are implicit.

identification of argumentative discourse units will be, in large part, grounded in a non-reconstructive analysis. The decision to do without a reconstructive analysis is motivated by pragmatic considerations. The reconstruction of arguments is time-consuming and demands extensive training. The question is whether the accompanying costs are worth the effort. As we saw in this chapter, the advantages of a reconstructive analysis in terms of narrowing down the interpretive distances between different analysts compared to a mere non-reconstructive analysis are unclear. I argued that analysts will differ in their assessment if their relevant background knowledge differs. We can, therefore, expect that different analysts will reconstruct arguments differently if they work independently from each other. Hence, reconstructive analysis does not necessarily have an advantage in terms of reliability.

Instead of reconstructing text segments, annotators are simply asked to identify text segments that are presented as having a justificatory role—that is, they will identify the relata of justificatory relations, which can be text segments or implicit statements. The identification of argumentative units and their justificatory relations will be based on what I called non-reconstructive analysis (see Section 4.2.1). In other words, the annotation will be grounded in analysing linguistic cues and the contextual understanding of texts. The annotation scheme will be confined to just two justificatory relations: a support relation and an attack relation. The idea is that many important natural-language concepts relevant for analysing argumentation structure can be captured with this simplistic model.  $y$  being presented as a reason or an argument for  $x$  will be modelled with a support relation between  $y$  and  $x$ ;  $y$  being presented as an objection to or refutation of  $x$  will be modelled by an attack relation between  $y$  and  $x$ ;  $z$  being a rebutting defeater of the justification of  $x$  by  $y$  will be modelled by an attack relation between  $z$  and  $x$ ;  $z$  being an undermining defeater of the justification of  $x$  by  $y$  will be modelled by an attack relation between  $z$  and  $y$ .

The annotation scheme does not introduce further subcategories of justificatory relations besides the attack and support relation. In particular, I refrain from defining categories that further qualify the intended probative force of support and attack relations. I also refrain from defining categories that distinguish between undercutting and undermining defeaters.

The intended probative force is usually difficult to determine in ordinary-language argumentation. Either linguistic cues uniquely express probative force, or a lack of such explicitness underdetermines it.<sup>227</sup> In addition, the decision to introduce different justificatory relations in terms of their probative force presupposes a particular argumentation-theoretic stance. While the applied logician interprets all or at least most support relations as (implicitly) truth-preserving—that is, as relations that are represented by deductively valid arguments—the applied logician interprets most support relations as defeasible.

Classifying attack relations into undercutting and undermining defeater is problematic for similar reasons. It presupposes the reconstruction of arguments and depends on the specific account of enthymemes the analyst endorses. For instance, while the applied logician

---

<sup>227</sup>This is in line with the explorative empirical findings in Habernal and Gurevych (2016). They used an adapted Toulmin model for their annotation of argument structure and decided to do without qualifiers of Toulmin's original model—which can be interpreted as qualifying the probative force of support relations (Freeman 1991, chap. 5)—since the absence of such qualifiers in their data. In a random sample of 40 documents, “authors [did not] state the degree of cogency (the probability of their claim, as proposed by Toulmin)” (142).

regards undercutting defeaters as irrelevant to most arguments (see Section 4.4.5), the informal logician often categorises attacking arguments as undercutting defeaters. However, even for the informal logician, the distinction between undermining and undercutting defeaters will sometimes be blurry (see Section 4.4.6). Since I am not aware of any suggestion to circumvent these problems without advocating some ambitious argumentation-theoretic stance,<sup>228</sup> I will omit the distinction between undercutting and undermining defeaters in the minimal annotation scheme.

One important aspect concerns the granularisation of argumentation or, in other words, the individuation of argumentative components, which can be claims, reasons or arguments. The annotator has to identify text segments corresponding to argumentative components; in particular, they must decide where they start and end. We saw that the individuation of arguments is slightly different than the individuation of reasons. We should, therefore, decide whether analysts should identify reasons or arguments or both.

Both the individuation of arguments and the individuation of reasons are beset with problems. Individuating arguments demands reconstructing arguments, and even then, the individuation of arguments will be underdetermined in some cases, as I argued in Sections 4.4.2 and 4.4.3. The individuation of reasons, on the other hand, can be accomplished by employing an intuition-based or the syntactical criterion, which I borrowed from Freeman (2011). The intuition-based criterion grounds the individuation of reasons in our judgements of independent probative relevance, and the syntactical criterion grounds the individuation of reasons in determining the inference rules of the arguer. I argued that the intuition-based criterion is helpful and easy to apply—in the sense of neither depending on sophisticated theoretical baggage nor the reconstruction of arguments. However, it cannot exclude the possibility of granularisation ambiguity. But neither can the syntactical criterion, which, additionally, requires reconstructing arguments. In particular, it demands that inference rules be attributed to the arguer, and there is no canonical way to determine these inference rules uniquely.

In sum, neither option—the individuation of arguments and the individuation of reasons—can guarantee that different analysts converge in their interpretation of the argumentation’s granularisation. Again, it might turn out that one option or the other is preferable in narrowing down interpretive distances between annotators. However, this is an empirical

---

<sup>228</sup>Peldszus, Warzecha, and Stede (2016) (used in Peldszus and Stede (2015)) ask annotators to distinguish between what they call rebutters and undercutters. Their rebutters encompass what I introduced as undermining and rebutting defeaters of arguments. Their definition of undercutters as attacking the relation between two propositions (Peldszus, Warzecha, and Stede 2016, 8) corresponds to what I introduced here as undercutting defeaters. Their rule of thumb to distinguish rebutters (of the argument’s conclusion) and undercutters is to ask whether the “attack still works effectively” (9) without the premise. If it doesn’t, the attack depends on the premise, and it can be regarded as an undercutter if it is not a rebutter of the premise in question (Peldszus, Warzecha, and Stede 2016, 8–9). While this rule of thumb might be very helpful in distinguishing between rebutting defeaters and defeaters that are either undermining or undercutting, it cannot distinguish between the latter two. If we hypothetically omit the premise in question and the attack ceases to work, the concluding explanation must be that the attack is targeting something that was implicitly meant with that premise. That can be another implicit premise or the inferential link between the premises and the conclusion. Their rule of thumb is, therefore, only able to discriminate between undermining defeaters of the premise the annotator is asked to omit hypothetically, rebutting defeaters of the argument and other possibly implicit parts of the argument—be they implicit premises or the inferential link between premises and conclusion.

question beyond the scope of this work. At this point, there is no convincing reason to use a reconstructive analysis for the granularisation of argumentation. Instead, the annotation scheme asks to individuate reasons and employs the discussed intuition-based criterion as a rule of thumb. The obvious pragmatic advantage of this suggestion is that annotators can do without reconstructing arguments.

Up to this point, I sidestepped one practical difficulty of annotating the argumentation structure in argumentative texts. Often, an author will reformulate a claim or reason that they already made. What is more, the corresponding text segments might be non-contiguous to each other. In these cases, the annotator will encounter different formulations of the same argumentative component at different locations in the text. There are several reasons for an author to reformulate or repeat a point. They might simply want to explain it in other words to foster a better understanding on the part of the audience, or they might want to add some detail by reformulating a point in a more precise way, or they might simply use a restatement as a rhetorical means to emphasize the importance of a point. Consequently, a reason for some statement or a statement that is being justified by a reason might correspond to different non-contiguous text segments. The decision of whether two text segments belong to the same argumentative component is relevant for the granularisation of argumentation, or in other words, for the number of annotated reasons and claims. There are two resulting issues for the design of the annotation scheme. First, it should provide criteria to determine whether two non-contiguous text segments correspond to one or different argumentative components. Second, we should decide whether these reformulations should be annotated themselves.

Whether two different argumentative units express the same argumentative component is often a non-trivial question. An author will rarely repeat the same point by using the exact same words. Thus, searching for different occurrences of identical sentences or clauses won't be enough. Instead, an author will at least slightly reformulate the points they repeat. Additionally, the phenomenon cannot be fully captured by deciding whether two argumentative units are logically equivalent in their literal meaning. For instance, the point of elaborating a claim might be to provide further details in the form of qualifications. The author might begin by stating their main claim unprecise and later formulate the qualifying antecedent conditions of that claim. For example, they might begin by claiming that taxation is a reprehensible state-sanctioned form of robbing citizens and later add that this does not hold for taxes only used to finance a country's infrastructure. In this case, the different text segments will not be logically equivalent. What is explicitly expressed will differ between both text segments. That is neither unusual nor does it necessarily violate any conversational rule of our natural language. Instead, the second text segment is intended to clarify what the author intended to express with the first text segment despite both being not equivalent in their literal meaning. The decision of whether two non-contiguous argumentative units belong to the same argumentative component can, therefore, be context-dependent, might demand the analysis of the author's intentions and should be guided by the principle of charity. There will be no simple criteria to guide the identification of intended reformulations and restatements but at the best rules of thumb.

The other question is whether the annotator should annotate all possibly non-contiguous argumentative units that belong to the same argumentative component. Alternatively, we could let the analyst annotate just one of the text segments. I will opt for the first option for

a simple reason. It will allow us to compare the possibly diverging decisions of different annotators with respect to the granularisation of argumentation. In particular, it will allow us to understand cases in which a difference in the chosen granularisation can be explained by differences in deciding whether text segments formulate new argumentative components or belong to argumentative components that are scattered over the text. On a technical level, annotators will be provided with labels for argumentative components. In the case that several argumentative units represent the same argumentative component, annotators are asked to use the same label for the different units.<sup>229</sup>

The resulting annotation scheme is minimal in two ways. It is not grounded in a reconstructive analysis of arguments, which has the advantage of training annotators without introducing advanced argumentation-theoretical concepts. Additionally, the annotation scheme is based on a minimal set of categories. It defines just two justificatory relations and one restatement relation. Despite this parsimony, the category system can be fruitfully used to analyse the macrostructure of argumentation. It can provide useful insights, especially if the data is complex regarding its argumentation structure.

---

<sup>229</sup>Other annotations studies adopt another strategy to annotate repetitions and reformulations. They introduce an additional relation between text segments that annotators are asked to use if two text segments repeat or reformulate a point. For instance, Peldszus, Warzecha, and Stede (2016) introduce a *restatement relation* and Kirschner, Eckle-Kohler, and Gurevych (2015) use a *detail relation*. This is, however, only a technical difference. Whether we use labels or introduce another relation has no practical relevance.



## 5. CONTENT ANALYSIS OF ARGUMENTATION STRUCTURES (CAAS)

The main aim of this chapter is to show how the proposed annotation scheme for analysing argumentation structure (see 4.5 and A.1) can be applied within the paradigm of reliability-orientated content analysis (see 3.2.3). In particular, I will provide answers to the following questions:

1. How can the content analyst measure the extent of hermeneutical underdetermination of a text's argumentation structure?
2. How can the content analyst satisfy the requirement of reliability even if the analysis of argumentation structure is faced with an irreducible degree of interpretation?

The intuitive answer to the first question is to assess the spread or variance of all correct interpretations. We might, for instance, use the spread within a sample to estimate the population variance. To that end, we had to calculate the mean difference or disagreement between the coding results of properly instructed annotators. Chance-corrected reliability measures draw similarly on the observed disagreement among reproduced annotations to estimate the reliability of the data-making process. We can therefore take advantage of the existing accounts and exploit them to devise a disagreement measure for CAAS (5.1).<sup>230</sup> After discussing several relevant accounts, I suggest using a graph-edit distance to compare coding results. Such a measure presupposes an alignment of coding units among annotators. We have to determine whether an argumentative unit identified by one annotator represents the same argumentative component as an argumentative unit of another annotator.<sup>231</sup> This inter-coder identification of argumentative units is a practical problem, which I will solve by adapting the alignment procedure developed by Mathet, Widlöcher, and Métivier (2015) (5.2).

To answer the second question, we have to apply the general considerations of Chapter 3 to the case of analysing argumentation structure. By relying on the measurement picture of content analysis, I showed that researchers can use the method of significance testing and are thereby able to detect differences in the phenomena of interest even if annotators diverge in their coding results due to interpretational differences. To that end, the data-making process must be conceptualised probabilistically. The proposed recipe for content analysis

---

<sup>230</sup>In a similar vein, the suggestions I make in this chapter can be used to devise a chance-corrected reliability measure by using the disagreement measure and the statistical model (5.3.1) I propose. For reasons outlined in Chapter 3, I do not pursue this line of reasoning here.

<sup>231</sup>I distinguish between argumentative units and argumentative components. Cf. Footnote 9 on page 13.

(see 3.4.3) provided a mere sketch of how to do this. In this chapter, I will elaborate more on this outline in the context of argumentation analysis (5.3).

## 5.1 A DISAGREEMENT MEASURE FOR CAAS

A natural way to assess the degree of interpretation, or the extent of interpretational underdetermination, is the following: First, introduce a measure that captures the notion of disagreement or, in other words, the distance between two interpretations. Then, determine the set of all correct interpretations and calculate the mean disagreement between all pairs of interpretations. The resulting value can be regarded as a measure of the degree of interpretation.<sup>232</sup> If it is not possible or feasible to determine the set of all interpretations, we might take the observed disagreement in a random sample of interpretations as an estimate for the degree of interpretation.

In argumentation analysis, an interpretation corresponds to a set of identified argumentative units and a set of justificatory relations between them (see Figure 5.1 for an example). The question is how to introduce a disagreement measure for such coding results.

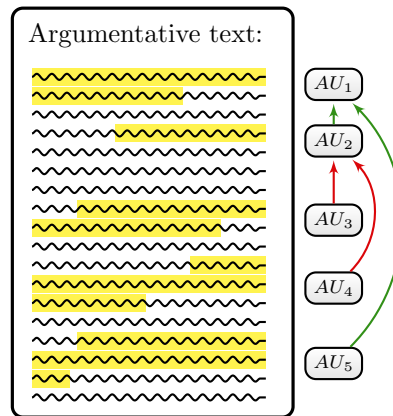


ABBILDUNG 5.1

A schematic example of annotating the argumentation structure of a text. There are five annotated argumentative units connected by support and attack relations.

The calculation of chance-corrected reliability coefficients relies on quantifying the disagreement between coding results. Let us, therefore, review these measures to assess whether they can help to design a disagreement measure for CAAS.

Chance-corrected reliability measures have the following general form:

$$R = 1 - \frac{D_o}{D_e} \quad (5.1)$$

$D_o$  denotes the observed disagreement within the reliability data—that is, the result of replicating the annotation by different and independently working annotators. The expected

<sup>232</sup>I do not claim this is the only reasonable way to introduce such a measure.

disagreement  $D_e$  measures the disagreement that would result if a pure chance mechanism were to perform the coding—which is of no interest to our purpose.

Under certain favourable circumstances, the definition of  $D_o$  is straightforward. Suppose that annotators do not have to determine the start and end points of coding units; instead, every annotator has to categorise the same set of coding units. For instance, the content analyst might predetermine the coding units for all annotators. Further, suppose there is only one category with a set of category values corresponding to a nominal scale. In other words, the task of annotators is to attribute one of the category values to each predefined coding unit. Finally, let's assume the researcher performs the reliability assessment based on two annotations. Under these circumstances, we can define the observed disagreement as the relative frequency of how often both annotators disagree in the annotation of a coding unit.

Unfortunately, the annotation procedure in CAAS differs from this simple situation in several important points.

First of all, it will, in most cases, be necessary to reproduce a coding effort more than once since the sensitivity to detect differences between texts will depend on the sample size. The more annotators are used to analyse the argumentation structure of different texts, the better we can detect differences.<sup>233</sup> Hence, we need a measure that is not confined to situations of two annotators.

Second, annotators have to identify argumentative units by themselves. They have to distinguish justificatory relevant text segments from other text segments. To that end, annotators must determine the boundaries of argumentative units, which may encompass a clause, a sentence, multiple sentences and everything in between. Following Krippendorff (2013), the identification of coding units by annotators prior to their categorisation is called *unitising*.<sup>234</sup> Apart from very specific contexts, unitising in CAAS is hermeneutically underdetermined. In other words, if annotators employ the suggested annotation scheme, we can expect them to diverge in their identification of argumentative units.<sup>235</sup>

Figure 5.2 illustrates some properties of unitising in CAAS that are of importance: Argumentative units of different annotators might coincide ( $AU_5$  and  $AU_7$ ) or only overlap ( $AU_1$  and  $AU_6$ ). The latter case can be interpreted by saying that different annotators might identify the same argumentative component but disagree in identifying the boundaries of the corresponding argumentative unit. It might also happen that an argumentative unit identified by one annotator does not overlap with any other argumentative unit of another annotator. For instance, it seems that the first annotator disagrees with the other two about whether the text segment beginning at position 340 and ending at 380 ( $AU_2$ ) is an argumentative unit. Furthermore, there might be gaps between argumentative units. In other words, texts will often have text segments that are not justificatory relevant. Finally,

<sup>233</sup>This claim cannot be proven rigorously here. For the case of using the normal distribution as the statistical model, see 3.4.2.

<sup>234</sup>That is not to say that unitising and categorisation are necessarily distinct processes—that, for instance, annotators have, first, to identify coding units before they can categorise them. It is rather a conceptual distinction relevant to the definition of reliability coefficients. CAAS is one example where the process of unitising is procedurally identical to the process of categorising: To identify a text segment as argumentatively relevant is to determine that the text segment is justificatory related to some other text segment (see A.1.7).

<sup>235</sup>Several specific examples can be found in the annotation guidelines (A.1).

different annotators might choose a diverging granularisation. For instance, one annotator might identify two different arguments while another annotator sees one complex argument ( $AU_1$  versus  $AU_3$  and  $AU_4$ ).<sup>236</sup>

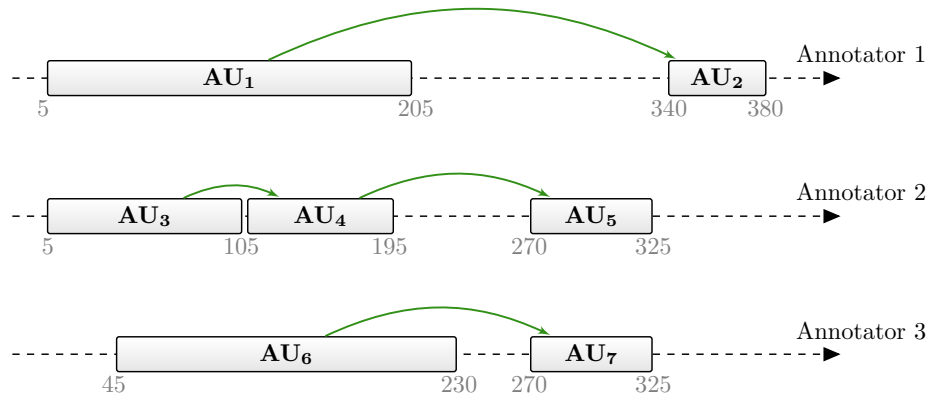


ABBILDUNG 5.2

Unitising in CAAS. The figure illustrates the identification of argumentative units by three different annotators. The x-axis represents positions in the continuum of the textual data, say, an ascending indexing of characters. Argumentative units are referred to by unique names  $AU_1, AU_2, \dots, AU_n$ . That is, two coding units of two different annotators receive different names even if they have the same start and end points (e.g.,  $AU_5$  and  $AU_7$ ).

The last important difference to the above-described simple case concerns the actual categorisation. In the simple example, annotators are asked to categorise coding units by attributing category values to single coding units. In CAAS, text segments are justificatory relevant in virtue of them standing in a justificatory relation to other text segments. There are, therefore, no monadic but only relational category values.

### 5.1.1 RELIABILITY OF UNITISING

If annotators identify coding units themselves, the resulting segmentations might diverge. Accordingly, there are two different types of how coding results can differ: There might be *positional disagreements*—that is, differences in the allocation of coding units—and *categorical disagreements*—that is, differences in the attribution of category values to coding units. But then there are also two corresponding dimensions of reliability: reliability of unitising and reliability of categorisation.

Reliability of unitising was surprisingly late addressed by content analysts. Important chance-corrected reliability measures for categorising, such as Scott's (1955)  $\pi$ , Cohen's (1960)  $\kappa$  and Krippendorff's (1970)  $\alpha$  were introduced decades earlier than the first suggestion of a chance-corrected reliability measure for unitising by Krippendorff (1995).<sup>237</sup>

This might be explained in two ways. First, in some settings, content analysts can preselect coding units. Under these circumstances, every annotator categorises the same coding units. Accordingly, there is no need to assess the reliability of unitising. Second, the introduction of reliability measures for unitising is mathematically intricate, as we will see shortly.

<sup>236</sup>What I referred to as hermeneutical underdetermination of granularisation (4.1.2).

<sup>237</sup>Krippendorff (1995, 54–55) mentions Guetzkow (1950) as an exception, whose measure is, however, not chance-corrected.

There are three general approaches to introducing reliability measures of unitising. The first suggestion is to conceptualise unitising as a token labelling task, which allows using existing reliability measures for categorisation. The second approach is based on quantifying the overlap between different coding units. This overlap-based approach results in measures that quantify both reliability of unitising and categorisation. Finally, alignment-based measures are grounded in an inter-annotator identification of coding units. Such an alignment allows for neatly separating reliability of unitising from reliability of categorisation.

One way to measure disagreements in unitising is to conceptualise unitising as a token labelling problem in the following way (see Figure 5.3): First, the continuum of data must be segmented into chunks that are the same for each annotator. These atoms must be chosen so that the boundaries separating identified coding units from their surrounding irrelevant text will always coincide with the boundaries of these atoms. For instance, the coding instructions for unitising usually guarantee that coding units will never sever words. In such a case, the continuum can be atomised into distinct words.

Under this construction, each identified coding unit corresponds to a sequence of atoms (e.g., in Figure 5.3,  $A_6$  and  $A_7$  correspond to the unit  $U_1$ ). Similarly, each text segment not identified as a coding unit is a sequence of atoms (e.g.,  $A_1 - A_3$ ). In this way, the task of unitising can be conceptualised as categorising predefined coding units—the given atoms—by two category values: being part of a coding unit and being irrelevant text. In other words, the construction transforms unitising into a binary classification task of predefined units.

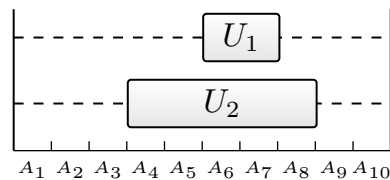


ABBILDUNG 5.3

Example of unitising as categorisation. The continuum consists of ten atoms. The second annotator identified a longer text segment  $U_2$  as a coding unit (spanning over the five atoms  $A_4 - A_8$ ), and the first identified a shorter coding unit  $U_1$  ( $A_6 - A_7$ ), which is contained in  $U_2$ .

Understanding the task of unitising as a binary classification task relieves the content analyst of devising special reliability measures for unitising; they can use existing measures for categorisation. Additionally, the resulting reliability values will not depend on the fineness of the chosen atomisation, which is desirable (Krippendorff 1995, 56).

However, there are also several problems. First, the measure is invariant under transposing what is considered a coding unit and an irrelevant text segment since the names of both introduced labels do not matter for calculating reliability values. This is regarded as an undesirable feature since distinctions within the set of identified coding units are generally of more importance than distinctions within the set of identified irrelevant text segments (Krippendorff 1995, 56).

Additionally, the measure is not responsive to certain relevant differences between unitised texts (Krippendorff 1995, 56–57; Mathet, Widlöcher, and Métivier 2015, 444–54). In

particular, the construction cannot distinguish differences in the number of identified coding units if they are contiguous. Consider, for instance, Figure 5.4. The first annotator has identified two coding units; the second has identified one long coding unit. However, both annotators are in perfect agreement according to the binary classification. There are no atoms that are categorised differently between both annotators. Accordingly, the resulting reliability values will be equal.<sup>238</sup>

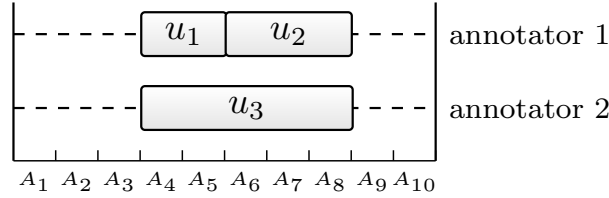


ABBILDUNG 5.4  
Example illustrating a difference in granularisation.

The last point is significant in CAAS since annotators can disagree on the granularity of reasons and arguments (see Chapter 4). If measuring positional disagreements is supposed to account for such differences, the discussed measure is unsuitable.

Let us, therefore, consider another idea of measuring disagreements in unitising. Overlap-based measures take the extent of overlap between text segments as a basis to conceptualise positional disagreement. This basic idea can be explicated in different ways. Here, I will only consider the  ${}_u\alpha$  coefficient from the family of  ${}_u\alpha$ -reliability measures by Krippendorff et al. (2016), who introduced some improvements over earlier suggestions by Krippendorff.<sup>239</sup> Additionally, I will confine the discussion to how  ${}_u\alpha$  measures the observed disagreement  $D_o$  without considering their conceptualisation of the chance correction.

Let  $S = s_1, s_2, \dots, s_n$  be the set of the identified segments by all annotators, including segments that represent gaps—that is, segments that are not marked as coding units but as irrelevant data between them.<sup>240</sup> Figure 5.5 illustrates such a case. The segments are numbered consecutively across all annotators. Let further  $C$  be the set of all category values and  $cat(s_i)$  the annotated category value of segment  $s_i$ . Gaps are denoted by an additional category value  $\phi$  (i.e.,  $cat(s_i) = \phi$  if  $s_i$  is a gap).

Let  $|s_i \cap s_j|$  be a measure of the overlap between two segments, which can, for instance, be quantified by counting the character tokens in the overlap  $s_i \cap s_j$ . The observed disagreement

<sup>238</sup>A simple suggestion to fix this problem is using three instead of two categories as, for instance, proposed by Ajjour et al. (2017). Although they do not intend to measure positional dissimilarities, their suggestion can be adapted to assess the reliability of unitising. The idea is to categorise atoms as either irrelevant text, as the first atom of a coding unit, or as some subsequent atom of a coding unit—known as the IOB labeling problem (Stede and Schneider 2018, 65). By using an extra label for the beginning of a coding unit, differences in granularisation entail a difference in the calculation of reliability values. However, the resulting differences might be tiny. What is more, they will depend on the fineness of the chosen partitioning, in contrast to the original suggestion to use only two category values.

<sup>239</sup>These include Krippendorff (1995), Krippendorff (2004c) and Krippendorff (2013).

<sup>240</sup>I will use a different notation than Krippendorff et al. (2016), one that is more in line with the notation used in the following parts of this chapter.

$D_o$  of unitising is now calculated by adding up  $|s_i \cap s_j|$  over a specific subset  $S_{u\alpha}$  of all segment pairs  $\langle s_i, s_j \rangle$ :<sup>241</sup>

$$D_o = c \sum_{\langle s_i, s_j \rangle \in S_{u\alpha}} |s_i \cap s_j| \quad (5.2)$$

The definition of  $S_{u\alpha}$  should depend on what we want to measure by  $D_o$ . Since we want to assess disagreements between different annotators, we should only consider segment pairs of different annotators. Let  $ann(s_i)$  the annotator who identified  $s_i$ . The first requirement can now be expressed as follows:  $S_{u\alpha}$  should only contain segment pairs  $\langle s_i, s_j \rangle$  with  $ann(s_i) \neq ann(s_j)$ .

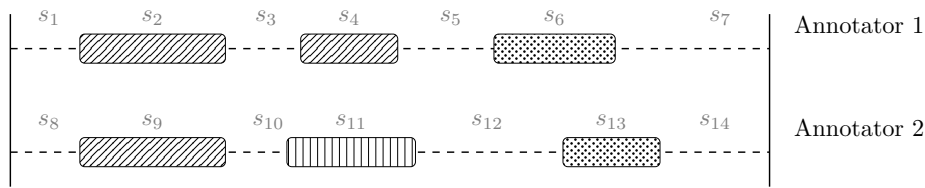


ABBILDUNG 5.5

A simple example of unitising a continuum of data by two annotators. Shaded boxes represent identified coding units. Different types of shading indicate category values.

The  $u\alpha$  coefficient is conceptualised to measure both categorical and positional dissimilarities. Accordingly, we should not include segment pairs that agree in their boundaries and their attributed category values (e.g.,  $s_2$  and  $s_9$  in Figure 5.5). However, if two segments are annotated with different categories and overlap to a large extent (e.g.,  $s_4$  and  $s_{11}$ ), their overlap should be considered in the calculation of  $D_o$ . Including such unit pairs into  $S_{u\alpha}$  will take care of measuring categorical disagreements.

Perhaps surprisingly, a slight generalisation of this suggestion already results in a proper quantification of positional dissimilarities. More specifically, if we define  $S_{u\alpha}$  by

$$S_{u\alpha} := \left\{ \langle s_i, s_j \rangle \mid ann(s_i) \neq ann(s_j) \text{ and } cat(s_i) \neq cat(s_j) \right\} \quad (5.3)$$

the observed disagreement defined by equation 5.2 will not only measure differences in categorisation but also in unitising.

Consider, for instance,  $s_6$  and  $s_{13}$  in Figure 5.5. They have the same category value but overlap only partially. One plausible interpretation is that both annotators identified the same phenomenon but slightly disagreed on the boundaries of the corresponding coding unit. This disagreement in unitising might be measured by counting those tokens of  $s_6$  and  $s_{13}$  on which they differ (i.e.,  $|(s_6 \cup s_{13}) \setminus (s_6 \cap s_{13})|$ ). But this is already gauged by defining  $S_{u\alpha}$  as suggested: The tokens of  $s_6$  that are not contained in  $s_{13}$  correspond to the overlap of  $s_6$  and the gap  $s_{12}$ , which are in  $S_{u\alpha}$  since they are categorised differently. Similarly, the tokens of  $s_{13}$  that are not in  $s_6$  correspond to the overlap  $s_{13} \cap s_7$ . Hence, the suggested

<sup>241</sup>The constant  $c$  serves for a correct normalisation of  $D_o$ . For the specifics, see Krippendorff et al. (2016).

construction of  $S_{u\alpha}$  will count observed differences in categorisations and unitising in a simple and neat way.

The described approach has, however, some problems. In some contexts—CAAS is one of them—a measure of observed disagreement should be sensitive to differences in the granularisation of coding units. The  $u\alpha$  coefficient is unfortunately unable to account for such differences.

Consider the four examples in Figure 5.10. Case A depicts a diverging granularisation. The first annotator identifies two different but contiguous segments; the second identifies only one longer segment at the same location. Suppose that both annotators attribute the same category value to these segments. The segments  $s_2$  and  $s_3$  have only an overlap with  $s_6$ . Since  $s_6$  is given the same category value as  $s_2$  and  $s_3$ , the pairs  $\langle s_2, s_6 \rangle$  and  $\langle s_3, s_6 \rangle$  are not in  $S_{u\alpha}$ . Additionally, there are no other segment pairs  $s_i, s_j$  with  $cat(s_i) \neq cat(s_j)$ . Hence, there is no observed disagreement according to  $u\alpha$  even though the chosen granularisation differs.

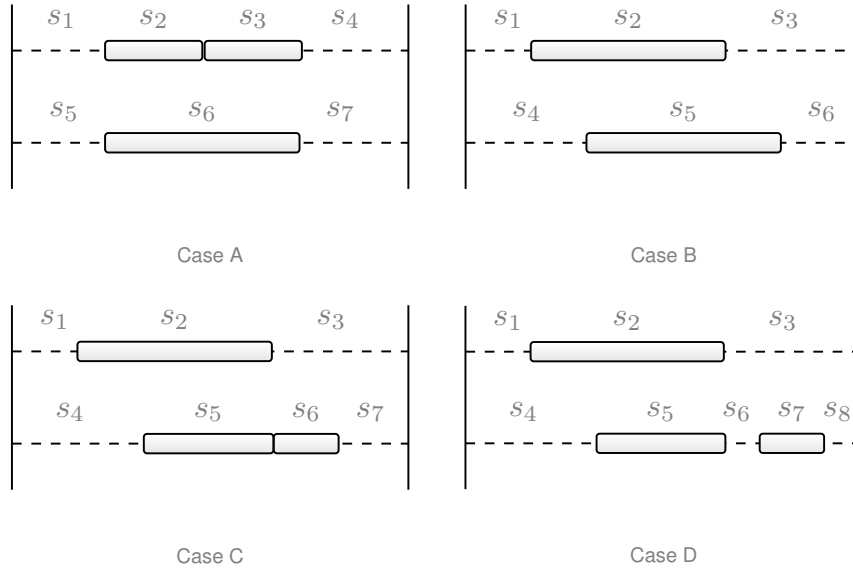


ABBILDUNG 5.10  
Examples to illustrate problems of  $u\alpha$ .

The other cases illustrate that  $u\alpha$  has more general difficulties in detecting certain differences in the individuation of coding units. In Case B,  $s_2$  overlaps with gap segment  $s_4$ , and  $s_5$  overlaps with  $s_3$ . Hence,  $u\alpha$  will measure the corresponding difference in unitising. The problem is that  $C$  and  $D$  will not be considered different to  $B$  by  $u\alpha$  since the contributing non-vanishing overlaps are the same in all cases. The overlap  $s_2 \cap s_4$  is the same in all cases. The following overlaps are equal as well:  $s_5 \cap s_3$  in B,  $s_6 \cap s_3$  in C and  $s_7 \cap s_3$  in D. However, there is clearly a difference in the individuation of coding units between all cases.

Additionally, Case B illustrates that this problem is not confined to contiguous text segments.<sup>242</sup>

<sup>242</sup>The authors are well aware of this behaviour and add that the connected calculation of expected



It is instructive to understand the basic idea of the discussed overlap-based measure from a different perspective. The contribution of one specific coding unit to  $D_o$ , for instance  $s_2$  in Figure 5.11, can be divided into the following parts: Its non-vanishing overlap to other coding units that are categorised differently ( $s_5$ ) and its non-vanishing overlap to gaps ( $s_6$ ,  $s_7$ ). Although  $s_8$  is not in  $S_{u\alpha}$ , it plays an indirect role for the contribution of  $s_2$  to  $D_o$ . It occupies space ( $s_2 \cap s_8$ ) that would contribute to  $D_o$  if a gap occupied it. On this view, we might reinterpret the calculation of  $D_o$  in the following way: We, first, align coding units that have a non-vanishing overlap ( $s_2$ ,  $s_5$  and  $s_8$ ). Then, we can regard the overlap between units in this alignment that are categorised differently ( $\langle s_2, s_5 \rangle$  and  $\langle s_5, s_8 \rangle$ ) as a measure of categorical dissimilarity. Similarly, the extent to which they do not overlap with each other (e.g.,  $\langle s_1, s_5 \rangle$ ,  $\langle s_2, s_6 \rangle$  and  $\langle s_5, s_7 \rangle$ ) can be regarded as a measure of positional dissimilarity. In other words, the calculation of  $D_o$  by  $u\alpha$  can be interpreted as relying on an implicit alignment of coding units and a subsequent calculation of  $D_o$  by the overlap within this alignment and by the overlap between elements of this alignment and the surrounding gaps.

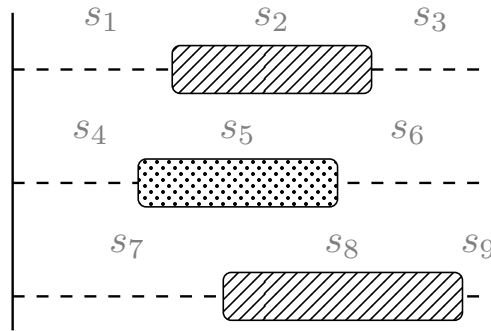


ABBILDUNG 5.11

Example of differences in unitising to illustrate that the overlap-based approach relies on an implicit alignment of coding units. In the example, the segments  $s_2$ ,  $s_5$  and  $s_8$  can be considered aligned.

The basic idea of alignment-based accounts of reliability is to ground the calculation of observed (and expected) disagreement on an *explicit* alignment of coding units. Instead of merely being an interpretation, the alignment of coding units results from explicit criteria and is then used to calculate disagreements. This idea is motivated by the evident problem of calculating categorical differences that arises when annotators identify coding units. If coding units are predefined, every annotator has to categorise the same text segments. Under these conditions, it is clear what is meant by comparing whether two or more annotators categorise a specific coding unit differently. However, if annotators identify coding units by themselves, there are not necessarily coding units that are the same for each annotator. Hence, what needs to be added is a criterion that identifies coding units of different annotators as representing the same phenomenon if they do not share the exact boundaries but only their bulk. This is what an alignment is supposed to achieve:

agreement is responsive to such differences, which renders  $u\alpha$  in its aggregated form responsive to such differences as well (Krippendorff et al. 2016, 2351–52). Here, we are interested in a suitable definition of observed differences without incorporating a chance correction. Consequently, the discussed behaviour is problematic for this purpose.

To determine those coding units of different annotators that represent the same (semantic) phenomenon.

If coding units are aligned in such a way, the content analyst can use the usual reliability coefficients to assess the reliability of categorisation. Moreover, they can use additional measures to assess the disagreement in unitising based on the alignment. Admittedly, much will depend on the specifics of the alignment, as we will see in the following sections.

### 5.1.2 RELATIONAL CATEGORIES

Assessing the reliability of categorisation is based on measuring the disagreement among different annotators. The reference points for these comparisons are category values of coding units. The observed disagreement can then be measured by the frequency of coding units that are categorised differently by different annotators. The categorisation of relations between coding units is different. Instead of assigning category values to individual coding units, annotators decide whether different coding units are related to each other in a specific way.

How do we measure disagreements in the annotation of relations—especially, in the context of CAAS?<sup>243</sup> The simple answer is to assess whether annotators attribute different relations to coding-unit pairs. In other words, instead of regarding individual coding units as reference points, we take coding-unit pairs as reference points for the comparison. On this view, the relations are category values annotators attribute to coding-unit pairs.

The only difference is that there are a lot more coding-unit pairs than coding units.<sup>244</sup> Additionally, there are usually more category values than relations. Suppose annotators are instructed to choose precisely one out of  $n_r$  relations or not to connect two coding units by a relation. In this case, there are  $n_r + 1$  possibilities to annotate a coding-unit pair since annotators might decide that two coding units are not related in any way.

This suggestion has the advantage that existing reliability measures for categorisation can be used without any adaption. However, there are two possible complications.<sup>245</sup> First, most of the common reliability measures assume that category values exclude each other. In other words, annotators must choose exactly one category value for an annotation. However, depending on the annotation scheme, coders might be allowed to assign more

<sup>243</sup>It is beyond this work to provide a general overview of reliability measures for the annotation of relations. This phenomenon plays an important role in different annotations tasks. One example is the annotation of anaphoric references. For an overview, see, e.g., Artstein and Poesio (2008), esp. Section 4.4.

<sup>244</sup>If there are  $n$  coding units, then there are  $\binom{n}{2} = \frac{n(n-1)}{2}$  (unordered) coding-unit pairs.

<sup>245</sup>Kirschner, Eckle-Kohler, and Gurevych (2015) claim that there is another problem. They analyse annotations of argumentative relations and observe that annotators categorise only a small amount of all units to be related to each other (1%-2%). The majority of unit pairs are categorised as non-related. The authors worry that this “may yield misleading high or low [reliability] values” (Kirschner, Eckle-Kohler, and Gurevych 2015, 3). They suggest omitting unit pairs in the calculation of reliability values. Expected agreement indeed becomes high if most data falls under one category, which might lead to low reliability values even if percentage agreement is high. This behaviour is, however, desired. A high percentage of agreement might, in this case, stem from indiscriminate annotation. If one category value is prevalent, we expect a reliability coefficient to measure annotators’ ability to agree on rare category values. But this is what chance-corrected coefficients like  $\kappa$  and  $\pi$  achieve. For details on this apparent prevalence problem, see Artstein and Poesio (2008, 573–74).

than one relation to a pair of coding units. While such configurations are unlikely in CAAS, they cannot be excluded from the outset (see A.1.7). There are reliability measures that are tailored to such cases.<sup>246</sup> However, here, we are only interested in an adequate conceptualisation of observed disagreement and can therefore widen the perspective to search for an apt measure. From a mathematical perspective, coding-unit pairs that are annotated by different relations have the structure of a graph. Coding units correspond to nodes in the graph, and the relations between coding units correspond to edges between these nodes. Accordingly, we might utilise measures that quantify distances between graphs as disagreement measures.

The second challenge prevails if annotators identify coding units themselves. If we want to compare coding results using graph-distance measures, we need to know which nodes and node pairs of different graphs must be compared. In other words, we need an inter-annotator identification of coding units. Consider the coding results in Figure 5.12a of two annotators that are represented by two graphs (5.12b and 5.12b). Both graphs agree on the relation of the two nodes  $N_1$  and  $N_2$ . However, in contrast to the first graph, the second contains an additional node ( $N_3$ ) that has a relation to the node  $N_2$ . This qualitative description of agreements and disagreements between two annotators in terms of graph differences is based on an alignment of coding units. In the example, the coding unit  $U_1$  is aligned with  $U_3$  and represented by the node  $N_1$ . Similarly,  $U_2$  is aligned with  $U_5$  and represented by the node  $N_2$ .

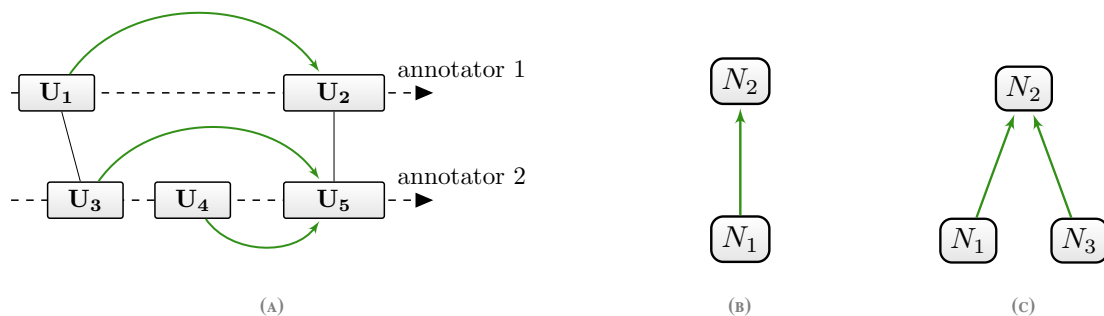


ABBILDUNG 5.12

Comparing categorisations of relations by using graph-based measures. The coding results of both annotators correspond to two graphs whose nodes and edges can be compared based on an inter-annotator alignment of coding units. Vertical lines indicate an inter-annotator identification of units.

Therefore, assessing the disagreement in categorising relations requires an alignment of coding units. The results of measuring disagreements will depend heavily on the specifics of the alignment procedure and, if it is based on comparing the resulting graphs, on the specifics of measuring differences between them. Let me illustrate these points with two specific examples.

Duthie et al. (2016) propose an integrated measure of inter-annotator agreement for the annotation of argumentation structure. This measure contains, among other dimensions, a suggestion to compare annotated relations between coding units. What they call a propositional content relation score is based on an alignment of coding units and a comparison of the resulting argument maps. Coding units of different annotators are matched to each other

<sup>246</sup>See, e.g., Krippendorff (2004c).

using the Levenshtein distance, which is an edit distance. It counts the minimal number of atomic alterations in terms of substitutions, deletions and insertions of characters needed to change one sequence of characters into another. For instance, the sequence ‘cat’ can be transformed into the sequence ‘catch’ by inserting two characters at the end of the string. Hence, the Levenshtein distance is two.

Three conditions determine the alignment of coding units: First, coding units of different annotators are only aligned if they have a non-vanishing overlap. Second, an annotator’s coding unit can maximally be aligned with one other coding unit of every other annotator. Third, coding units are aligned in such a way that the overall Levenshtein distance of all aligned coding units is minimised.<sup>247</sup>

Figure 5.13 illustrates some consequences of such an alignment procedure. Both units  $U_3$  and  $U_4$  overlap with  $U_1$ . Since the overlap of  $U_3$  with  $U_1$  is larger than the overlap of  $U_4$  with  $U_1$ ,  $U_1$  will be aligned with  $U_3$  but not with  $U_4$ . Additionally,  $U_2$  will be aligned with  $U_5$  since the Levenshtein distance is zero between them. Comparing the resulting graphs (5.13b and 5.13c) shows that they differ from each other by one node ( $N_3 \rightarrow N_2$ ).<sup>248</sup>

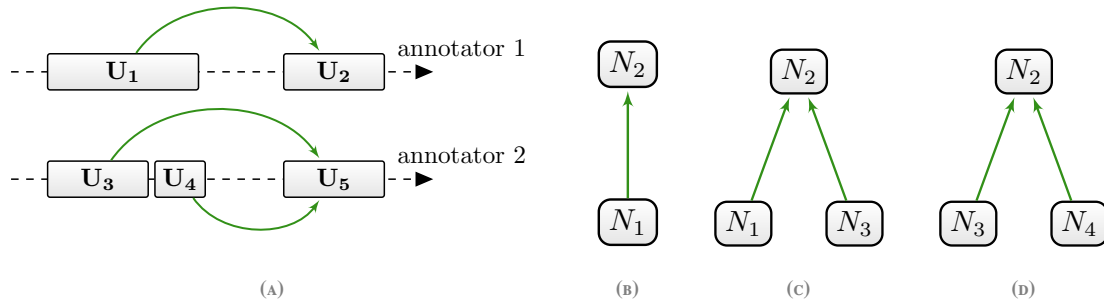


ABBILDUNG 5.13

Comparing consequences of different alignment procedures for the resulting argument maps. The argument maps 5.13b and 5.13c correspond to the resulting alignment by minimising the Levenshtein distance, which leads to an alignment of  $U_3$  with  $U_1$ . The maps 5.13b and 5.13d result if neither  $U_3$  nor  $U_4$  is aligned with  $U_1$ .

However, this alignment might seem counterintuitive from an argumentation theoretic perspective. Both annotators disagree on the granularisation of supporting reasons. This is a case of what I referred to as horizontal underdetermination of individuation (see 4.4.2). The first annotator identified one reason or argument ( $U_1$ ) in favour of another text segment ( $U_2$ ). In contrast,  $U_3$  and  $U_4$  can be considered as two parts of  $U_1$  that the second annotator identified as independent reasons or arguments. On this view, however, neither unit should be aligned with  $U_1$ . Neither represents the same reason or argument as  $U_1$  but parts of  $U_1$ . If we follow this interpretation, the resulting argument maps (5.13b for the first annotator, 5.13d for the second) differ from the ones suggested by Duthie et al. (2016). The difference between the arguments maps (5.13b and 5.13d) amounts to two nodes ( $N_3$  and  $N_4$ ) and two

<sup>247</sup>The given explications of the alignment procedure are sparse in Duthie et al. (2016). What I describe here is the most plausible interpretation from my perspective, which might still diverge from what the authors intended to convey. Fortunately, it matters little whether I misinterpret the authors since I aim to illustrate some points.

<sup>248</sup>In the following, ‘ $\rightarrow$ ’ and ‘ $\leftarrow$ ’ will refer to support relations and ‘ $\rightsquigarrow$ ’ and ‘ $\leftrightsquigarrow$ ’ to attack relations.

relations ( $N_3 \rightarrow N_2$  and  $N_4 \rightarrow N_2$ ), instead of one node and one relation (between 5.13b and 5.13c).

How can we move from a mere qualitative description of differences between annotation graphs to a quantitative assessment? There are several possibilities, which will suit different purposes. For instance, Kirschner, Eckle-Kohler, and Gurevych (2015) use a measure that quantifies the degree to which one graph is included in another one. Let  $E_A$  be the edges of Graph A and  $SP_B(N_i, N_j)$  the number of edges of the shortest path from node  $N_i$  to node  $N_j$  in Graph B (without considering the direction of edges and with  $SP_B(N_i, N_j) := \infty$  if there is no path from  $N_i$  to  $N_j$ ). The extent  $\mathcal{I}(A, B)$  to which A is contained in B can be expressed by:

$$\mathcal{I}(A, B) := \frac{1}{|E_A|} \sum_{N_i, N_j \in E_A} \frac{1}{SP_B(N_i, N_j)} \quad (5.4)$$

For instance, Graph A in Figure 5.14 contains two edges, which can also be found in Graph B. Accordingly,  $\mathcal{I}(A, B)$  yields  $\frac{1}{2}(\frac{1}{1} + \frac{1}{1}) = 1$ . A similar calculation for  $\mathcal{I}(B, A)$  yields  $\frac{1}{3}(\frac{1}{1} + \frac{1}{1} + \frac{1}{\infty}) = \frac{2}{3}$ . Since two graphs are maximally similar to each other if they are contained in each other, the arithmetic mean  $\frac{1}{2}(\mathcal{I}(A, B) + \mathcal{I}(B, A))$  is a natural suggestion to measure the similarity between two graphs. In the same vein,

$$\mathcal{D}(A, B) := 1 - \frac{1}{2}(\mathcal{I}(A, B) + \mathcal{I}(B, A)) \quad (5.5)$$

can be used to measure differences between both graphs.

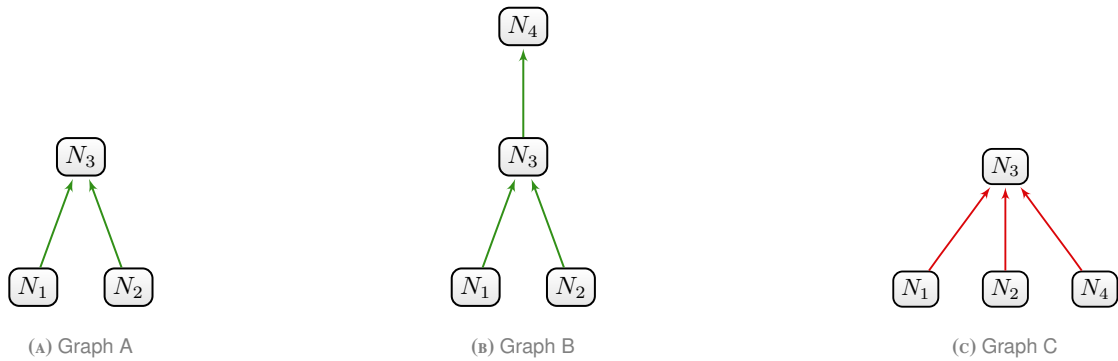


ABBILDUNG 5.14

Simple annotation graphs to illustrate using  $\mathcal{D}$  (Definition 5.5) to quantify the dissimilarity between two graphs. While  $\mathcal{D}(A, B)$  yields  $1/6$ , the differences between B and C are not captured ( $\mathcal{D}(B, C) = 0$ ).

However, as Kirschner, Eckle-Kohler, and Gurevych (2015) note, their measure is neither responsive to the edges' direction nor the type of the relations. Accordingly, the suggested graph-similarity measure does not detect the differences between Graph B and Graph C. This is far from a terminal objection against using this measure. However, it shows that the analyst must be aware of what is measured with a specific quantity and decide whether the used measure serves the intended purpose.

### 5.1.3 MEASURING DISAGREEMENT BETWEEN ARGUMENTATION STRUCTURES

Let me briefly recapitulate the preliminary findings from the last two sections. There are different approaches of measuring differences in unitising. The overlap-based family of  $_{\alpha}$  coefficients and the reformulation of unitising as a token-labelling problem cannot account for differences in the granularisation of argumentation. An alignment-based approach, on the other hand, is responsive to such differences if the underlying alignment procedure refrains from aligning coding units that correspond to different granularisations.

What is more, by using an alignment-based approach, we can separate two concepts: A measure of disagreements in unitising  $\Delta_u$  and a measure of disagreements in categorisation  $\Delta_r$ . I consider this an important feature in the context of CAAS since both types are conceptually very distinct, and the content analyst will be interested in assessing them separately. Let us assume that the argumentative units of different annotators are only aligned with each other if they represent the same argumentative component (e.g., the same reason or the same argument). Let us further assume that  $\Delta_u$  aggregates positional disagreements of aligned units. Under these assumptions,  $\Delta_u$  measures the fuzziness of the boundaries of argumentative components. But fuzzy boundaries should not influence differences in the general structure of argumentation, which is supposed to be measured by  $\Delta_r$ . If we want to assess the degree of hermeneutical underdetermination that encompasses all interpretational indeterminacies I discussed in Chapter 4, we need a disagreement measure that abstracts away from mere positional disagreements between identified argumentative units.

How do we conceptualise a proper measure  $\Delta_r$  that can quantify the degree of hermeneutical underdetermination? The preceding considerations suggest to use an approach that is based on an alignment of argumentative units. Accordingly, we have to accomplish two tasks: First, we have to think about a proper procedure of aligning argumentative units. As we saw in the last section, such an alignment allows to properly compare the resulting argumentation graphs. The second task is to define a measure  $\Delta_r$  that quantifies categorical disagreements between these graphs.

The alignment of argumentative units should be based on a semantic criterion. The basic idea is simple. The units of different annotators should be aligned with each other if and only if they represent the same argumentative components. These include all elements that are either the source or the target of the introduced justificatory relations—such as reasons, arguments, objections or central claims. For instance, if argumentative units of two different annotators represent the same reason, they should be aligned with each other.

The suggested criterion has immediate consequences for the treatment of differences in granularisation. Argumentative units that result from a diverging interpretation of argument or reason individuation should not be aligned. Let's reconsider the above-discussed example of Figure 5.13. Whereas the first annotator interpreted a text segment as one compound argumentative component ( $U_1$ ), the second annotator saw two independent reasons or arguments ( $U_3$  and  $U_4$ ). But neither single part is to be identified with the whole compound component. Consequently, neither  $U_3$  nor  $U_4$  should be identified with  $U_1$ .

Unfortunately, the suggested semantic criterion faces practical difficulties. If we want to determine whether two overlapping text segments represent the same reason or argument,

we have to analyse these text segments under the consideration of their content. In other words, it demands a content analysis on its own. In most contexts, such a superimposed content analysis won't be feasible. In the following, I will rely on the next best thing: An approximation of the suggested alignment criterion. Similarly to the already discussed alignment approach of Duthie et al. (2016), I will suggest grounding the alignment on syntactical criteria. The inter-annotator identification of argumentative units will be decided by considering their positions and widths only. The hope is that this approach will lead to sufficiently similar results as applying the semantic criterion.<sup>249</sup> To that end, I will discuss and adapt the alignment procedure developed in Mathet, Widlöcher, and Métivier (2015) in the following sections (5.2.1 and 5.2.2).

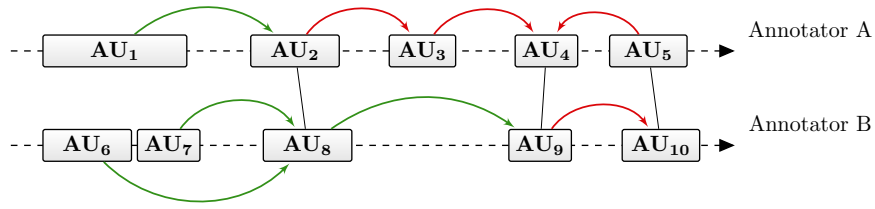
One intended purpose of using  $\Delta_\tau$  is to assess the degree of hermeneutical underdetermination in the analysis of argumentation structure. In other words, if we let  $n_s$  coders annotate the argumentation structure of a text and if we can assume, first, that they follow the coding instructions correctly and, second, that their annotated argumentation graphs  $\tau = \{\tau_1, \tau_2, \dots, \tau_{n_s}\}$  sufficiently encompass the space of all interpretations,  $\Delta_\tau(\tau)$  should quantify the degree of interpretational leeway. Accordingly,  $\Delta_\tau$  is supposed to quantify all differences between argumentation graphs that might result from the discussed interpretational indeterminacies (see 4.1). Annotators might diverge in the identification of argumentative units, in the determination of the type and direction of their justificatory relations to other components and the individuation of arguments and reasons.

Fortunately, these requirements can be met using a simple graph-edit distance. The example in Figure 5.15 illustrates all mentioned indeterminacies: Annotator *B* chose a different granularisation of *A*'s argumentative unit  $AU_1$  ( $AU_6$  and  $AU_7$ ). Since the alignment procedure won't align these units with each other, this difference in interpretation corresponds to a difference between the resulting argumentation graphs of *A* ( $\tau_A$ ) and *B* ( $\tau_B$ ). In contrast to  $\tau_A$ ,  $\tau_B$  has additional edges ( $AC_6 \rightarrow AC_2$  and  $AC_7 \rightarrow AC_2$ ). Furthermore, *A* identified a text segment as an argumentative unit ( $AU_3$ ), which *B* did not. More specifically, *A* interpreted their unit  $AU_2$  as a rebuttal of an objection ( $AU_3$ ), which attacks another text segment ( $AU_4$ ). *B*, on the other hand, regarded their corresponding argumentative unit  $AU_8$  as directly supporting  $AU_9$ . Again, this difference corresponds to a difference in node-edge compounds: Whereas  $\tau_A$  contains a chain of attacking edges from  $AC_2$  to  $AC_4$  ( $AC_2 \rightsquigarrow AC_3$  and  $AC_3 \rightsquigarrow AC_4$ ),  $\tau_B$  contains one supporting edge from  $AC_2$  to  $AC_4$ . Finally, the annotators disagree about the direction of the justificatory relation between their last two argumentative units. Again, the difference is mirrored in the argumentation graphs. Whereas  $\tau_A$  contains an attack relation from  $AC_4$  to  $AC_5$ ,  $\tau_B$  contains the same relation but in the opposite direction.

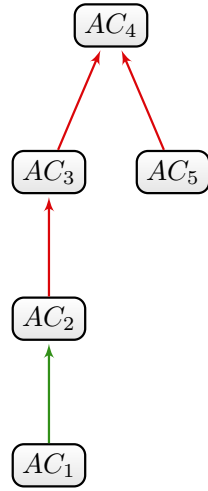
In sum, all interpretational differences can be found in the argumentation graphs resulting from the argumentative units' alignment. Moreover, I described the differences solely in terms of comparing the edges (together with their sources and targets) of the different graphs.

This motivates defining the disagreement  $\delta_\tau$  between two graphs  $\tau_1$  and  $\tau_2$  in terms of differences in node-edge compounds. The point is that there is no need to consider nodes

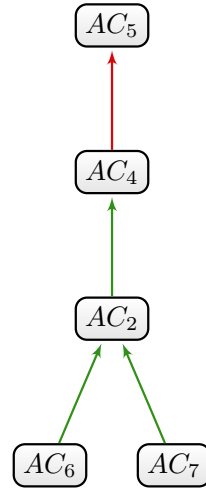
<sup>249</sup>Both the methodological explication of applying the semantic criterion and the empirical endeavour to compare the performance of both approaches are beyond this work.



(A) Annotations of A and B. Vertical lines indicate alignments.



(B) Map  $\tau_A$ .



(C) Map  $\tau_B$ .

ABBILDUNG 5.15

Annotation example that illustrates differences in interpretation.



and edges separately since the coding scheme does not allow to annotate a text segment as an argumentative unit without any justificatory relation to some other unit (see A.1.7). Hence, we can define  $\delta_\tau$  by counting edges that are in  $\tau_1$  and not in  $\tau_2$  and the other way around. This can be considered a graph-edit distance since it counts the number of alterations—in terms of adding and deleting edges—to transform  $\tau_1$  into  $\tau_2$ .

More formally, let  $E(\tau)$  the edges of an argumentation graph  $\tau$ , then we define the pairwise distance  $\delta_\tau$  between two arguments maps by

$$\delta_\tau(\tau_1, \tau_2) := |E(\tau_1) \setminus E(\tau_2) \cup E(\tau_2) \setminus E(\tau_1)| \quad (5.6)$$

In the example of Figure 5.15, both argumentation graphs do not share any of their edges ( $E(\tau_A) \setminus E(\tau_B) \cup E(\tau_B) \setminus E(\tau_A) = \{AC_1 \rightarrow AC_2, AC_2 \rightsquigarrow AC_3, AC_3 \rightsquigarrow AC_4, AC_5 \rightsquigarrow AC_4, AC_6 \rightarrow AC_2, AC_7 \rightarrow AC_2, AC_2 \rightarrow AU_4, AU_4 \rightsquigarrow AU_5\}$ ) and, hence,  $\delta_\tau(\tau_A, \tau_B) = 8$ . In the case of more than two annotation graphs, we can now define  $\Delta_\tau$  as the mean pairwise distance between the argumentation graphs:

$$\Delta_\tau(\tau) := \frac{2}{|\tau|(|\tau| - 1)} \sum_{\tau_i, \tau_j \in \tau} \delta_\tau(\tau_i, \tau_j) \quad (5.7)$$

This measure is not only able to quantify the degree of interpretation, but it is also easy to interpret since it is based on counting those edges that are not shared by two argumentation graphs. If needed, the measure can be adapted (e.g., by normalising it in some way) and complemented by additional measures.

## 5.2 ALIGNMENT-BASED APPROACHES

Mathet, Widlöcher, and Métivier (2015) introduced a chance-corrected reliability measure, the  $\gamma$  coefficient, which has a number of attractive features. It can be used with more than two coders and measures reliability of both unitising and categorisation. Additionally, the approach does not impose any specific requirements on the process of unitising. For instance, it allows gaps and overlaps between coding units. Finally, the measure draws on an alignment of coding units, which can be adapted by different parameters and by substituting the used distance functions.

In the following sections, I will explain the basic ideas of the alignment procedure on which  $\gamma$  is based (5.2.1).<sup>250</sup> After revealing some problems of directly using this alignment procedure in the context of CAAS, I will adapt the measure accordingly (5.2.2).

### 5.2.1 THE GAMMA APPROACH

The basic idea of the alignment procedure on which  $\gamma$  is based is simple and the same as in Duthie et al. (2016): Coding units are aligned in such a way that the overall alignment

<sup>250</sup>I will closely follow the notation and terminology introduced by Mathet, Widlöcher, and Métivier (2015).

minimises the aggregation of certain distances between coding units. However, the elaboration of this idea is more intricate than in Duthie et al. (2016), who minimised the overall Levenshtein distance (see page 206).

Let's begin with some definitions, which will help to express all ideas rigorously enough. As before, let  $\mathcal{U} = U_1, U_2, \dots, U_n$  be the set of identified coding units by all annotators. A unitary alignment  $\check{a}$  is a set of aligned coding units from different annotators. In every unitary alignment is at most one coding unit of each annotator. A complete alignment  $\bar{a}$  of coding units is a set  $\{\check{a}_1, \check{a}_2, \dots, \check{a}_p\}$  of unitary alignments if, first, every unit in  $\mathcal{U}$  is in exactly one unitary alignment and, second, every unitary alignment contains at least one unit of  $\mathcal{U}$ . Let further  $\delta_o(\bar{a})$  be the overall aggregated disagreement between all aligned coding units. This function measures the observed disagreement between annotations based on the alignment  $\bar{a}$ .

Based on which alignment should the observed disagreement  $D_o$  be calculated? The answer is straightforward: On the alignment that minimises  $D_o$ . More formally, let  $\mathcal{A}_{\mathcal{U}}$  be the set of all possible complete alignments of all coding units from  $\mathcal{U}$  and  $\bar{a}^*$  the alignment which minimises  $\delta_o$ , that is

$$\delta_o(\bar{a}^*) \leq \delta_o(\bar{a}) \text{ for all } \bar{a} \in \mathcal{A} \quad (5.8)$$

then the observed disagreement  $D_o$  is defined as

$$D_o := \delta_o(\bar{a}^*) \quad (5.9)$$

The elaboration of this idea demands defining  $\delta_o(\bar{a})$  in an appropriate way. The suggestion is to aggregate a pairwise distance (or dissimilarity)  $d(U_i, U_j)$  over all unit pairs in a unitary alignment and over all unitary alignments in the following way:<sup>251</sup>

$$\delta_o(\bar{a}) := \frac{2}{|\bar{a}|n_s(n_s - 1)} \sum_{\check{a}_i \in \bar{a}} \sum_{U_i, U_j \in \check{a}_i} d(U_i, U_j) \quad (5.10)$$

with  $n_s$  being the number of annotators.

Before properly defining  $d$ , we need to account for failed attempts to align coding units in the formalism. Since annotators might find coding units at different positions and identify a different number of coding units, not every coding unit can or should be aligned with a corresponding coding unit of every other annotator. In other words, a coding unit of one annotator may not be aligned with any coding unit of another annotator. Consider, for instance, the simple case depicted in Figure 5.16. It seems natural not to align  $U_1$  with  $U_3$  since it is better to align the former with  $U_2$ . But then there wouldn't be any other unit left that can be aligned with  $U_3$ .

However, if the alignment procedure is allowed to not align coding units with each other up to the point that there is only one coding unit in every unitary alignment, it will do exactly

<sup>251</sup>The normalisation is slightly different than in Mathet, Widlöcher, and Métivier (2015). Here, we take the average over unitary alignments in  $\bar{a}$  ( $|\bar{a}|$ ) and over unit pairs in an alignment with  $n_s$  units.

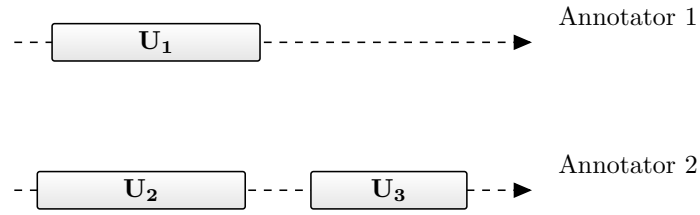


ABBILDUNG 5.16

This example illustrates that not every coding unit ( $U_3$ ) can be aligned with a coding unit of every other annotator.

that. If there is only one coding unit and, hence, no unit pair in a unitary alignment, the aggregated distance of unit pairs within the unitary alignment will be zero. The problem is that not aligning a coding unit of one annotator with any coding unit of another annotator comes with no costs. Accordingly, we need to adapt the formalism to incorporate such costs.

To that end, so-called pseudo units  $U_0$  are introduced, which fill empty slots in unitary alignments (see Figure 5.17). In this way, every unitary alignment has the same number of units, and the costs of not aligning a coding unit  $U_i$  with any coding unit of another annotator are charged with  $d(U_i, U_0)$ .

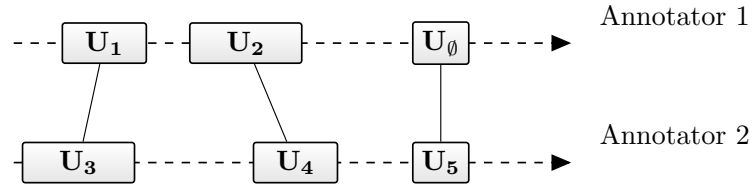


ABBILDUNG 5.17

Simple example of an alignment. Vertical lines connecting coding units indicate unitary alignments.  $U_1$  is aligned  $U_3$ ,  $U_2$  with  $U_4$  and  $U_5$  with a pseudo unit.

The dissimilarity function  $d$  is required to be symmetric ( $d(U_i, U_j) = d(U_j, U_i)$ ), and the constant  $\delta_0$  quantifies the costs of aligning a coding unit with the pseudo unit:<sup>252</sup>

$$d(U_i, U_0) := \delta_0 \text{ and } d(U_0, U_0) := \delta_0 \quad (5.11)$$

Mathet, Widlöcher, and Métivier (2015) define  $d$  as a linear combination of two dissimilarity types: Positional dissimilarities  $d_{pos}$  and categorical dissimilarities  $d_{cat}$ .<sup>253</sup>

$$d(U_i, U_j) := \alpha d_{pos}(U_i, U_j) + \beta d_{cat}(U_i, U_j) \quad (5.12)$$

The positional dissimilarity measures differences with regard to the locations of  $U_1$  and  $U_2$ . Let  $start(U)$  and  $end(U)$  be the start and end points of a unit  $U$ . The positional dissimilarity  $d_{pos}$  is then defined by

<sup>252</sup>Mathet, Widlöcher, and Métivier (2015) set  $\delta_0 = 1$  in their implementation.

<sup>253</sup>The authors use  $\alpha = \beta = 1$  (Mathet, Widlöcher, and Métivier 2015, 453).

$$d_{pos}(U_1, U_2) = \left( \frac{|start(U_1) - start(U_2)| + |end(U_1) - end(U_2)|}{end(U_1) - start(U_1) + end(U_2) - start(U_2)} \right)^2 \delta_0 \quad (5.13)$$

Categorical dissimilarity, on the other hand, quantifies differences in the categorisation of aligned units. Similarly to Krippendorff's family of  $\alpha$  coefficients, it can be defined in variable ways to deal with different data types (i.e., nominal, ordinal, interval and ratio). In the case of nominal data, the authors suggest using a simple binary function. Let  $cat(U)$  denote the category of code unit  $U$  then the categorical dissimilarity is defined as follows:

$$d_{cat}(U_1, U_2) := \begin{cases} \delta_0 & \text{if } cat(U_1) \neq cat(U_2) \\ \delta_0 & \text{if } U_1 = U_\emptyset \text{ or } U_2 = U_\emptyset \\ 0 & \text{if } cat(U_1) = cat(U_2) \end{cases} \quad (5.14)$$

There are two problems using the described alignment approach in the context of CAAS. First, argumentative units might be aligned even if they do not overlap at all. Second, the approach will align argumentative units that correspond to a mere difference in granularisation.

In some cases, the  $\gamma$  approach will align non-overlapping coding units (Mathet, Widlöcher, and Métivier 2015, 450). More specifically, if there are no better alternatives for aligning a coding unit with another one, it will rather be aligned with a non-overlapping coding unit than with a pseudo unit. For instance, in the example of Figure 5.18 all three coding units will be aligned with each other despite their lack of overlap.

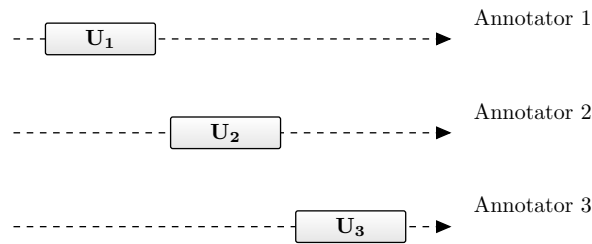


ABBILDUNG 5.18

Example of non-intersecting units, which will be aligned in the  $\gamma$  approach.

Why is that a problem in the context of CAAS? Mathet, Widlöcher, and Métivier (2015) provide two examples in which such an alignment is expected and conclude that “units by different annotators may correspond to the same phenomenon though they do not intersect” (Mathet, Widlöcher, and Métivier 2015, 450). Admittedly, the same can happen during the process of identifying argumentative units. The described annotation scheme does not erase all ambiguity on where argumentative units start and end. For instance, it can happen that someone interprets an explanatory or illustrative repetition of an argument as the argument itself, whereas another annotator only identifies the first occurrence of that argument as the argumentative unit. The resulting units would not overlap, although they refer to the same argument.

First and foremost, this example indicates that the question of aligning units cannot successfully be reduced to a consideration of the position of coding units. Sometimes, the meaning of the identified text segments is crucial to decide whether they represent the same phenomenon.

As described above, due to feasibility, I rely on an approach that aligns units according to their positions. Admittedly, this simplification will sometimes produce erroneous alignments. However, if we decide to align coding units based on syntactic criteria, it is not that relevant that *there are* cases of non-overlapping units that should be aligned according to semantic criteria. Instead, we should ask *how often* such configurations occur. How should we design the syntactic alignment procedure so that it approximates a hypothetical semantic procedure sufficiently? How can we design it to produce as few mistakes as possible? Is the lack of overlap between coding units a good enough indicator for them not representing the same phenomenon? I side here with other authors that answer this question in the affirmative (e.g., Duthie et al. 2016; Krippendorff et al. 2016).<sup>254</sup>

As the above-discussed accounts of unitising and aligning coding units, the  $\gamma$  approach is not able to properly (not) align coding units in the case of a diverging choice as to the granularisation of argumentation. The case of Figure 5.19a can be interpreted as two annotators disagreeing in the granularisation of argumentation. The configuration suggests that the second annotator identifies two different arguments, which the first interprets as one compound argument. Accordingly, the alignment should neither align  $U_1$  with  $U_2$  nor with  $U_3$ .

However, the  $\gamma$  approach will align  $U_1$  with  $U_2$ . Even though the approach is very versatile, it is not possible to nudge it into the desired behaviour by adjusting the given free parameters without producing other problems. A minimalistic example illustrates this difficulty. Case *B* (5.19b) is a slight modification of Case *A* (5.19a), which differs from the former only with regard to the location of  $U_3$ . In the first case, we want none of the units to be aligned with each other. However, Case *B* suggests aligning  $U_1$  with  $U_2$  due to their high overlap. In contrast to Case *A*, there is no reason to assume that  $U_2$  is part of a partitioning of  $U_1$ . The problem is that the alignment procedure cannot distinguish both cases. Accordingly, it will, in both cases, either align or not align  $U_1$  with  $U_2$ .<sup>255</sup>

## 5.2.2 ALIGNING ARGUMENTATIVE UNITS

The following alignment procedure is based on the discussed approach by Mathet, Widlöcher, and Métivier (2015). In particular, I will follow the idea of aligning units by optimising the aggregated overall disagreement  $\delta_o$  of the alignment. I will depart from the  $\gamma$  approach in two points. First, I will adapt their used pairwise distance function  $d$ . Second,

<sup>254</sup>Admittedly, this is an empirical hypothesis, which I cannot systematically substantiate here.

<sup>255</sup>The  $\gamma$  approach decides on an alignment by comparing the overall aggregated disagreement  $\delta_o$  of the different full alignments (see Definition 5.8). In the described situation we have to compare the alignment  $\bar{a}_1 := ((u_1, u_0), (u_2, u_0), (u_3, u_0))$  with the alignment  $\bar{a}_2 := ((u_1, u_2), (u_3, u_0))$ . What we need for the desired alignment is that  $\delta_o(\bar{a}_1) < \delta_o(\bar{a}_2)$  in Case *A* and  $\delta_o(\bar{a}_1) > \delta_o(\bar{a}_2)$  in Case *B*. However, the only difference between both cases is the position of  $U_3$ , and  $U_3$  contributes to  $\delta_o$  only the costs  $d(U_3, U_0)$ , which are independent of  $U_3$ 's position (see Definition 5.10 and Definition 5.11). The overall disagreements  $\delta_o(\bar{a}_1)$  and  $\delta_o(\bar{a}_2)$  do, therefore, not vary between both cases—independent of how we set the parameters  $\alpha, \beta$  and  $\delta_0$ . Hence, the alignment will be same in Case *A* and *B*.

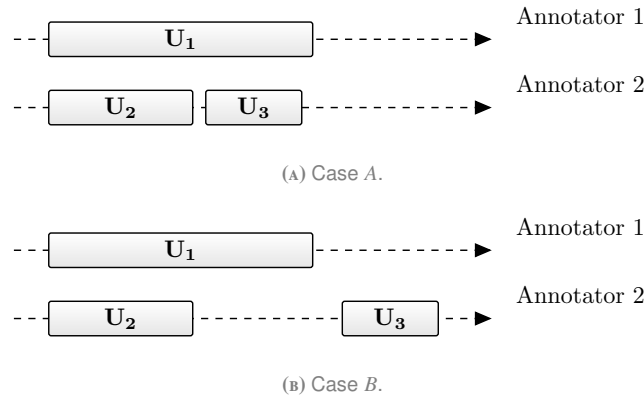


ABBILDUNG 5.19

Annotation example to illustrates difficulties of the  $\gamma$  approach to deal with diverging granularisations of argumentation.

I will formulate some constraints the alignment candidates have to fulfil to participate in the contest of optimising  $\delta_o$ .

In the  $\gamma$  approach, the pairwise disagreement function  $d$  contains a positional and a categorical part (see Definition 5.12). In other words, the alignment will depend on the location of coding units and their categorisation. Annotators who use the suggested annotation scheme (see Appendix A.1) do not attribute any category values to argumentative units besides identifying them as justificatory relevant. We could, alternatively, consider the annotation of justificatory relations to be relevant for the alignment procedure, for instance, by using the introduced graph-edit distance (see Definition 5.7). This would, however, complicate the optimisation procedure considerably.<sup>256</sup> Therefore, I will ground the argumentative units' alignment solely in terms of their positions and widths.

More specifically, I will suggest a simplified measure for the positional (dis-)agreements between argumentative units. Let  $|AU|$  be the length of an argumentative unit  $AU$  in terms of its character tokens and  $\max(|AU_1|, |AU_2|)$  the maximum of the lengths of  $AU_1$  and  $AU_2$ . I will define  $|AU_1 \cap AU_2|_{\max}$  as the length of the overlap between  $AU_1$  and  $AU_2$  normalised by the width of the larger code unit:

$$|AU_1 \cap AU_2|_{\max} := \frac{|AU_1 \cap AU_2|}{\max(|AU_1|, |AU_2|)} \quad (5.15)$$

The adapted pairwise disagreement  $d$  is defined by:

$$d(AU_1, AU_2) := \begin{cases} \delta_0 & \text{if } AU_1 = U_0 \text{ or } AU_2 = U_0 \\ (1 - |AU_1 \cap AU_2|_{\max})^2 \delta_0 & \text{otherwise} \end{cases} \quad (5.16)$$

<sup>256</sup>Instead of having a contributing term  $d_{cat}(U_i, U_j)$  of unit pairs, we would have a general contributing term of the form  $\delta_{graph}(\bar{a})$ . From an algorithmic perspective, the  $\gamma$  approach is easy to implement since it relies on a linear optimisation (Mathet, Widlöcher, and Métivier 2015, 456–58). Including the calculation of a graph-wide disagreement into the alignment procedure would, however, transform the linear problem into a non-linear one.

The overall aggregated (positional) disagreement within an alignment can now be written as:<sup>257</sup>

$$\delta_o(\bar{a}) := \frac{2}{|\bar{a}|n_s(n_s - 1)} \sum_{\check{a}_i \in \bar{a}} \sum_{AC_i, AC_j \in \check{a}} d(AC_i, AC_j) \quad (5.17)$$

However, this alternative to defining the overall aggregated disagreement within an alignment does not account for the discussed difficulties. In particular, the requirement of not aligning argumentative units representing diverging granularisations cannot be satisfied by substituting the used pairwise-disagreement function or tweaking the available parameters.<sup>258</sup> The general idea of solving the discussed problem is, fortunately, quite simple: Instead of searching for an optimal alignment  $\bar{a}^*$  among all possible alignments  $\mathcal{A}_U$ , we confine the search area to a subset  $\mathcal{A}_U^* \subset \mathcal{A}_U$ . In other words, coding units should be aligned to optimise  $\delta_o$  over  $\mathcal{A}_U^*$  instead of over  $\mathcal{A}_U$ .

The idea is then to define  $\mathcal{A}_U^*$  such that the excluded alignments ( $\mathcal{A}_U \setminus \mathcal{A}_U^*$ ) comprise (and only comprise) those alignments that do not satisfy the formulated requirements of aligning argumentative units. In the last section, I formulated two such desiderata: Argumentative units should not be aligned with each other if, first, they do not overlap and, second, if they represent different granularisations of one argumentation.

The first requirement does not need any further clarification. However, I suggest formulating an even stronger requirement. It is reasonable to demand that argumentative units should overlap to a greater extent than an arbitrarily small one. Consider the units in Figure 5.20a. Although they intersect to some extent, this seems to be a case of different argumentative units whose boundaries overlap due to interpretational differences. Even in the absence of other units, they should, therefore, not be aligned with each other.

The other two Figures 5.20b and 5.20c suggest that it is not the overlap relative to the larger unit that counts, which is roughly the same in all figures. The units in both figures can be regarded as presenting the same phenomenon and should, therefore, be aligned with each other—in contrast to the units in Figure 5.20a. What the latter distinguishes from the former two is the overlap of the units relative to the smaller unit. In other words, the extent to which a smaller unit is contained in a bigger one relative to the smaller one constrains whether both should be aligned. More formally, let

$$|AU_1 \cap AU_2|_{min} := \frac{|AU_1 \cap AU_2|}{\min(|AU_1|, |AU_2|)} \quad (5.18)$$

and  $OV(\bar{a})_{ref}$  the property that each unit pair in each unitary alignment of  $\bar{a}$  satisfies this minimal-overlap requirement, which is defined by:  $OV_{req}(\bar{a})$  if and only if:

$$\text{For all } \check{a} \in \bar{a} : |AC_i \cap AC_j|_{min} \geq \alpha_{\cap} \text{ for all } AC_i, AC_j \in \check{a} \quad (5.19)$$

<sup>257</sup>I introduce this measure first and foremost to facilitate an alignment of coding units. It can, of course, also be used to measure positional dissimilarities based on an alignment.

<sup>258</sup>See Footnote 255.

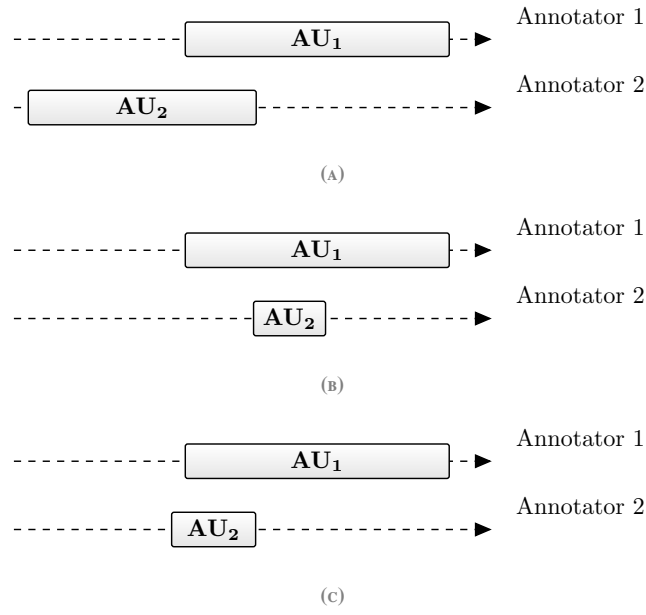


ABBILDUNG 5.20

Examples of relative overlap. The overlap of  $AU_2$  relative to the larger unit  $AU_1$  (i.e.,  $|AU_1 \cap AU_2|_{max}$ ) is in all three examples the same.

then we demand that

$$\text{For all } \bar{a} \in \mathcal{A}_{\mathcal{U}}^* : OV_{req}(\bar{a}). \quad (5.20)$$

If we only want to exclude the alignment of non-overlapping units, we can set  $\alpha_{\cap}$  to zero. If we want to impose stronger requirements, as suggested, we have to set  $0 < \alpha_{\cap} \leq 1$ .<sup>259</sup>

The second requirement is more complicated since we have to specify under which conditions a set of argumentative units of one annotator represents a different granularisation of the unit of another annotator. A natural thought is to think about differences in granularisation in terms of partitioning a long interval into smaller ones. The units  $P := \{AU_2, AU_3, AU_4\}$  in Figure 5.21a illustrate what can be dubbed a clean partitioning of the longer unit  $AU_1$ . The elements in  $P$  do not overlap, and their union coincides with  $AU_1$ . This example is a clear case of two annotators who differ in their interpretation of the individuation of argumentative components. Hence, we should demand that whenever units of one annotator form a clean partition of another annotator's unit, the former units should not be aligned with the latter.

While this requirement is a good starting point, it needs further refinements since it is too weak in its current form. There are other configurations that indicate divergent granularisations and do not violate the suggested criterion.

First, we must decide how to judge situations where we find partitions by putting together units from different annotators (see Figure 5.21b). I suggest interpreting such configurations similarly to indicate diverging granularisations. The alternative is to interpret such cases as

<sup>259</sup>Exact values must be found in practice. But a  $\alpha_{\cap} = 0.5$  seems to be a reasonable choice.



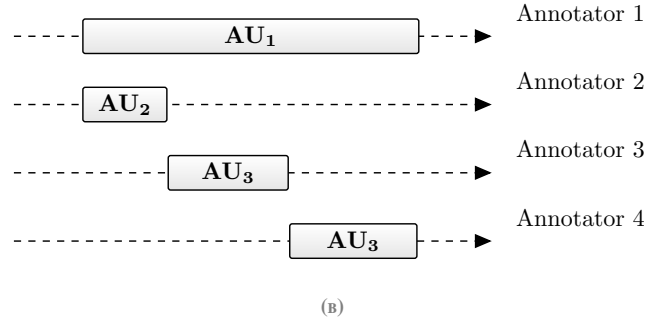
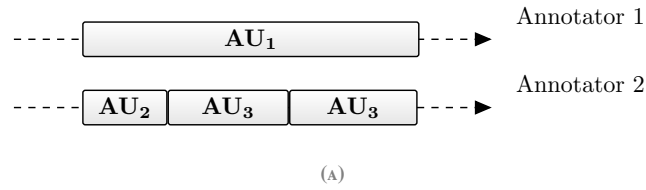


ABBILDUNG 5.21  
Examples of "clean" partitions.

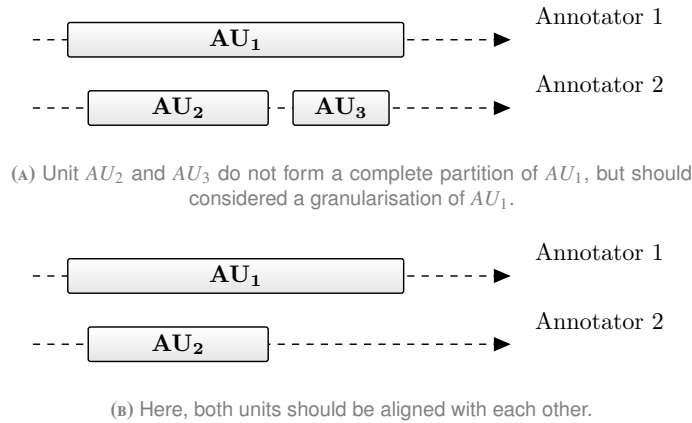


ABBILDUNG 5.22

a mere difference in identifying a unit's boundaries, which coincidentally looks similar to an alternative granularisation.<sup>260</sup>

Second, even if a set of units do not form a proper partition of a unit  $AU_1$ , in the sense of not completely covering  $AU_1$ , they might correspond to different granularisation of  $AU_1$ . In Figure 5.22a, it seems that the second annotator construes what the first one identified as one longer argumentative component ( $AU_1$ ) as two different argumentative components ( $AU_2$  and  $AU_3$ ). So, we should neither align  $AU_2$  nor  $AU_3$  with  $AU_1$  (even if  $AU_2$  does not violate the minimal overlap constraint).

Obviously, we need at least two argumentative units in  $\mathcal{U}$  to interpret them as being a different granularisation of another unit. If we remove  $AU_3$  in Figure 5.22a, we end up with a situation as depicted in Figure 5.22b. Without any other indication, the latter configuration should allow for an alignment of  $AU_1$  and  $AU_2$ .

<sup>260</sup>Similarly to the decision regarding non-overlapping units, this decision should be justified by empirical means, which is beyond this work.

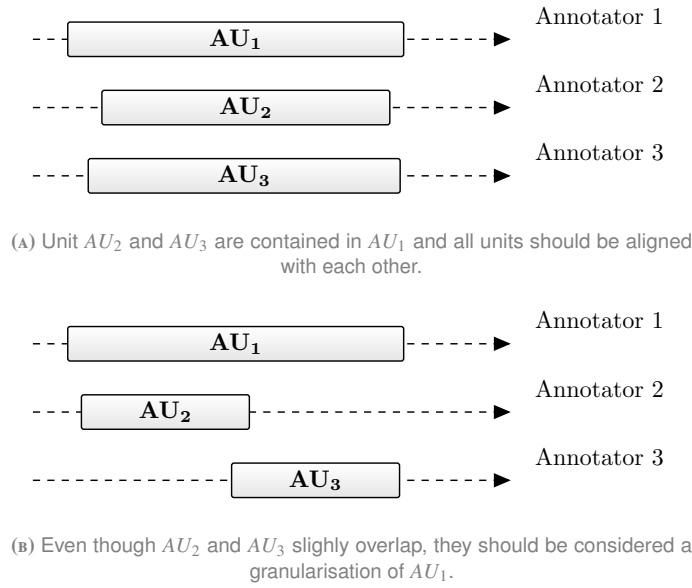


ABBILDUNG 5.23

But does every pair of argumentative units contained in another unit correspond to an alternative granularisation of the latter? Figure 5.23a shows this is not the case. The units  $AU_2$  and  $AU_3$  are both contained in  $AU_1$ . However, the configuration suggests that all units represent the same argumentative component, although the annotators disagree on its precise boundaries. Hence, we should interpret two coding units  $AU_i$  and  $AU_j$  only then as suggesting an alternative granularisation if their overlap is not too large (see Figure 5.23b). In other words, we demand that their overlap relative to the larger one of both is below some threshold  $\alpha_{\cap}^{max}$  ( $|AU_i \cap AU_j|_{max} \leq \alpha_{\cap}^{max}$ ).

Finally, we do not have to demand that two units are fully contained in what they appear to partition. Consider the case of Figure 5.24. Even if  $AU_2$  and  $AU_3$  are not fully contained in  $AU_1$ , they seem to represent a different granularisation of  $AU_1$ . This can be considered a result of two combined interpretational differences: A different choice regarding the granularisation and the units' boundaries. So, instead of requiring a full containment of  $AU_2$  and  $AU_3$  in  $AU_1$ , we demand a minimal containment above some threshold  $\alpha_{\cap}^{min}$  relative to themselves. More formally, let  $|AU_i \cap AU_j|_{AU_i} := \frac{|AU_i \cap AU_j|}{|AU_i|}$ , then we demand  $|AU_1 \cap AU_2|_{AU_2} \geq \alpha_{\cap}^{min}$  and  $|AU_1 \cap AU_3|_{AU_3} \geq \alpha_{\cap}^{min}$ .

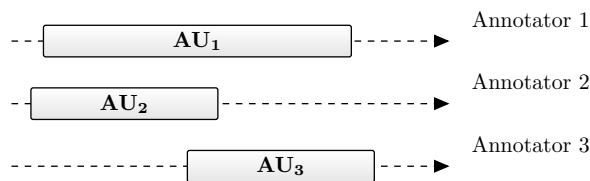


ABBILDUNG 5.24

Even though  $AU_2$  and  $AU_3$  are not contained in  $AU_1$ , they should be considered a granularisation of  $AU_1$ .

Putting all these considerations together, we can now define  $\mathcal{A}_{\mathcal{U}}^*$  in the following way. We will say that two units  $AU_i$  and  $AU_j$  are a granularisation of a third unit  $AU_k$ , in short

$GR(\{AU_i, AU_j\}, AU_k)$ , if and only if

$$\begin{aligned} 1. & |AU_i \cap AU_j|_{\max} \leq \alpha_{\cap}^{\max}, \\ 2. & |AU_k \cap AU_i|_{AU_i} \geq \alpha_{\cap}^{\min} \text{ and} \\ 3. & |AU_k \cap AU_j|_{AU_j} \geq \alpha_{\cap}^{\min} \end{aligned} \quad (5.21)$$

Let now  $GR_{req}(\bar{a})$  be the property of  $\bar{a}$  to satisfy the requirement of not aligning argumentative units that correspond to alternative granularisations. More specifically, will say that  $GR_{req}(\bar{a})$  if and only if the following holds:

$$\begin{aligned} & \text{If there are } \check{a} \in \bar{a}, AU_k \in \check{a} \text{ \& } AU_i, AU_j \in \mathcal{U} \\ & \text{such that } GR(\{AU_i, AU_j\}, AU_k) \\ & \text{then } AU_i \notin \check{a} \text{ and } AU_j \notin \check{a}. \end{aligned} \quad (5.22)$$

We can now define  $\mathcal{A}_{\mathcal{U}}^*$  by

$$\mathcal{A}_{\mathcal{U}}^* := \{\bar{a} \in \mathcal{A}_{\mathcal{U}} | OV_{req}(\bar{a}) \text{ and } GR_{req}(\bar{a})\} \quad (5.23)$$

The suggested alignment procedure, which searches for an  $\bar{a}^* \in \mathcal{A}_{\mathcal{U}}^*$  that optimises  $\delta_o$ , can be implemented as a linear optimisation problem.<sup>261</sup> The resulting alignment can then be used to calculate the annotator wide disagreement of relational categorisation (see Definition 5.7) and will be further used in the next sections to conceptualise a statistical model of CAAS.

### 5.3 A PROBABILISTIC CONCEPTUALISATION OF CAAS

In the last chapter, I motivated that the analysis of argumentation structure is hermeneutically underdetermined. The identification of argumentative units and their justificatory relations to each other can differ between different annotators. I argued that argumentation theory does not help to eliminate this degree of interpretation in every case.

In Chapter 3, I proposed a general framework that deals with such an irreducible degree of interpretation in content analysis without eschewing the requirement of reliability. I suggested conceptualising the space of all interpretations by a statistical model, which allows the researcher to employ statistical methods to infer something about the phenomena of interest. In analogy to the error-based approach in natural sciences, I provided an outline of how significance testing can be used to decide whether observed differences between two samples are significant enough to justify the conclusion that there is a difference between the phenomena of interest.

In this section, I intend to elaborate on how these preliminary ideas can be applied in the context of analysing argumentation structure. The general idea is this: The result of annotating the argumentation structure of a text is represented by an argumentation

<sup>261</sup> Similar to Mathet, Widlöcher, and Métivier (2015), Section 5.

graph—that is, a directed graph consisting of argumentative components as well as attack and support relations between them. Since such an annotation might be hermeneutically underdetermined, different annotators might end up with different argumentation graphs without doing anything wrong. The space of all interpretations is then supplemented with a probability function, which assigns a probability value to each argumentation graph. In other words, the data-making process for a specific text is modelled by a probability function over the outcome space of coding results.

On an abstract level, this probability function represents phenomena, which the social scientist might want to investigate. Or to put it differently, the probability function can be used to infer relevant information about social phenomena. A simple example illustrates how this abstract picture translates to concrete questions that might be of relevance in CAAS. For instance, the researcher might want to know the number of argumentative components in a text (perhaps in comparison to another text). However, confronted with hermeneutical underdetermination there is not one unique number that represents *the* number of argumentative components, but different numbers since the number might vary among different interpretations. Hence, there is not a properly defined phenomenon that corresponds to *the* number of argumentative components.

However, that does not mean that every natural number is on par with every other number in this example. Rather, by providing a probabilistic conceptualisation of the data-making process the numbers might differ in their likelihood. The probability function that assigns probability values to argumentation graphs can be used to calculate probability values for the different numbers. In this way the probabilistic conceptualisation can be used to properly define a phenomenon that is the next best cousin to ‘*the* number of argumentative components’, namely, the expected number of argumentative components—that is, the probability-weighted mean of these numbers. In a similar vein, we cannot properly compare *the* number of argumentative components between two texts, but we can compare the expected number of argumentative components.

This reasoning can be generalised to other argumentative properties. Besides the number of argumentative components, the researcher might want to estimate the argumentative complexity in a text, which might be measured by counting the justificatory relations in relation to the number of argumentative components (or some other graph measure), or they might want to know to what extent a text provides a balanced exposition of reasons for and objections against the central claim. From a mathematical perspective, these properties are functions that attribute a number to an argumentation graph. Again, there is not necessarily one such number for a specific text in the case of hermeneutical underdetermination. But as before, the phenomenon can be identified with the expected number of the respective measure. In this sense, many relevant phenomena can be captured even if there is some interpretational leeway concerning the argumentation structure.

The probability function is unknown before the data-making process—that is, prior to the annotation—and must be statistically inferred from a sample and by using statistical assumptions about the data-making process. The required sample is generated by independently working annotators and represented by a set of argumentation graphs. If annotators are randomly chosen from the population of properly instructed annotators, different statistical methods can be used to infer something about the phenomenon of

interest.

The elaboration of this general idea demands specifying the following elements.

First, we need a mathematical specification of the outcome space and the statistical model (5.3.1). Statistical inferences deduce properties of a probability function from an observed sample. These inferences are usually based on statistical assumptions, which are referred to as the statistical model. For instance, it might be assumed that certain events are statistically independent. If you toss two different coins, you can usually assume that the probabilities for the outcomes of the first coin do not depend on the probabilities for the second coin. The outcome space describes the set of all possible outcomes of the data-making process. In the simple example we have four outcomes, which can be written as the Cartesian product of the two different outcome spaces: The outcome space  $\Omega_1$  of the first coin with two outcomes  $H_1, T_1$  ( $H$  denotes heads and  $T$  tails) and the outcome space  $\Omega_2 = \{H_2, T_2\}$  of the second coin. Now, we can write the whole outcome space as  $\Omega_{12} = \{(H_1, H_2), (H_1, T_2), (T_1, H_2), (T_1, T_2)\}$ . The construction of the outcome space and the formulation of the statistical model might be intertwined. In the simple example, the decomposition of  $\Omega_{12}$  into the Cartesian product  $\Omega_1 \times \Omega_2$  together with the independence assumptions imply that the (marginal) probabilities of events in  $\Omega_1$  are identical to probabilities of these events conditional on events in  $\Omega_2$ :  $P(H_1) = P(H_1|T_2) = P(H_1|H_2)$ . The same applies to marginal probabilities in  $\Omega_2$ . Such a decomposition of  $\Omega_{12}$  simplifies the statistical treatment of the case considerably. In our case, we will equate the outcome space with a finite set of argumentation graphs. Similarly to the given example, we will try to rewrite this space as a cartesian product in a way that similar independence assumptions can be made.

Second, we have to choose a specific statistical method that is suitable for our research context (5.3.2). As elaborated in Chapter 3, satisfying reliability—in the form of weak phenomenon sensitivity—demands deciding whether observed differences are significant enough to justify the conclusion that there is a difference in the phenomena. Suppose, for instance, the amount of argumentative components is of relevance and we want to investigate whether two texts differ with respect to their expected number of argumentative components. To that end, we have to generate a sample of argumentation graphs for each text and decide whether the observed difference between the mean number of argumentative components is significant enough. We will closely follow the description from Section 3.4.2 by applying the method of significance testing. The main conceptual challenge is that the null hypothesis—that is, the hypothesis that both texts are not different—is complex. In contrast to a simple hypothesis, a complex hypothesis is not represented by one specific probability function but by a whole set of probability functions. Consequently, the concept of the  $p$ -value must be generalised to such cases.

Finally, the mathematical considerations to describe the outcome space and the statistical model as well as the elaboration of the statistical inference method will reveal a certain complexity. Both the size of the model's parameter space and the size of the sample space will render the actual calculation of  $p$ -values computationally complex. I will sketch some solutions for how to deal with these complexities (5.3.3). However, it is beyond this work to sufficiently solve these problems. Instead of demonstrating that the discussed statistical methods can be applied to realistic cases, I will point to further needed research.

### 5.3.1 A STATISTICAL MODEL FOR CAAS

Let  $\Omega_\tau$  the set of all argumentation graphs that represent a correct interpretation of the argumentation structure of a specific text.  $P(\tau)$  will denote the probability of an argumentation graph  $\tau$  to be the coding result of a randomly chosen annotator.<sup>262</sup> There are two important problems we have to solve for the formulation of a statistical model. First, how do we determine the set of all correct interpretations? Second, are there any independence assumptions we can use to rewrite the outcome space as a Cartesian product such that the subspaces are statistically independent of each other?

Let us first address the problem of determining  $\Omega_\tau$ . The annotation scheme formulates and clarifies how annotators should analyse the argumentation structure of a text. In a way, it determines all relevant standards for a correct interpretation. But  $\Omega_\tau$  is not determined by the annotation scheme alone since  $\Omega_\tau$  will differ between different texts. Hence,  $\Omega_\tau$  must be constructed by (correctly) applying the annotation guidelines to the text at hand and varying everything that is not uniquely determined. Can we perhaps use the coding results of annotators to construct  $\Omega_\tau$ ? After all, we expect properly instructed annotators to generate argumentation graphs that represent correct interpretations of the text's argumentation structure. There are two immediate worries.

First, annotators can make mistakes due to various reasons. They might misinterpret the coding instructions, suffer from fatigue, lack focus or what have you. Accordingly, coding results would have to be checked for correctness and incorrect argumentation graphs would have to be removed. This is surely feasible in the case of obvious mistakes that can be detected automatically by an algorithm. However, usually mistakes demand a human analysis of the text and the coding results. In practice, such laborious assessments are too time-consuming. And who should do this? Other annotators, who might make mistakes themselves? So instead of checking every coding result for correctness, the content analyst has to promote correctness by indirect means. For instance, they can increase the likelihood of correct annotations by eliminating typical causes for mistakes, such as fatigue. Although these measures cannot guarantee that annotators do not make mistakes, they are often the best that can be done in practice. So while this worry must be taken seriously, I do not take it as a deciding objection against the suggestion to construct  $\Omega_\tau$  based on coding results.

The second worry is that a finite amount of coders might miss elements of  $\Omega_\tau$  even if we aggregate their results in some way. How many replicated annotations are necessary to guarantee that every correct interpretation is among them? We need at least as many annotations as there are correct interpretations. But how many correct interpretations are there? We can safely assume that  $\Omega_\tau$  is finite. There will be a finite amount of argumentative components in a text, and, accordingly, only a finite amount of possibilities to connect them via justificatory relations. But if  $\Omega_\tau$  is finite, then there will be a non-empty subset of  $\Omega_\tau$  such that the probability for each argumentation graph  $\tau$  within that subset is greater than zero ( $P(\tau) \geq 0$ ).<sup>263</sup> For these argumentation graphs, we can expect that they will

<sup>262</sup>To be more precise,  $P$  is a probability mass function. The probability function should be defined over an algebra of the outcome space—the event space. Since the outcome space is finite, I will not bother to thoroughly distinguish outcome space and event space.

<sup>263</sup>This is a simple implication of the fact that the probability over a finite outcome space must satisfy  $\sum_{\tau \in \Omega_\tau} P(\tau) = 1$ .

eventually be the result of an annotation if we increase the number of repeated annotations further and further.<sup>264</sup>

For the remaining argumentation graphs—that is those  $\tau \in \Omega_\tau$  with  $P(\tau) = 0$ —we could argue that they are irrelevant. We might dispose of them in the construction of  $\Omega_\tau$ . Although they represent correct interpretations, they are irrelevant since they are (almost) never the result of an annotation. To not consider such an argumentation graph  $\tau$  in the construction of  $\Omega_\tau$  is from a practical point of view the same as including  $\tau$  into  $\Omega_\tau$  and setting  $P(\tau) = 0$ .

These considerations suggest that we could construct  $\Omega_\tau$  by accumulating an infinite amount of annotations. For obvious practical reasons, this is not feasible. Not only is it impossible to reproduce an infinite amount of annotations, but it also is often not even possible to reproduce many of them since the instruction of coders and the actual annotation is time-consuming. I can think of two alternatives to construct  $\Omega_\tau$ .

First, we might use an algorithmic procedure to construct  $\Omega_\tau$ . The basic idea is to automatically identify those text segments that are possibly argumentative units from a linguistic point of view. Argumentative components are expressed by statements. Hence, only text segments that can be interpreted as referring to or representing a statement, can be argumentative components. Although argumentative components cannot be assumed to correspond to exactly one sentence, not every arbitrary word token or every aggregate of contiguous word tokens is a potential candidate for being an argumentative unit.

Let's suppose we have such an algorithm. We might now use all possible subsets of these units together with all combinatorial possibilities to connect them via justificatory relations as a basis to construct  $\Omega_\tau$ . Admittedly, such construction will most probably include a lot of incorrect interpretations. But even if this problem could be resolved, the sheer size of the resulting outcome space would prevent us from using  $\Omega_\tau$  in any practical context.<sup>265</sup>

The second alternative is to employ some gold standard to construct  $\Omega_\tau$ : A super annotator who is able to determine the set of all correct interpretations by analysing the text at hand. However, this suggestion is faced with practical and methodological difficulties.<sup>266</sup> Although I am inclined to think that these hurdles can be cleared to some extent, I propose to do without a super annotator and suggest a simpler and more pragmatic solution. We will use the given annotations of a finite set of annotators to estimate  $\Omega_\tau$ .

<sup>264</sup>This statement can be given a precise mathematical form and can then be proved by the law of large numbers (see, e.g., DeGroot and Schervish (2012), Chapter 6).

<sup>265</sup>Suppose we have a small text of say 10 sentences. At least every sentence in a text is a candidate to be an argumentative unit. Elements in  $\Omega_\tau$  are argumentation graphs. So how many argumentation graphs can we construct with 10 potential argumentative units? An argumentation graph is a set of unit pairs each of which corresponds to two argumentative components that are connected by a justificatory relation. For simplicity let's assume there is only one such justificatory relation. There are  $\binom{10}{2} := \frac{10!}{k!(n-k)!} = 45$  unit pairs that can be constructed with the 10 sentences. Every subset of unit pairs from these 45 unit pairs corresponds to one argumentation graph. This leads to  $2^{45}$  (roughly  $3 \cdot 10^{13}$ ) argumentation graphs.

<sup>266</sup>This task is from a methodological point of view a content analysis. Accordingly, we would have to formulate coding instructions for analysing the space of all correct annotations of a text's argumentation structure. What is more, the resulting annotations would have to be similarly checked for reliability. Additionally, one could argue that the correctness question reemerges at this level anew since these super annotators might make mistakes as well. Finally, we would have to think about a proper mathematical representation of the space of correct interpretations. In Cacean (2020), I suggested a solution for the last challenge.

Let  $\hat{\Omega}_\tau$  be our estimate of  $\Omega_\tau$  based on a finite sample of coding results. The construction of  $\hat{\Omega}_\tau$  will be driven by the following considerations: We will, first, try to avoid including incorrect categorisations. Second, we want to include as many correct categorisations as possible. Finally, we will prefer to minimise the size of  $\hat{\Omega}_\tau$ , which will be important to perform calculations that are necessary to apply statistical methods. There is, however, as we will see, a trade-off between these desiderata. So, how do we construct  $\hat{\Omega}_\tau$ ?

The leading idea is to base the construction of  $\hat{\Omega}_\tau$  on the argumentation graphs that represent the coding results. In other words, we will assume an alignment of argumentative units that determines which argumentative units of different annotators represent the same argumentative component (see 5.2). Consequently, the positions of argumentative units will only play an indirect role in the construction of  $\hat{\Omega}_\tau$  inasmuch as they are relevant for the alignment procedure. We now have to think about how to construct  $\hat{\Omega}_\tau$  based on the argumentation graphs  $\tau$  generated from the sample of coding results.

Instead of taking every unit *candidate* as a possible argumentative component—as suggested above—we will base  $\hat{\Omega}_\tau$  on those argumentative components that were discovered by annotators. That is, we consider the components that we find in the sample  $\tau$  as our best estimate to identify all argumentative components. Let  $\mathcal{AC}_\tau$  be the set of  $n_\tau$  argumentative components by pooling together every component of the argumentation graphs in  $\tau$ . We can now construct an argumentation graph by forming pairs of units from  $\mathcal{AC}_\tau$  and by connecting the elements of each pair by justificatory relations. This suggests the following construction of  $\hat{\Omega}_\tau$  by  $\tau$ : Let  $\Omega_{ij}$  be the possible outcomes of how two components  $AC_i$  and  $AC_j$  can be related to each other. Then we can construct  $\hat{\Omega}_\tau$  as the cartesian product  $\Omega_{12} \times \Omega_{13} \times \dots \Omega_{1n_\tau} \times \Omega_{21} \times \Omega_{23} \times \dots$  of every pair of non-identical components in  $\mathcal{AC}_\tau$ .

For instance, let there be  $n_\tau = 4$  different argumentative components  $AC_1, AC_2, AC_3$  and  $AC_4$  in a sample  $\tau$ . The argumentation graph depicted in Figure 5.25 can now be written as the outcome:  $(AC_1 \rightarrow AC_2, AC_3 \rightarrow AC_2, AC_2 \rightsquigarrow AC_4)$ .

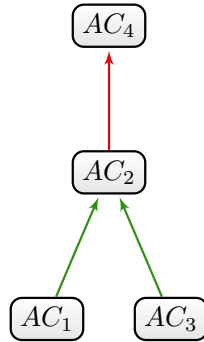


ABBILDUNG 5.25  
Simple argumentation graph.

But how many outcomes are in each subspace  $\Omega_{ij}$ ? There are four different possibilities to connect two components  $AC_1$  and  $AC_2$  by one relation:  $AC_1 \rightarrow AC_2$ ,  $AC_1 \rightsquigarrow AC_2$ ,  $AC_1 \leftarrow AC_2$  and  $AC_1 \leftrightsquigarrow AC_2$ . If we include the possibility that they are not connected at all, we arrive at five different possibilities. But isn't it possible to connect two components by more than one justificatory relation or even three? There are, in fact, only seven possibilities how arguers can connect two components by justificatory relations in a



rational way, two of which connect them by two relations (see A.1.7). However, we cannot exclude the other combinatorial possibilities from the outset. What the coder is supposed to annotate are intended justificatory relations (see page 250). And it can, in principle, happen that the provided linguistic cues must be interpreted as indicating, for instance, three justificatory relations between two argumentative components. Hence, if we want to include every outcome that is in principle possible and allowed according to the annotation scheme, we would have to include all 16 combinatorial possibilities to connect two components.<sup>267</sup>

How does this suggestion perform with respect to the formulated desiderata? Admittedly, the estimation of  $\Omega_\tau$  might miss some argumentative components. All coders may decide to not annotate a text segment as justificatory relevant, even if it would be a correct interpretation to do so. At least the suggested construction of  $\hat{\Omega}_\tau$  cannot miss any justificatory relations between the identified argumentative components since all combinatorial possibilities are included. However, we might worry that  $\hat{\Omega}_\tau$  includes a lot of argumentation graphs that do not represent correct interpretations. The construction of the subspaces  $\Omega_{ij}$  was motivated by not using a super annotator. Consequently, we included every combinatorial possibility to connect argumentative components. But the text at hand might exclude many of these relations. We can surely assume that in most cases an interpretation that connects two argumentative components by, for instance, three justificatory relations is not correct. Most people are sufficiently competent to not claim any nonsensical combinations of justificatory relations. It would therefore not be charitable to interpret them in that way.  $\hat{\Omega}_\tau$  will, therefore, most likely contain incorrect interpretations.

In some way, the intrusion of incorrect interpretations into  $\hat{\Omega}_\tau$  is not a problem. If we assume that coders annotate correctly, incorrect elements will never occur in annotation samples. In other words, their observed frequencies will be zero. Accordingly, the estimated probabilities for these incorrect outcomes will be comparably small.<sup>268</sup> Hence, when balancing to miss correct interpretations with the intrusion of incorrect elements into  $\hat{\Omega}_\tau$ , we might think to play it safe by preferring to construct  $\hat{\Omega}_\tau$  rather bigger than smaller.

However, this strategy comes at the cost of combinatorial complexity. The worry is that the number of possible outcomes is vast even if  $n_\tau$  is comparably small. Let's reconsider the simple example to have four argumentative components. There are  $\binom{n_\tau}{2}$  ways to produce pairs of elements from  $n_\tau$  elements. In the example with four elements, we have already six different pairs. Consequently, we have  $16^6 = 16.777.216$  combinatorial possibilities to construct argumentation graphs—even though we have only four nodes! If we would confine the outcomes in  $\Omega_{ij}$  to the seven reasonable outcomes, we still would have  $7^6 = 117.649$  argumentation graphs.

Accordingly, I will prefer an alternative construction of  $\hat{\Omega}_\tau$  that includes fewer elements. In particular, I suggest estimating correct justificatory relations similar to the estimation of

<sup>267</sup>There are  $\sum_{i=0}^4 \binom{4}{i} = 16$  ways of how two argumentative components can be connected by the four justificatory relations: One possibility to connect them by no relation, four to connect them by one relation, six to connect them by two relations, four to connect them by three relations and one to connect them by all four relations.

<sup>268</sup>A further specification of this statement depends on the used statistical method. Suffice it to say that observed frequencies are good point estimates (see page 76). In this case, the estimated probabilities would equal zero. However, the corresponding confidence intervals will usually contain other values as well.

argumentative components. The preceding construction idea had to rely for the estimation of nodes ( $\mathcal{AC}_\tau$ ) on what can be found in the sample  $\tau$ . The same can be done with justificatory relations: If we find in the whole sample  $\tau$  only a subset of, say, three relations (out of the 16 possibilities) between two argumentative components  $AC_i$  and  $AC_j$ , we construct  $\Omega_{ij}$  based on these three relations.

Consider the example in Figure 5.26. Suppose the sample comprises three correct annotations, represented by the three argumentation graphs 5.26a–5.26c. The first two annotators have identified an argumentative component that the third one has not ( $AC_3$ ). Similarly, the third one has identified an argumentative component that the other two have not ( $AC_4$ ). Pooling them together leads to four argumentative components ( $\mathcal{AC}_\tau = \{AC_1, AC_2, AC_3, AC_4\}$ ). The construction of the subspaces  $\Omega_{ij}$  proceeds similarly: The sample contains only three possibilities of how  $AC_2$  and  $AC_3$  are related to each other: By a support relation from  $AC_2$  to  $AC_3$  (5.26a), by an attack relation between  $AC_2$  and  $AC_3$  (5.26b) and by a lack of a relation between both nodes (5.26c). Accordingly, we construct  $\Omega_{23}$  to have three outcomes ( $\Omega_{23} = \{AC_2 \rightarrow AC_3, AC_2 \rightsquigarrow AC_3, AC_2 \leftrightarrow AC_3\}$ ).

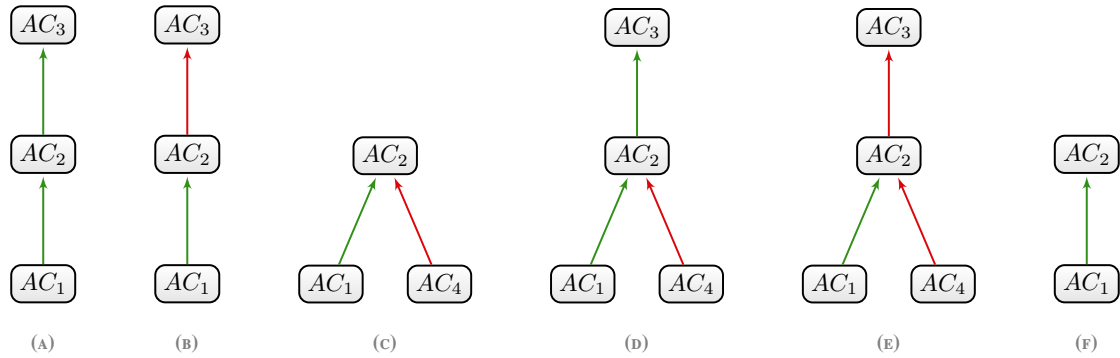


ABBILDUNG 5.26

Different argumentation graphs to illustrate the construction of  $\hat{\Omega}_\tau$ . The first three argumentation graphs represent the sample  $\tau$ .

Some of the subspaces  $\Omega_{ij}$  have even only one outcome. In every argumentation graph of the sample,  $AC_1$  and  $AC_2$  are connected by a support relation. Similarly, there is no justificatory relation between  $AC_1$  and  $AC_3$  in all maps. Hence,  $\hat{\Omega}_{12} = \{AC_1 \rightarrow AC_2\}$  and  $\hat{\Omega}_{13} = \{AC_1 \leftrightarrow AC_3\}$ . As a consequence, these subspaces do not increase the number of outcomes in  $\hat{\Omega}_\tau$  as compared to not including them at all. Only the subspaces  $\hat{\Omega}_{23}$  and  $\hat{\Omega}_{24}$  have more than one element.  $\Omega_{23}$  has three elements ( $\{AC_2 \rightarrow AC_3, AC_2 \rightsquigarrow AC_3, AC_2 \leftrightarrow AC_3\}$ ) and  $\Omega_{24}$  two elements ( $\{AC_2 \rightsquigarrow AC_4, AC_2 \leftrightarrow AC_4\}$ ). So instead of having  $16^6$  outcomes, we have only  $3 * 2 = 6$  outcomes (5.26a–5.26f).

This suggestion has the advantage of reducing the combinatorial complexity but comes at the cost of possibly missing correct justificatory relations between the given nodes in  $\tau$ . What is more, the corresponding uncertainty cannot be further qualified. We only know that the chance of missing elements of  $\Omega_\tau$  decreases with the size of the sample  $\tau$ . As a rule of thumb, we should therefore try to use as many annotators as possible—a recommendation, which is in line with estimating the probability function itself.

Let us now assess whether the suggested construction of  $\hat{\Omega}_\tau$  from the sample  $\tau$  satisfies the independence assumptions we want to exploit. That is, we ask whether  $P(e|e^*) = P(e)$  for all  $e \in \Omega_{ij}$  and  $e^* \in \Omega_{mn}$  with  $i \neq m$  or  $j \neq n$ .

To answer this question, it is useful to uncover a more specific independence assumption we used in the construction of  $\hat{\Omega}_\tau$ . The basic idea is to construct correct interpretations by slicing the given argumentation graphs from the sample  $\tau$  into their constituent smaller components and to rearrange these pieces. By using this procedure the resulting outcome space might contain more elements than  $\tau$  itself. We assume that this slicing and piecing-together procedure generates correct argumentation graphs from a given sample of correct argumentation graphs. In the example of Figure 5.26, it seems plausible that the resulting additional argumentation graphs are correct interpretations as well: For instance, there are four argumentative components in the sample, but none of the argumentation graphs comprises all four elements. But if the attacking edge from  $AC_4$  to  $AC_2$  is part of a correct interpretation (5.26c), a hypothetical fourth annotator might have come to an interpretation that includes all four argumentative components (5.26d or 5.26e) without doing anything wrong. Hence, the underlying assumption is that the correctness of a justificatory relation between one pair of argumentative components is independent of the interpretive decisions regarding justificatory relations between other pairs.

This general assumption does, however, not hold in CAAS, which implies a violation of statistical independence. There are two relevant exceptions: Sink-source ambiguity and different choices regarding the granularisation of argumentation.

Sink-source ambiguity prevails if the source or the target of a justificatory relation is ambiguously formulated (see 4.1.3). Consider the following abstract example: An author formulates an objection  $AC_1$  but is ambiguous to what exactly the corresponding text segment objects. Suppose further that the text is quite clear that the target of the objection can only be one of two different argumentative components  $AC_2$  and  $AC_3$ . In this case, the annotator has to decide on one interpretation. They cannot add both possible attack relations to their annotation (see A.1.6). Either  $AC_1 \rightsquigarrow AC_2$  or  $AC_1 \rightsquigarrow AC_3$ . In other words, the correctness of annotating an attack relation between  $AC_1$  and  $AC_2$  depends on the decision to annotate an attack relation between  $AC_1$  and  $AC_3$ . Since these different choices exclude each other,  $P(AC_1 \rightsquigarrow AC_2 | AC_1 \rightsquigarrow AC_3) = 0$  even though  $P(AC_1 \rightsquigarrow AC_2)$  is surely greater than zero.

The general strategy to deal with this problem is twofold. First, we have to decide whether two or more justificatory relations that differ either in their sinks or their sources exclude each other. As before, we will use the sample to decide these questions. Consider the argumentation graphs in Figure 5.27. Suppose the sample includes the first two argumentation graphs. In this case, every argumentation graph in the sample has either a support relation between  $AC_1$  and  $AC_2$  or between  $AC_1$  and  $AC_3$ . Accordingly, we interpret this sample as evidence for a target ambiguity that excludes the possibility to connect  $AC_1$  to both other components simultaneously. If, on the other hand, the sample contains the third argumentation graph (5.27c) as well, it includes a case in which both support relations occur simultaneously. If this argumentation graph is correct, it exemplifies how both relations do not exclude each other.

Second, we will rewrite the outcome space to include the observed dependencies in a way that the resulting marginal probabilities do not violate statistical independence. In the simple example: If  $\tau$  contains only the first two argumentation graphs, the subspaces  $\Omega_{12}$  and  $\Omega_{13}$  are not independent of each other. Knowing the precise nature of their dependence,

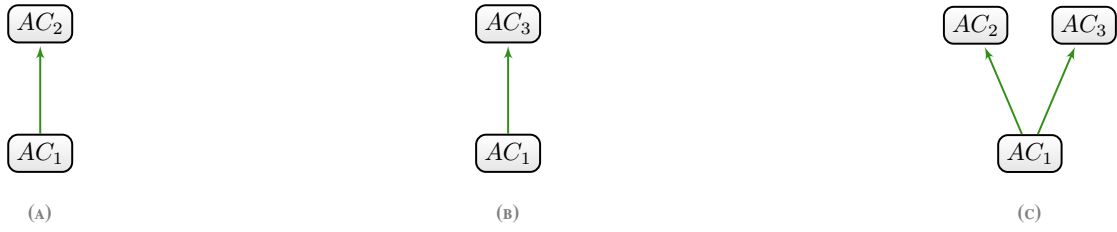


ABBILDUNG 5.27

Different argumentation graphs to illustrate the construction of  $\hat{\Omega}_\tau$  in case of sink-source ambiguity.

we can rewrite the aggregated outcome space as  $\Omega_{123} = \{AC_1 \rightarrow AC_2, AC_1 \rightarrow AC_3\}$ .

Different choices as to the granularisation of argumentation lead also to a violation of statistical independence. Consider the situation in Figure 5.28. Here, the second annotator chose a different granularisation (two argumentative units  $AU_3$  and  $AU_4$ ) of the supporting reasoning than the first one (one argumentative unit  $AU_1$ ). The alignment procedure will not align argumentative units with each other if they correspond to different granularisations. Since annotators are asked to decide on one granularisation (see Section A.1.5), the resulting argumentative components exclude each other. But then the corresponding subspaces are not statistically independent. For instance, in the given example  $P(AC_1 \rightarrow AC_2 | AC_3 \rightarrow AC_2) = 0$  since  $AC_1 \rightarrow AC_2$  and  $AC_3 \rightarrow AC_2$  exclude each other. But surely,  $P(AC_1 \rightarrow AC_2) \neq 0$ .

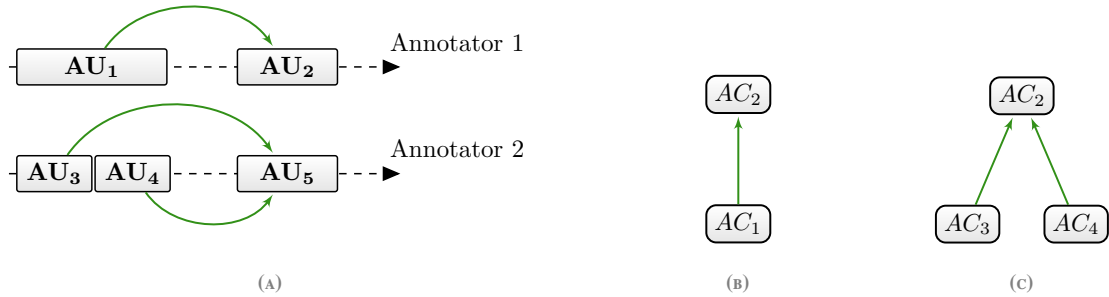


ABBILDUNG 5.28

Case of diverging granularisation.  $AU_2$  and  $AU_5$  will be mapped to the same argumentative component ( $AC_2$ ) by the alignment procedure. Argumentative units that correspond to different granularisations will not be aligned with each other.

Again, a simple suggestion to circumvent this problem is to rewrite the outcome space in a way that takes account of these dependencies. As the example suggested, we know which events exclude each other in the context of diverging granularisations. We can simply enumerate all possibilities of compounding different justificatory relations that exclude each other and define a corresponding subspace for these complex events. In the example above we can describe the situation in the following way: Either  $AC_1 \rightarrow AC_2$ ,  $AC_3 \leftrightarrow AC_2$  and  $AC_4 \leftrightarrow AC_2$ , or  $AC_1 \leftrightarrow AC_2$ ,  $AC_3 \rightarrow AC_2$  and  $AC_4 \rightarrow AC_2$ . In other words, we construct a subspace  $\hat{\Omega}_i = \{(AC_1 \rightarrow AC_2 \ \& \ AC_3 \leftrightarrow AC_2 \ \& \ AC_4 \leftrightarrow AC_2), (AC_1 \leftrightarrow AC_2 \ \& \ AC_3 \rightarrow AC_2 \ \& \ AC_4 \rightarrow AC_2)\}$  that incorporates the dependencies resulting from the different granularisations.

In the following, will assume that there are no other statistical dependencies we have to consider. Based on these assumptions and the suggested construction of  $\hat{\Omega}_\tau$ , we can now

formulate the outcome space as the Cartesian product of  $n_\tau$  subspaces:

$$\hat{\Omega}_\tau = \Omega_1 \times \cdots \times \Omega_{n_\tau} \quad (5.24)$$

Every subspace  $\Omega_i = \{a_{i1}, \dots, a_{in_i}\}$  is either a set of observed justificatory relations between a pair of argumentative components, or, as described with the last examples, a set of compound events. The number of outcomes can vary between the different subspaces (i.e., not necessarily  $n_i = n_j$ ).

Now let  $(e_1, \dots, e_{n_\tau})$  an outcome  $\tau \in \hat{\Omega}_\tau$  with  $e_i \in \Omega_i$ . Since the subspaces are statistically independent to each other, the probability of  $\tau$  can be written as a product of marginal probabilities:

$$P(\tau) = P(e_1) \times \cdots \times P(e_{n_\tau}) \quad (5.25)$$

Finally, let  $\vec{\pi}_i = (\pi_{i1}, \dots, \pi_{in_i})$  probabilities for an outcome space with  $n_i$  elements (i.e.,  $\sum_{j=1}^{n_i} \pi_{ij} = 1$  and  $0 \leq \pi_{ij} \leq 1$ ) and  $\pi = (\vec{\pi}_1, \dots, \vec{\pi}_{n_\tau})$  a set of these probability sets. By setting  $P(a_{ij}) = \pi_{ij}$ , we can finally identify the statistical model with the parameter space  $\Pi$  of all  $\pi$  that satisfy these conditions.

### 5.3.2 SIGNIFICANCE TESTING WITH CAAS

In Chapter 3, I showed how the statistical method of significance testing is used in natural sciences to decide whether observed differences are significant enough to justify the conclusion that there is a difference in the phenomena. I suggested that the same statistical framework can be used in content analysis to meet the challenge of hermeneutical under-determination. The main question of this section is whether this claim can be substantiated in the context of CAAS by using the suggested statistical model.

Let's, first, recapitulate the basic idea of significance testing.<sup>269</sup> A significance test aims to decide whether a hypothesis—the so-called null hypothesis  $h_0$ —is to be rejected given the observed data. The null hypothesis corresponds to the statement that observed differences in the data are sufficiently explainable by chance and have nothing to do with differences in the phenomena. In CAAS, the null hypothesis corresponds to the claim that observed differences between the annotations of different texts can be sufficiently explained by interpretational differences. Accordingly, significance testing demands a probabilistic conceptualisation of the target system: An outcome space  $\Omega$  that describes all possible outcomes  $\omega \in \Omega$  of a chance set-up and a statistical model that summarises all relevant probabilistic assumptions.<sup>270</sup>

<sup>269</sup>I will confine the discussion to *pure significance testing*, which can be considered a subset of the more general method of hypothesis testing. The former is confined to the consideration of type I errors, while the latter includes type II errors as well. For more details on pure significance testing, see Cox, Hinkley, and Hinkley (1974).

<sup>270</sup>For simplicity of exposition, I will assume the outcome space  $\Omega$  to be finite. As elaborated above, this assumption is satisfied in CAAS.

The statistical model corresponds to the set of all probability functions that satisfy the probabilistic assumptions. Often it is possible to describe the statistical model by a parametrised probability function. In this way, the statistical model can be equated with a parameter space  $\Theta$ —that is, a set of parameters together with a specification of their permissible value ranges. For instance, the normal distribution is defined by two parameters: The expectation value  $\mu$  and the variance  $\sigma^2$  (with  $\mu, \sigma^2 \in \mathbb{R}$ ). A specific probability function from the statistical model is then represented by one specific set of parameter values  $\theta \in \Theta$ . A *simple null hypothesis* corresponds to the hypothesis that the target system can be described by one specific probability function and, hence, by one specific set of parameter values  $\theta_0 \in \Theta$  ( $h_0 : \theta = \theta_0$ ).

Significance tests are based on drawing a finite sample  $\omega^{n_s} = (\omega_1, \dots, \omega_{n_s})$  by independently generating outcomes  $\omega \in \Omega$  with the chance mechanism. Accordingly, the sample space can be described by the Cartesian product  $\Omega^{n_s} = \Omega \times \dots \times \Omega$  ( $n_s$ -times). Since the different outcomes of the sample are generated independently, the probability of generating a specific sample can be calculated if we know or hypothesise about the probability function  $P$  that governs the chance mechanism:  $P(\omega^{n_s}) = P(\omega_1) \times \dots \times P(\omega_{n_s})$ . In particular, we can calculate sample probabilities under a given simple null hypothesis  $h_0 : \theta = \theta_0$ , which will be denoted by  $P(\omega^{n_s}; \theta_0)$ .

A significance test uses the observed data  $\omega^{n_s}$  together with the probability function  $P(\cdot; \theta_0)$  to decide whether  $h_0$  is to be rejected. Technically, we have to determine a subspace  $\mathcal{C} \subset \Omega^{n_s}$ —the so-called rejection region, or critical region—that includes all events that lead to a rejection of  $h_0$ . But how do we determine  $\mathcal{C}$ ? The general idea is to determine how “extreme” the observed sample is under the null hypothesis. More specifically, if the probability of observing as extreme events as the observed sample  $\omega^{n_s}$  under  $h_0$  is below a certain threshold, we regard  $h_0$  as rejected.

A specification of this general idea demands, first, to specify the set of events  $\Omega^e(\omega^{n_s}) \subset \Omega^{n_s}$  that are considered at least as extreme as the sample  $\omega^{n_s}$ . The probability to observe events that are at least as extreme as the sample (under  $h_0$ ) is called the  $p$ -value of the sample:

$$p(\omega^{n_s}; \theta_0) := P(\Omega^e(\omega^{n_s}); \theta_0) \quad (5.26)$$

Second, we need a specification of a threshold  $\alpha$ —the significance level—that determines which observed  $p$ -values justify a rejection of  $h_0$ . A typical value for  $\alpha$  is 0.05, which can be justified by its relation to the probability of erroneously rejecting the null hypothesis.

The *rejection rule* of significance testing can now be formulated in the following way:<sup>271</sup>

$$\text{If } p(\omega^{n_s}; \theta_0) \leq \alpha, \text{ then } h_0 \text{ is considered as rejected.} \quad (5.27)$$

It remains to specify  $\Omega^e(\omega^{n_s})$  for all  $\omega^{n_s} \in \Omega^{n_s}$ . As the terminology suggests, the idea is to rank the outcomes of the sample space according to some criterion. This criterion is usually formulated with the help of a suitable function that maps outcomes to numbers

<sup>271</sup>Under this construction the critical region  $\mathcal{C}$  is the set  $\{\omega^{n_s} \in \Omega^{n_s} | p(\omega^{n_s}; \theta_0) \leq \alpha\}$ .

such that natural ordering of numbers corresponds to the desired ranking of the outcomes. More specifically, this so-called test statistic  $t : \Omega^{n_s} \rightarrow \mathbb{R}$  is typically defined such that:

$$\Omega^e(\omega^{n_s}) = \{\nu^{n_s} \in \Omega^{n_s} \mid |t(\nu^{n_s})| \geq |t(\omega^{n_s})|\} \quad (5.28)$$

This construction is very neat from a mathematical point of view since a probability function  $P$  on  $\Omega^{n_s}$  determines a corresponding probability function  $P_t$  on the image of the test statistic. Accordingly, the significance test can be performed by considering  $p$ -values of the form  $p(t(\omega^{n_s}); \theta_0)$  that are calculated by using  $P_t(\cdot; \theta_0)$ .

The definition of the test statistic depends on the specific context and, in particular, on the given statistical model. To take an illustrating example: Suppose we know the variance  $\sigma^2$  of a normally distributed quantity but do not know its expectation value  $\mu$ . Now we want to test whether  $\mu$  equals a specific value  $\mu_0$  ( $h_0 : \mu = \mu_0$ ). In this case, it seems reasonable to compare sample means  $\bar{X}$  of this quantity with the hypothesised population mean  $\mu$ . Every outcome of the sample space that results in a larger deviation from the hypothesised mean than the observed sample mean is considered more extreme than the observed sample under the assumption  $h_0$ . The Z statistic  $Z := \frac{\bar{X} - \mu_0}{s}$  (with  $s = \sigma / \sqrt{n}$ ) provides the desired ordering of outcomes in the sample space and is distributed by the standard normal distribution.

How do we apply this general framework in the context of CAAS? If we simply want to test whether a specific hypothesis about specific parameter values ( $h_0 : \pi = \pi_0$ ) is considered to be rejected by an observed sample  $\tau$  of annotations, we can proceed as suggested: Let  $\Omega_\tau^{n_s} = \Omega_\tau \times \dots \times \Omega_\tau$  ( $n_s$ -times) be the sample space of producing  $n_s$  annotations by randomly drawing  $n_s$  annotators from the population of properly instructed annotators who annotate a text independently from each other. We might now define  $\Omega_\tau^{extr}$  by

$$\Omega_\tau^e(\tau) := \{\omega^{n_s} \in \Omega_\tau^{n_s} \mid P(\omega^{n_s}; \pi_0) \leq P(\tau; \pi_0)\} \quad (5.29)$$

and calculate the corresponding  $p$ -values by using their definition (5.26).

However, a specific hypothesis about the parameter values is usually not that interesting in itself. Rather, we might want to know something about the number of argumentative components in a text or its argumentative complexity. More generally, we are interested in argumentative quantities that are defined by attributing a specific value to an argumentation graph. As explained above we have to resort to expectation values of these quantities since there is not necessarily one value in the case of hermeneutical underdetermination.

More formally: Let  $X$  be a function that assigns each  $\tau \in \Omega_\tau$  a specific number (e.g.,  $X : \Omega_\tau \rightarrow \mathbb{R}$ ). The expectation value, or the probability-weighted mean of  $X$ , is defined by:

$$E(X) := \sum_{\tau \in \Omega_\tau} P(\tau) X(\tau) \quad (5.30)$$

Hence, typical null hypotheses of interest will have the form:  $h_0 : E(X) = v_0$ . For instance, let  $N$  be the number of nodes in an argumentation graph. Then we might want to know whether a specific hypothesis about the expected number of argumentative components

$N_0$  is rejected by an observed sample—that is, we test  $h_0 : E(N) = N_0$ . Or, we might want to know whether the expected number of argumentative components differs between two texts. In this case, the null hypothesis might claim that there is no difference between the expected number of argumentative components of the first text  $E_1(N)$  and of the second text  $E_2(N)$  ( $h_0 : E_1(N) - E_2(N) = 0$ ).

At first glance, there seems to be no relevant difference to the example of the normal distribution from above. Let's suppose we want to test  $h_0 : E(N) = N_0$ . Accordingly, we might want to use a similar test statistic by measuring the difference between the sample mean  $\bar{N}$  and the expected number under  $h_0$ :

$$t(\tau) := |\bar{N} - N_0| = \left| \frac{1}{n_s} \sum_{\tau \in \tau} N(\tau) - N_0 \right| \quad (5.31)$$

Given such a statistic, we must then calculate the corresponding  $p$ -values to apply the rejection rule 5.27 by calculating the probability to observe larger or equal differences from  $N_0$  than the observed sample mean:

$$p(\tau; h_0) = P_t(t \geq t(\tau); \pi_0) \quad (5.32)$$

The problem in CAAS is that hypotheses of the form  $E(X) = v_0$  do not necessarily correspond to one specific probability function, but many. Depending on  $\Omega_\tau$ , there will be different  $\pi \in \Pi$  that will lead to  $E(N) = N_0$ . In other words, in contrast to a simple null hypothesis, which corresponds to *one* complete set of parameter values  $\pi \in \Pi$ , such a *complex null hypothesis* is consistent with different parameter value sets  $\pi$ .

The method of significance testing on which I have elaborated so far is not able to deal with such complex null hypotheses. The  $p$ -values are not properly defined because there is not one specific probability function  $P_t(\cdot; \pi_0)$  but many. Accordingly, there is not one particular  $p$ -value of the observed data, but as many  $p$ -values as there are probability functions consistent with  $h_0$ .

On what  $p$ -value should we base the inference rule of significance testing (5.27)? There are at least two different suggestions: First, we might pick one specific  $p$ -value. In particular, there are good reasons to base the inference on the maximal  $p$ -value. Second, we might aggregate all  $p$ -values to one overall  $p$ -value, as, for instance, the mean  $p$ -value.

Both suggestions can be justified by adapting the common justification of choosing a specific significance level, which is based on quantifying and controlling type I error probabilities. A type I error occurs if the researcher rejects the null hypothesis even though it is true. In the case of a simple hypothesis, it is easy to show that the chosen significance level  $\alpha$  is an upper bound for the probability of erroneously rejecting the null hypothesis (see A.2.6). The choice of  $\alpha$  will therefore depend on the willingness to accept a certain chance of type I errors.

According to the first suggestion, we will base the rejection rule on the maximum  $p$ -value among all  $p$ -values that are based on parameters  $\pi$  that are consistent with  $h_0$ . This suggestion can be given a precise form in the following way: Let  $h_0$  be a complex



hypothesis that is represented by a subset  $\Pi_0$  of the parameter space  $\Pi$ . If we define the  $p$ -value for a complex hypothesis as

$$p_{\Pi_0}(\omega^{n_s}) := \sup\{p(\omega^{n_s}; \pi) | \pi \in \Pi_0\} \quad (5.33)$$

the reformulated rejection rule for complex hypotheses becomes:

$$\text{If } p_{\Pi_0}(\omega^{n_s}) \leq \alpha, \text{ then } h_0 \text{ is considered as rejected.} \quad (5.34)$$

In other words, we demand that the  $p$ -value is smaller than the significance level for all parameter value sets  $\pi \in \Pi_0$ . Accordingly, the probability of type I errors will be guaranteed to be smaller than  $\alpha$  because it is smaller for all specific probability functions that are consistent with the null hypothesis.<sup>272</sup>

If the suggested generalisation of  $p$ -values to complex hypothesis guarantees that  $\alpha$  is an upper bound for type-I-error probabilities, why should we even consider an alternative to this suggestion? The worry is that we might diminish our chance to detect significant differences. Considering type-I-error probabilities cannot be the only relevant criterion. Otherwise, we could simply set  $\alpha = 0$ . Since  $p$ -values are probabilities they will never be smaller than 0. Hence, setting  $\alpha = 0$  would never lead to a rejection of  $h_0$ . Hence, the price for excluding type I errors completely is that false null hypotheses will not be rejected as well. There is a trade-off between the acceptance of type-I-errors probabilities and the chance to reject false null hypotheses.<sup>273</sup>

This consideration is relevant for the suggested generalisation of  $p$ -values in the following way: The generalised Rejection Rule 5.34 results only in a rejection if  $p(\omega; \pi)$  is below  $\alpha$  for *every*  $\pi \in \Pi_0$ . Hence, a specific  $\omega$  might not lead to a rejection, even if it would lead to a rejection when using the rejection rule for simple hypotheses  $\pi \in \Pi_0$ . In other words, there might be fewer events that lead to an actual rejection of  $\Pi_0$  than the events that would lead to a rejection of a simple hypothesis  $\pi \in \Pi_0$ .<sup>274</sup>

The worry is therefore that the Rejection Rule 5.34 is too conservative in rejecting  $h_0$ . An alternative might be to base the rejection of complex hypotheses on the mean  $p$ -value instead of the maximum  $p$ -value. Let the mean  $p$ -value  $\bar{p}$  be defined by:

$$\bar{p}_{\Pi_0}(\omega^{n_s}) := \frac{1}{|\Pi_0|} \int_{\Pi_0} p(\omega^{n_s}; \pi) d\pi \quad (5.35)$$

Using the mean  $p$ -value might in some cases  $\pi \in \Pi_0$  result in type-I-error probabilities greater than  $\alpha$  but will guarantee that the mean type-I-error probability will be below  $\alpha$ .

<sup>272</sup>See A.2.10 for a more rigorous proof.

<sup>273</sup>In the simple framework of pure significance testing, this trade-off cannot be further quantified. However, the more general framework of hypothesis testing includes a conceptualisation of type II errors—that is erroneously accepting an alternative of  $h_0$ —and their probabilities.

<sup>274</sup>Or to put it in terms of critical regions: The size of the critical region of rejecting a complex hypothesis  $\Pi_0$  is a lower bound for all critical regions of rejecting simple hypotheses  $\pi \in \Pi_0$ . See, A.2.9 for details.

We do not have to decide here, which of the two suggestions is more appropriate or if other statistical accounts are even more suitable for CAAS. Rather, I provided a mere conceptual exploration of using statistical methods in the context of CAAS. The resulting sketch established that statistical methods are in principle possible to deal with hermeneutical underdetermination while being able to reliably detect differences in phenomena.

### 5.3.3 CONSIDERATIONS ABOUT COMPUTATIONAL COMPLEXITY

Let me end this chapter with a few remarks on the practical problems that are connected to using the sketched suggestion to calculate  $p$ -values.

As described above, the social scientist will generally not be interested in determining whether an observed sample is statistically consistent with a specific set of parameter values  $\pi$ , but whether it is statistically consistent with the expectation value of some property of the argumentation structure. For instance, they might want to know whether an observed mean number of argumentative components  $\bar{N}$  rejects the hypothesis that the expected number of argumentative components equals some specific value  $N_0$  (i.e.,  $h_0 : E(N) = N_0$ ).

A reasonable test statistic  $t$  is in this case the distance between the expectation value  $N_0$  under  $h_0$  and the observed mean  $\bar{N}$ :

$$t(\tau) := |\bar{N} - N_0| \quad (5.36)$$

The  $p$ -values of interest correspond in this case to the probability to observe at least as large a deviation from  $N_0$  as the observed deviation  $t(\tau)$ :

$$p(\tau; \pi) = P_t(t \geq t(\tau); \pi) \quad (5.37)$$

Consequently, we have to determine the probability functions  $P_t(\cdot; \pi)$  ( $\pi \in \Pi_p$ ) and then either find the maximum or the mean  $p$ -value. There are several practical challenges.

First, there is usually no easy analytical solution to deduce  $P_t(\cdot; \pi)$  from the distribution  $P(\cdot; \pi)$  on the sample space. The problem is that the mapping  $t$  from  $\Omega^{n_s}$  to  $\mathbb{N}$  depends crucially on  $\Omega$ . To put it simply: The number of argumentative components depends on the specific argumentation graph at hand. In particular,  $N$  is usually not bijective: Different argumentation graphs will have the same number of argumentative components. We might apply a brute force strategy to determine  $P_t$ . Since  $\Omega^{n_s}$  is finite, we can simply iterate over the whole sample space and determine  $P_t(\cdot; \pi)$  by using

$$P_t(N; \pi) = \sum_{v \in \{\tau \in \Omega^{n_s} \mid t(\tau) = N\}} P(v; \pi) \quad (5.38)$$

If, however, the sample space is large, this suggestion might be computationally unfeasible. In this case, we might estimate  $P_t(\cdot; \pi)$  by, first, using  $P(\cdot; \pi)$  to generate a sample of samples and then use the resulting empirical distribution of  $N$  as an estimation of  $P_t(\cdot; \pi)$ .

Either way, we have to identify those parameter sets  $\pi$  that are consistent with the null hypothesis  $\Pi_0$ . This can even be done analytically. A null hypothesis of the form  $E(X) = X_0$

with  $X : \Omega^{n_s} \rightarrow \mathbb{R}$  can be regarded as one additional constraint on the different parameters  $\pi_{ij}$ . Accordingly, we can rewrite the null hypothesis to eliminate one of the free parameters  $\pi_{ij}$ .<sup>275</sup>

Now we have to determine the maximum  $p$ -value  $p(\tau; \pi)$  over  $\Pi_0$ , or the mean  $p$ -value respectively. The challenge is that the parameter space  $\Pi_0$  is continuous and can, presumably, not be assessed analytically. Finding a maximum is an optimisation problem that might be approached by, for instance, a random walk on  $\Pi_0$ . That in turn demands generating random  $\pi \in \Pi_0$  accordingly. Determining the mean  $p$ -value is similarly difficult. Instead of analytically determining  $\bar{p}$  by solving the Integral 5.35, we might estimate the integral by, for instance, generating a random sample  $\pi^m$  of  $m$  parameters  $\pi \in \Pi_0$  and use the mean as an estimate of  $\bar{p}$

$$\hat{p}_{\Pi_0}(t(\tau)) = \frac{1}{m} \sum_{\pi \in \pi^m} p(t(\tau); \pi) \quad (5.39)$$

With both strategies, we will not only have to determine  $\pi$  consistent with  $\Pi_0$ , but have to generate these parameter sets randomly.

These considerations show that the calculation of  $p$ -values relies on numerical methods that might be computationally demanding.

## 5.4 SUMMARY

In this chapter, I have tied together the different strands of the other chapters. In particular, I showed that *content analysis of argumentation structures* (CASS) can satisfy the requirement of reliability, which I elaborated in Chapter 3, even though it is hermeneutically underdetermined (4).

To that end, I introduced the simple disagreement measure  $\Delta_\tau$  (5.1) that measures the interpretational distance (for those indeterminacies I defined in Section 4.1) between annotation results and which can be used to assess the overall degree of interpretation. The measure  $\Delta_\tau$  is a simple graph edit distance but relies on an inter-annotator identification of argumentative units (i.e., an alignment). I adopted the  $\gamma$  approach of Mathet, Widlöcher, and Métivier (2015), which is, without alterations, not able to properly (not) align argumentative units that correspond to diverging choices as to the granularisation of argumentation.

The introduction of  $\Delta_\tau$  is but the first step of instantiating the described general recipe to deal with hermeneutical underdetermination within a reliability-orientated paradigm (3.4.3). It also needs a statistical model for CAAS and statistical inference patterns to decide whether observed differences in annotation results can be interpreted as differences in the phenomena.

It turned out that combinatorial and computational complexity are the main challenges to devising a statistical model and using statistical inference patterns. I suggested estimating the sample space based on a sample of annotation results, which increased the size of the

<sup>275</sup>See Appendix A.2.2.

sample space to a manageable extent (5.3.1). Following the used analogy from the error-based approach in natural sciences (3.4.2), I fathomed the prospects of using significance testing as the statistical inference pattern in CAAS (5.3.2). In contrast to pure significance testing, CAAS researchers will paradigmatically be interested in complex null hypotheses. While this hurdle can be cleared in principle, the accompanying calculations of  $p$ -values might be numerically infeasible (5.3.3). Therefore, further research is needed to enable the actual application of statistical inference patterns in CAAS.

## 6. CONCLUSION

This work introduced *content analysis of argumentation structures* (CAAS) as a reliability-orientated and value-neutral content analysis, which can be employed in different socio-empirical contexts to analyse argumentation structure. I also proposed an annotation scheme for CAAS, which can be used immediately or as a starting point for a more extensive category system to analyse argumentation structure.

The main aim of this work was to provide a framework that meets the challenge of hermeneutical underdetermination in the context of analysing argumentation. Based on theoretical considerations and different examples, I argued that the analysis of argumentation structure will lead, in some cases, to different interpretations even if we ground the annotation scheme on insights from argumentation theory. In other words, the results of annotators can differ.

This hermeneutical underdetermination is a challenge for CAAS since it threatens the reliability of the annotation. Reliability measures the extent to which the content analyst can trust the results of the annotation process. This trustworthiness view of reliability is conceptualised as a form of reproducibility (or intersubjectivity) and demands that annotations should not depend on the particular annotator. If annotation results depend on possibly diverging interpretational choices of annotators and, thus, may vary between annotations of the same material, it is difficult to ascertain whether observed differences are the results of mere interpretational differences or are explainable by differences in the phenomena. In other words, violating the reliability requirement imposes certain difficulties on the researcher to draw implications from the data concerning their research questions.

Hitherto, content analysis employed two strategies to address the challenge of hermeneutical underdetermination. The *received view of (reliability-orientated) content analysis* (also called *quantitative content analysis*) introduces chance-corrected reliability measures, which provide a graded conceptualisation of reliability and measure the degree to which annotators perform better than chance. If the corresponding reliability values of an annotation process exceed recommended thresholds, the data-making process is considered sufficiently reliable. Otherwise, the annotation scheme should be improved until sufficiently high reliabilities are reached. In other words, the strategy of the received view is to minimise the degree of interpretation by devising a sufficiently precise annotation scheme.

The problem is that a research question might impose certain boundaries on the prospects of improving the annotation scheme in this way. Depending on the research question, it may not be possible to narrow down the interpretation leeway sufficiently (without

compromising the validity of the data-making process). *Qualitative content analysis* is well aware of such cases. In the face of such an irreducible degree of interpretation, qualitative content analysis suggests employing alternative requirements, such as transparency of the annotation process and consensus coding to solve conflicts between diverging interpretations.

In sum, content analysis suggests minimising interpretational freedom to the point that reliability is satisfied or giving up on reliability if the former strategy is unsuccessful. In other words, hitherto, content analysis could not combine irreducible degrees of interpretation with fidelity to reliability. The main contribution of this work to content analysis, and CAAS in particular, is to provide an alternative framework of a reliability-orientated content analysis that allows an irreducible degree of interpretation.

My suggestions in this work draw on an analogy of measurements in natural sciences. While measurement instruments do not arrive at diverging interpretational choices, measurement results vary to a certain extent—even under conditions of repeatability. This variability of measurement results is often explained in terms of random measurement errors and conceptualised in a probabilistic way. Similar to the challenge of hermeneutical underdetermination, the question is under which conditions differences in measurement results can be taken to imply differences in the observed phenomena.

The probabilistic conceptualisation helps in this connection since it allows the use of statistical methods to answer this question. The paradigm of significance testing introduces the notion of significance that lets the researcher decide whether differences in measurement results cannot be explained by random errors alone but can be taken to indicate a difference in the phenomena.

While there are several differences between measurements in natural sciences and the annotation of argumentation structure, I argued that a probabilistic conceptualisation of reliability, similar to the one of measurements, is able to meet the challenge of hermeneutical underdetermination in CAAS. An elaboration of this suggestion requires a probabilistic conceptualisation of the annotation process together with a specification of a corresponding statistical model and apt statistical inference patterns that specify conditions under which inferences from differences in annotations to differences in the phenomena are (statistically) valid.

To that end, I suggested understanding the annotation process as a random process. According to this view, one annotation results from randomly picking an annotator from the population of trained coders and letting them annotate a text. In this way, the space of all interpretations is supplemented with a probability function, which assigns a probability value to each argumentation graph. In other words, the data-making process for a specific text is modelled by a probability function over the outcome space of coding results. I further suggested estimating the outcome space and the corresponding distribution based on a sample of annotations and elaborated on how significance testing can be employed to license statistical inferences from observed differences in annotation results to differences in the phenomena.

I have thereby shown that an irreducible degree of interpretation is compatible with satisfying the reliability requirement. The proposed framework and the suggested minimal

annotation scheme can be applied in different research contexts (and every other context in which reliability is of concern) to analyse argumentation structure.

The resulting argumentation graphs can be used to gauge different argumentative quantities. One paradigmatic example is undoubtedly the amount of reasons in a debate. Additionally, the researcher can analyse argumentative complexity, bias (or balance), argumentative depth, argumentative density and every other measure based on argumentation graphs.

Take, for instance, balance (or bias) within a debate analysed as an argumentation graph  $\tau$ . With respect to a central claim  $C$ , represented by one specific node in  $\tau$ , balance can be understood as the ratio of pro and con reasons with respect to  $C$ . Every reason node in  $\tau$  can be categorised as neutral, pro or con. Without giving a precise mathematical definition, the basic idea is this: If there is no directed path from a reason  $R$  to  $C$ ,  $R$  is considered neutral concerning  $C$ . If such a path exists, the amount of attacking edges determines whether  $R$  is a con or a pro reason. If  $R$  directly attacks  $C$  ( $R \rightsquigarrow C$ ),  $R$  is obviously a con reason. But the same holds if  $R$  attacks reasons that support  $C$  (for instance,  $R \rightsquigarrow A \rightarrow C$  or  $R \rightsquigarrow A \rightarrow B \rightarrow C$ ). If, on the other hand,  $R$  refutes objections by attacking con reasons against  $C$ ,  $R$  is considered a pro reason since it indirectly supports  $C$  by answering to objections (for instance,  $R \rightsquigarrow A \rightsquigarrow C$  or  $R \rightsquigarrow A \rightsquigarrow B \rightarrow C$ ). If  $R$  supports  $C$ ,  $R$  is a pro reason. Finally, if  $R$  supports another reason, it depends on whether this reason is a con or a pro reason: If  $R$  supports a con reason  $A$  (for instance,  $R \rightarrow A \rightsquigarrow C$ ),  $R$  is itself a con reason. Similarly, if  $R$  supports a pro reason  $A$  (for instance,  $R \rightarrow A \rightsquigarrow B \rightsquigarrow C$ ),  $R$  is a pro reason.

It is not only possible to compare different quantity values of these measures between different texts but also to estimate the degree of interpretation within a text or bigger corpus. Since the annotation process is conceptualised as a probabilistic process, the degree of interpretation is determined by the distribution of all interpretations. For a specific measure, the degree of interpretation can be quantified by specifying the dispersion of this measure. For instance, the degree of interpretation with respect to the number of argumentative components can be quantified by the (statistical) variance of the number of argumentative components, which must be estimated based on a sample of annotations.

I argued on conceptual grounds and by using specific examples that there can be a degree of interpretation in analysing argumentation structure. However, these results cannot be generalised to other texts and corpora. The extent of hermeneutical underdetermination is, therefore, open to empirical scrutiny.

It stands to reason that the degree of interpretation depends on the text at hand. Authors can make their argumentation structure maximally explicit by using linguistic cues or can even obfuscate their argumentation on purpose. We should, therefore, expect that the degree of interpretation varies to some extent between different texts. The question is whether the degree of interpretation depends on specific properties of argumentative text and their production. The degree of interpretation might be connected to the text genre, the educational background of arguers, their intentions and many other features.

One could, for instance, begin by equating the notion of argumentative clarity with the degree of interpretation. The more explicit and less ambiguous an argumentation is in terms of interpretational leeway, the more precise and clear it is. We could now hypothesise

that argumentative clarity depends (among other things) on argumentative competency in the following way: Argumentative clarity of arguers increases with an increase in their argumentative skills. More concretely, a researcher could investigate whether educational interventions such as critical thinking courses lead to a joint increase in argumentation skills and argumentative clarity in their essays.

Besides investigating the dependence of the degree of interpretation on other properties, the degree of interpretation might influence, for instance, the reception of argumentative text. In sum, analysing the role of interpretational leeway (as it pertains to argumentation structure) as an independent and dependent variable might be worthwhile. CAAS is, to the best of my knowledge, the first methodological framework that facilitates this form of empirical research.

Finally, CAAS can be combined and complemented with other research methods. The most straightforward suggestion to complement CAAS is to use the suggested minimal annotation scheme as a starting point to devise a more ambitious annotation scheme. The bare structural analysis of the suggested annotation scheme can be extended with topical features of argumentative units by introducing subcategories that distinguish between different types of argumentative units. The researcher could, for instance, categorise reasons according to argument schemes and assess what type of argumentation occurs and how often. It is similarly possible to introduce evaluative subcategories that assess, for instance, argument strength or rhetorical style of reasons and whole argumentations. The structural analysis can also be complemented with an analysis of dialectical aspects based on categories that track the genesis and the authors of reasons.

Let me close this conclusion by describing some limitations of CAAS and directions for further (methodological) research into CAAS.

The framework I devised is quantitative since it enables quantitative assessment of the degree of interpretational leeway; it is also structural to the extent that the minimal annotation scheme is confined to a mere structural analysis of argumentation. The quantitative and structural perspective abstracts away from the content of argumentation and the hermeneutical process of understanding this content. That may be unsatisfactory to some researchers who are interested in analysing argumentation. They might remark that numbers and structure cannot exhaustively describe hermeneutical underdetermination. Without considering the process of analysing the meaning of argumentative units, the methodological reflection on underdetermination and its assessment by applying CAAS are ignorant about some relevant aspects of hermeneutical underdetermination.

While I can sympathise with these considerations, I do not regard these points as problematic for CAAS. Admittedly, CAAS adopts a mere structural and quantitative perspective on the degree of underdetermination in its suggested form. However, CAAS is consistent with additional suggestions to conceptualise and analyse interpretational leeway.

First, the hermeneutical process of understanding argumentation represents an integral part of CAAS. Obviously, annotators have to understand argumentation to annotate its structure. Additionally, the process of understanding is described and, in some way, normalised by the annotation scheme. Finally, CAAS can be complemented with additional suggestions on how to account for these content aspects of underdetermination more explicitly. For



instance, annotators might be asked to describe and explain their interpretations. These descriptions could then be used to search for explanations of the observed interpretational leeway. I regard this kind of transparency as very important for the above-described empirical research about hermeneutical underdetermination since such descriptions might be used to formulate promising hypotheses about the source of hermeneutical underdetermination and its dependence on other text features.

More pressing issues are the identified challenges of applying CAAS. Compared to the received view of reliability-orientated content analysis, there are (at least) three hurdles to using CAAS.

First, CAAS demands the use of more than two annotators. According to the received view, two annotators are often sufficient to calculate reliability values. In CAAS, annotators are used to estimate the outcome space and the probability distribution. To that end, the annotators should be a representative sample of the population of trained annotators. It is not difficult to realise that  $n = 2$  will often be a too poor basis. The described statistical model corresponds to a set of different multinomial distributions. It is like having several different dice with unknown probabilities for their outcome. The question of how many annotators suffice to estimate these probabilities corresponds to how often you have to throw these dice to estimate their probabilities.

Second, the distribution of interpretations will often depend on the particular text at hand. In other words, the degree or variance of hermeneutical underdetermination might differ between texts. A comparison of texts in terms of their argumentation structure demands, therefore, to estimate the probability distributions for both texts separately.

Finally, the needed statistical methods for CAAS are mathematically complex and computationally demanding. While the considerations of the last chapter show how significance testing can, in principle, be used in CAAS to satisfy requirements of reliability, we also saw that the combinatorial complexity of the statistical model leads to computational challenges in calculating the relevant  $p$ -values. Further research is needed to address this issue. Small case studies should be used to better estimate whether the suggested statistical methods are feasible for the annotation of ordinary-language argumentation. Depending on the results, alternative statistical methods and tools must be devised to make the corresponding statistical inferences feasible.

Addressing these challenges represents a critical aspect of advancing CAAS. There are at least two additional directions to furthering CAAS by broadening its scope.

First, some basic aspects of comparing the argumentation structure between different texts are beyond the capabilities of the hitherto devised methods. While it is, for instance, possible to compare the number of argumentative components in different texts, it is not possible to assess the number of argumentative components in which they differ. For simplicity, let us, for the moment, disregard interpretational leeway. Suppose there are ten argumentative components in one text and fifteen in another. While the difference in the number of argumentative components is fixed by these number (five), they tell us nothing about the overlap of argumentative components between these texts. The second text could contain the ten of the first text and formulate five additional ones, or, to take the other extreme, it could contain fifteen new components—that is, not repeating any of the

components from the first text.

Such an assessment demands an inter-text (or intra-corpus) identification of argumentative components, which was not addressed in this work. The inter-annotator identification of argumentative units was confined to identifying argumentative units of different annotators in the same text based on mere syntactical properties. An inter-text identification of argumentative units, on the other hand, demands the consideration of semantical and pragmatic aspects. In this sense, it is an additional content analysis that must be based on identity criteria for argumentative components. For instance, the analyst has to formulate criteria that determine under which conditions two utterances express the same reason or the same argument.

Another direction to advance the scope of CAAS is to use reconstructive methods. I argued at length that using methods to reconstruct arguments does not *necessarily* provide any advantage in narrowing down the degree of interpretation connected to the argumentation structure. However, the discussed examples might not be representative and the provided general considerations might be irrelevant in actual debates. In other words, the reconstruction of arguments might in some context decrease the degree of interpretation. But even if the reconstruction of arguments does not decrease interpretational leeway, there are research contexts in which an annotation scheme based on argument reconstructions has advantages. First, a categorisation of argument types (e.g., based on argumentation schemes) might demand the identification of implicit premises to decide whether the set of all premises fits a particular argument type. Similarly, the evaluation of arguments, as argumentation theorists envisage it, requires the identification of all premises of arguments. Assessing arguments' rational strength demands assessing the plausibility of their premises and the inferential link between premises and conclusion. A reconstruction of arguments is necessary for these purposes since it makes the internal structure explicit. Without reconstructing arguments, the annotator might miss relevant implicit premises.

A combined and dual usage of the minimal argumentation scheme together with an additional one based on argument reconstruction could also be used to reveal a form of hermeneutical underdetermination that cannot be detected with the hitherto described methods. Suppose that a set of annotators employs the minimal annotation scheme and all agree in their categorisation of a specific text segment as an argument (including its granularisation and all justificatory relations having it as source or sink). Accordingly, there is no interpretational leeway with respect to this text segment—at least not in terms of the introduced structural types of hermeneutical underdetermination. However, other relevant forms of interpretational indeterminacies cannot be captured by mere structural aspects. If different annotators reconstruct one argument differently, there is an interpretational leeway of understanding the argument even if there is no ambiguity in terms of structural aspects.<sup>276</sup> A comparison of these different reconstructions can be used to describe and assess this type of hermeneutical underdetermination.

Finally, the reconstruction of arguments could be used for an inter-text identification of arguments. One suggestion to decide whether argumentative units in different texts express

---

<sup>276</sup>A reconstruction can, of course, also lead to differences in structural aspects even if there is agreement with respect to the annotation results of applying the minimal annotation scheme. For instance, a reconstruction might lead to a reassessment of justificatory relations.

the same argument is to base such a comparison on their reconstructions. In other words, the identity criteria of ordinary-language arguments can be linked to the identity criteria of reconstructed arguments. The basic idea is to regard two ordinary-language arguments as the same if their reconstructions are sufficiently similar.

In sum, there are several suggestions to apply CAAS in different research contexts, advance CAAS as a method by meeting the identified challenges, and combine and complement CAAS with additional methods.



## A. APPENDICES

### A.1 ANNOTATION GUIDELINES FOR A MINIMAL ANALYSIS OF ARGUMENTATION STRUCTURE

The following guidelines provide clarifications, maxims and hints that will help you analyse the argumentation structure in argumentative texts.

So, what are argumentative texts and their argumentation structures? We will say that a text is argumentative if it contains, among other things, reasons and arguments in favour of one or more claims, possibly objections to these arguments and claims, and possibly refutations of the objections. In other words, argumentative texts contain arguments and reasons, which can be connected in different ways.<sup>277</sup> The argumentation structure of a text is simply the set of these elements together with the different relations between them.

#### A.1.1 VISUALIZING ARGUMENTATION STRUCTURES WITH REASON MAPS

Let's have a look at the transcription of the following hypothetical dialogue, which is about vegetarianism:

*Clint:* I think we all should change to a vegetarian lifestyle.

*Audrey:* Why?

*Clint:* Because most mass meat-farming techniques are torture to animals.

*Audrey:* Is that so?

*Clint:* Yes! There is good evidence that animals feel fear and pain like us. Mass meat-farming techniques force those animals to have an unnaturally short, miserable and confined life.

Clint begins by formulating a claim about vegetarianism, and Audrey challenges this claim by asking for a justification. Clint answers this request by providing a reason or, in other words, an argument in favour of that claim. In the following, we will not distinguish between reasons and arguments. We will regard both concepts as more or less

---

<sup>277</sup>Essays are typically argumentative. But argumentation can also be found in other text types, such as newspaper articles, scientific articles, transcribed interviews, forum entries, text messages, tweets, blog articles, books, or what have you.

interchangeably.<sup>278</sup> What is important, though, and exemplified in the dialogue is that arguments and reasons are always relational. A reason is always a reason for something; similarly, an argument is always an argument in favour of something. Therefore, a complete description of an argument or a reason should be formulated according to the following standard form:

*x is a reason/argument for y.*

Identifying an argument or a reason demands identifying both relata *x* and *y*. In the above example, we can consider Clint's answer to Audrey's first question as a reason (*x*) for his initial claim (*y*).

Audrey reacts to Clint's answer by demanding another justification, and Clint willingly provides a further reason. But for what exactly is Clint's last answer a reason? Without the cotext of Clint's second answer—that is, all that was said besides this answer—we can easily imagine the answer to be another reason for vegetarianism. However, the way the dialogue proceeded, another interpretation is more appropriate. Audrey's second question refers to Clint's first answer. She asks whether mass meat farming is the torture of animals. Clint answers in the affirmative and provides a justification for why mass meat farming should be considered torture. In this particular dialogue context, the second answer should be, therefore, considered a reason for the first answer.

We can describe the structure of the given argumentation thus far as follows: The argumentation contains a claim for which a reason is presented, and this reason is itself backed up by an additional reason. Figure A.1 is a graphical depiction of that structure, which we will call a *reason map* since it visualizes reasons and their relationships to each other and to claims. As we will see, reason maps can get more complex than this simple example.

The nodes of a reason map represent text segments of the dialogue; the green arrows represent the relation of *x* being a presented reason for *y*, which is visualized by a green arrow from *x* to *y*. In the following, we will refer to this relation as the *support relation*.

So far, the argumentation structure of this dialogue is quite simple and not very controversial. Let's look at the following more challenging proceeding of the exchange:

*Audrey:* But even if that were true, there are alternative ways to breed animals. Grass-fed meat comes from animals that are raised in open grass pastures and that are not confined to these cramped living spaces.

*Clint:* Yes, but there are other reasons to become a vegetarian. The farming of animals produces more greenhouse gas emissions than the world's entire transport system, and these greenhouse gases are causing dangerous climate change.

Audrey begins by raising an objection. Similar to reasons, objections are relations: An objection is always an objection to something and should be described in some standard

---

<sup>278</sup>In argumentation theory, an argument is usually considered a complex entity with an internal premise-conclusion structure. The conclusion is what is supposed to be justified with the argument, and the premises are the sentences that formulate the justification. According to this understanding, reasons are premises as parts of arguments.

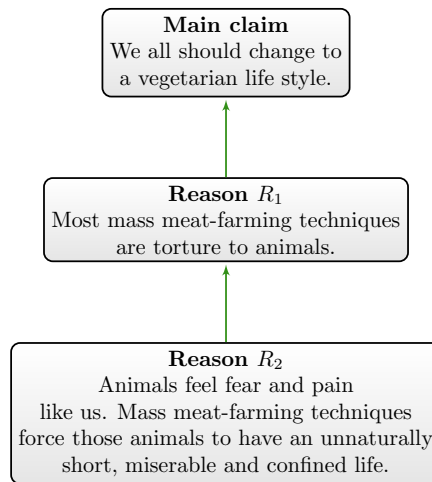


ABBILDUNG A.1  
Graphical representation of supporting reasons as a *reason map*.

form (e.g., ' *x is an objection to y.*'), which makes the relation of that relation explicit. Often, an objection to *y* is a reason that *y* is false or implausible.

So, what is the nature and target of Audrey's objection? She does not question that animals feel fear and pain when they grow up in meat farms. As a consequence, the target of her objection is not the second reason ( $R_2$ ). She also does not question the truth or plausibility that most mass meat-farming techniques can be regarded as torture to animals ( $R_2$ ). But she formulates a reason that animal suffering is not a sufficient reason for becoming a vegetarian since there are alternatives to mass-produced meat, which avoids the suffering of animals, or so her reasoning goes. Hence, we could say that the target of Audrey's objection is Clint's first reason ( $R_1$ ) because her objection questions the relevance or sufficiency of  $R_1$ .

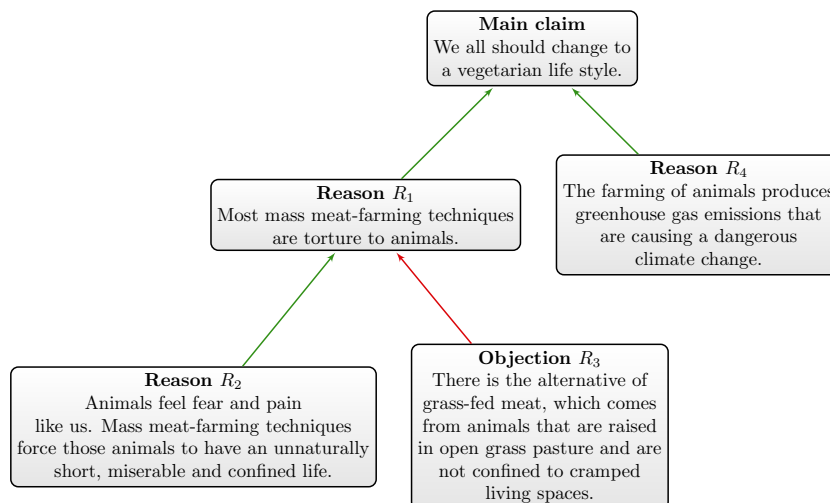


ABBILDUNG A.2  
Graphical representation of the dialogue's argumentation structure as a *reason map*.

Objections will be visualized in a reason map by an attack relation represented by a red arrow as illustrated in Figure A.2.

Clint's final utterance illustrates that argument structures can be more branched than the simple serial reasoning we began with. His reply to Audrey does not refute her objection ( $R_3$ ). On the contrary, he seems to concede her point. Hence, Clint's utterance is not an objection to something. Instead, he presents an alternative reason to the main claim as visualized in Figure A.2.

### A.1.2 TWO TASKS OF ANALYSING ARGUMENTATION STRUCTURE

This simple dialogue exemplifies that analysing the argumentation structure of a text involves two interdependent tasks:

*1. Annotation of reasons and claims:* You should annotate a text segment as *justificatory relevant* if it is being presented as a justification or if some other text segment is presented as a justification for or as an objection to it.

What we usually identify as arguments, reasons, objections, and refutations is, according to this definition, justificatory relevant. For instance, Audrey's critical questions are not justificatory relevant since they do not formulate reasons or objections on their own. But Clint's replies are justificatory relevant. Similarly, text segments that are the target of arguments, reasons, objections and refutations are justificatory relevant according to this definition. While Clint's first utterance is not presented as a justification for something else, other utterances are presented as a justification for it.

In the following, we will simply speak of reasons since objections and refutations can always be understood as *reasons against* something. If  $x$  is an objection to  $y$ , it can usually be understood as a reason that  $y$  is false, implausible, not acceptable, irrelevant or something similar. Additionally, we will neither distinguish between reasons and arguments on the one hand nor between different kinds of objections on the other hand.

As the example illustrates, a text segment can have both roles. It can be presented as a justification for something else, and, at the same time, there can be another text segment that is presented as a justification for it (e.g., Clint's second answer). In the following, we will refer to text segments that are not used as a justification for something but that are the target of presented justifications as claims or main claims.

It is also important to keep in mind that the definition of *justificatory relevance* first and foremost demands understanding the author's intentions. It is about understanding what is *presented* as a justification—independent of how you or others would evaluate the success of this presented justification.

Finally, the annotation of a text segment demands individuating reasons. What is meant by that? The annotation of a text segment requires determining the start and endpoint of the text segment—often by highlighting the corresponding text segment in some way. The text segments we are interested in can encompass sentences, clauses or even smaller units but will not split up words. Each annotated text segment should correspond to exactly one reason or one target of a reason. If a longer text segment contains more than one justification, you have to split up the longer text segment into smaller units of text



segments. The process of identifying justificatory relevant text segments can, therefore, be regarded as the process of individuating reasons in a text.

The first task asks you to select and annotate text segments as justificatory relevant; the second task asks you to categorise the exact role of these text segments.

2. *Identifying support and attack relations:* Categorise for each justificatory relevant text segment its ingoing and outgoing support and attack relations.

A text segment is justificatory relevant because it is presented as a reason for some other text segment or as the target of a reason. The second task demands categorising these relationships between justificatory relevant text segments. You should only use two fundamental relations: You have to categorise text segments as supporting something (e.g., reasons and arguments) or as attacking something (e.g., objections and refutations), and you have to identify their targets—that is, what they are supporting or attacking.

A text segment  $x$  that is presented as a reason for some other text segment  $y$  should be categorised via the *support relation* between  $x$  and  $y$  (represented by a green arrow from  $x$  to  $y$ ), and a text segment  $x$  that is being presented as a reason against some other text segment  $y$  will be categorised via an *attack relation* between  $x$  and  $y$  (represented by a red arrow from  $x$  to  $y$ ).

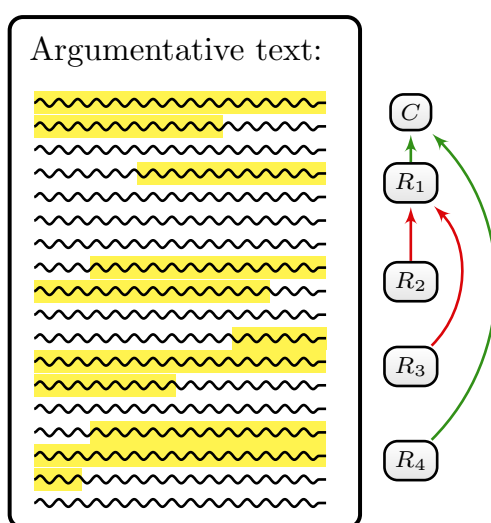


ABBILDUNG A.3

A schematic example of annotating the argumentation structure of a text. There are five annotated text segments: The first text segment represents the main claim ( $C$ ), which is supported by two reasons that are formulated by the second ( $R_1$ ) and the fifth text segment ( $R_4$ ). Both the third ( $R_2$ ) and the fourth text segment ( $R_3$ ) represent reasons against—that is, objections to the supporting reason  $R_1$ .

Figure A.3 schematically illustrates the result of completing both tasks: an annotated text that identifies justificatory relevant text segments and the categorisation of the interrelationships between these text segments as support and attack relations.

The analysis of argumentation structures is often more complicated and challenging than the simple example suggests:

- First, you might find it difficult to decide whether a particular text segment is meant as a reason, objection or something else. You will often find text segments with other functions than simply formulating justifications for some other text segment. The line between what counts as justificatory relevant and justificatory irrelevant is partly blurry and continuous.
- Second, you might ask yourself where a reason starts and ends.
- Third, what exactly is supposed to be justified by a reason is sometimes ambiguous. It might even happen that the claim that is justified by a text segment is itself implicit—that is, not represented by another text segment.
- Finally, you will often wonder whether a text segment that is formulated as a justification represents just one or several reasons. We will refer to this latter challenge as the *problem of reason individuation*.

It is not possible to understand the argumentation structure of a text without understanding the conveyed meaning of the text. Since the understanding of meaning is always an interpretation, which is to some extent subjective and can differ from person to person, the understanding of argumentation structure is also a process of interpretation. Since we can seldom ask the authors for clarifications of what they meant—at least not without considerable effort—the problem of multiple possible interpretations is particularly challenging in our context. However, the degree of interpretation can be increased to some extent by being more precise and explicit about the two mentioned tasks of analysing the structure of argumentation.

In the following, you will be provided with clarifications and maxims that will help you alleviate the problems mentioned. First, we will have to understand what makes text segments justificatory relevant. In other words, we will have to say more about what we mean with reasons, arguments and objections and how they can be distinguished from text segments with functions other than justifying something else. We will then go on to address problems of individuating reasons. Finally, we will look closer at how to determine the targets of the support and attack relations. More minor challenges will be dealt with in the form of frequently asked questions at the end of these guidelines (see page 273).

### A.1.3 USED EXAMPLES: GERMLINE EDITING AND AFFIRMATIVE ACTION

To give you a realistic impression of how reasons are formulated, we will use examples from several online articles about two different debates:

1. **Germline editing (GE):** The GE debate is about a biotechnology called CRISPR, which can alter genetic material more accurately than older gene-modifying technologies. It can be used to create genetically modified crops, develop vaccines, and treat genetic diseases. In contrast to somatic gene therapies, germline gene therapies target reproductive cells, which eventually form embryos, sperms and oocytes. One possible scenario is to use CRISPR in combination with in vitro fertilisation (IVF). After placing a single sperm into a single egg in vitro, CRISPR is used to modify genes of the embryo that cause genetic diseases such as Huntington's disease, Tay–Sachs disease and Duchenne muscular dystrophy. The important difference between somatic and germline gene therapies is that alterations of germline cells will be passed to further generations. The GE debate gained prominence in 2018

when the Chinese researcher He Jiankui claimed that he used CRISPR to edit the genomes of embryos and that two non-identical twin girls had been born as a result. Scientists worldwide condemned He's actions. Most ethicists and researchers believe that human germline editing should not be used as a therapy at this time because of its unknown risks and side effects. The current GE debate is about the following central questions:

- Should we use GE in the future for therapeutic means? Under which conditions should GE be permitted as a therapy?
  - Should we research germline editing technologies by using embryos? Under which conditions should we allow scientists to alter the genes of embryos for purposes of research? How should we regulate such research?
2. **Affirmative action at universities in the United States:** Affirmative action is an umbrella terminus for practices and regulations that seek to eliminate underrepresentation that resulted from historical injustices based on gender, race or religion of certain groups in education, employment and politics. Affirmative action was first introduced in the United States in 1961 by President John F. Kennedy's Executive Order 10925, which required federal contractors "to ensure that applicants are treated equally without regard to race, colour, religion, sex, or national origin." Measures of affirmative action can differ with respect to the means they employ and with respect to the groups they target.

The text snippets we will analyse are about affirmative action in college admissions. The main question of this debate is whether it is permissible to include race-based categories in admission criteria to promote diversity on campus and to counteract the underrepresentation of racial minorities at elite universities.

#### A.1.4 JUSTIFICATORY RELEVANCE – SOME IMPORTANT CLARIFICATIONS AND MAXIMS

The given definition of what counts as justificatory relevant (see page 250) covers a lot of different ways to formulate reasons and objections. The notions of reason and objection carry different connotations to different people, and some are irrelevant to the defined concept of argumentative relevance. In particular, you do not have to worry about the following aspects:

*Clarification 1:* You do not have to evaluate presented reasons and objections.

A text segment counts as justificatory relevant independent of whether the presented justification succeeds in its justification. It does not matter whether a presented reason is a good reason. What counts is the author's intention. You only have to ask yourself if the author meant the text segment to represent a justification.

Often, an author will formulate their own stance. They will present *their* arguments and possibly try to refute other's objections to their own position. However, you can also encounter argumentative texts in which the author presents reasons and objections without revealing their own position. What is essential here: Someone can present reasons without

taking a stance. Understanding the author's stance might, in some cases, help to identify the presented reasons. However, whether some text segment counts as justificatory relevant is independent of whose reasons are presented.

*Clarification 2:* It is not necessarily about who puts forward a reason.

Consider the following example:

*Example 1:* Some people worry that it is impossible to obtain informed consent for germline therapy because the patients affected by the edits are the embryo and future generations. The counterargument is that parents already make many decisions that affect their future children, including similarly complicated decisions such as PGD [preimplantation genetic diagnosis] with IVF [in vitro fertilization]. (NHGRI 2017)

The authors present two arguments—one pro germline therapy and one against it—without explicitly adopting their own stance. They merely describe two existing arguments without evaluating them and without revealing their own position towards germline therapy.

Consequently, we cannot determine whether the authors want to persuade the audience of something, which brings us to the following clarification.

*Clarification 3:* It is not necessarily about persuasion.

Usually, the author of an argumentative text or utterance formulates reasons with the intention to persuade the audience of their main claims. However, an author might present reasons without such an intention. They might, for instance, simply want to summarize a debate. Identifying justificatory relevant text segments is, therefore, independent of why the author presents something as a reason.

The following clarifications will show that the suggested notion of justificatory relevance will include a wide variety of reasons, arguments, objections and refutations.

*Clarification 4:* Reasons and objections can be presented indirectly.

The following example contains two different types of indirect presentation of reasons:

*Example 2:* For many people, the very idea [of germline editing] feels unnatural and wrong, and I was one of those people when I first started thinking about the issue. Humans have been reproducing for millennia, aided only by the DNA mutations that arise naturally, and for us to begin directing the process—similar to the way that plant biologists might genetically modify corn—seems almost perverse at first glance. As National Institutes of Health director Francis Collins put it, “Evolution has been working toward optimizing the human genome for 3.85 billion years. Do we really think that some small group of human genome tinkerers could do better without all sorts of unintended

consequences?” (Doudna and Sternberg 2017)

The first two sentences appeal to emotions. It is about what people feel toward the idea of germline editing. There is a common misconception that emotions and arguments exclude each other. While it is true that emotions are no reasons on their own, emotions can express or point to reasons. The reason that is implicitly formulated in the above text snippet is known as the argument from naturalness: We should not pursue germline editing because it is unnatural.

The second reason is formulated as a question in the last sentence. The question is clearly meant as a rhetorical question. The presented reason points to the possibility of severe side effects: Germline editing will have unintended side effects because human genome editing cannot be better than evolutionary optimization of the human genome.

Again, it is not important how we evaluate the argumentative strength of these reasons to categorize them as justificatory relevant (see *clarification 1*). What the example shows is that reasons can be formulated in indirect ways, for instance, by an appeal to emotions and by rhetorical questions.

### MAXIM 1: USING LINGUISTIC CUES

How do we recognize text segments as justificatory relevant? Often, authors employ linguistic cues to tell the audience what they intend to be a reason or objection.

*Maxim 1:* Search for linguistic cues that indicate what text segments are presented as justifications!

Linguistic cues are indicator words, phrases, clauses, or whole sentences used by an author (or interlocutor) to help the audience interpret the text’s (or utterance’s) intended meaning. Linguistic cues for reasons and objections act as direct or indirect signs that characterize text segments as justificatory relevant.

Typical phrases that indicate a support relation between x and y include:

- *x is true. Hence/therefore, y.*
- *y because x.*
- *y since x.*
- *y for the reason that x.*
- ...

Attack relations can be formulated, for instance, by the following phrases:

- *x is false/implausible/irrelevant because y.*
- *x. This is obviously false since y.*
- *x. This argument is based on the assumption that ..., which is false since y.*
- *An important objection to x is that y.*
- ...

Often, text segments are presented as reasons by attributing claims together with their reasons and objections to specific groups or people:

- *s argues for x by y.*
- *The main reason for x is put forward by s. According to them . . . y.*
- *s objects to x by saying that y.*
- . . .

One important question is whether you should annotate linguistic cues as justificatory relevant—in other words, whether you should count them as belonging to the formulation of reasons. There is no particular reason to do it one way or the other. However, to compare the annotations of different analysts, we should adopt one consistent rule for all analysts. In the following, we will, in most cases, not annotate linguistic cues as justificatory relevant (for more details, see question 7 on page 276). Nevertheless, searching for linguistic cues and annotating them is very helpful. In the following examples, we will either underline or italicize linguistic cues.

### MAXIM 2–3: USING THE COTEXT

While recognizing linguistic cues is an indispensable means to identify justificatory relevant text segments, you should not rely on finding linguistic cues. Often, linguistic cues will be amiss. Fortunately, the cotext will usually provide enough information to decide on the justificatory relevance of a text segment.

*Maxim 2:* Consider the cotext of a text segment to decide on its justificatory relevance!

The cotext of a text segment *S* in a text *T* is *T* minus *S* itself.<sup>279</sup> Let's have a second look at the first example:

*Example 1:* *Some people worry that* it is impossible to obtain informed consent for germline therapy because the patients affected by the edits are the embryo and future generations. *The counterargument is that* parents already make many decisions that affect their future children, including similarly complicated decisions such as PGD [preimplantation genetic diagnosis] with IVF [in vitro fertilization]. (NHGRI 2017, emphasis added)

We already characterized the first sentence as a reason against germline therapy. How do we know that? The phrase '*Some people worry that x.*' is not a unique linguistic cue for *x* being a reason. It simply presents *x* as being a worry. But often, such ascriptions serve as indirect means to express a reason. In this example, the cotext—here, the second sentence—corroborates this interpretation. It contains a linguistic cue ('*The counterargument is*') for an attack relation. It not only tells us that the second sentence expresses an objection but that some of the preceding sentences must contain a reason to which it is an objection.

There is another difficulty with linguistic cues besides their sometimes sparse usage by authors. You should not mindlessly follow them because they can stand for different

<sup>279</sup>Depending on the resources, it might even be possible to invoke contextual knowledge of *T*. The context includes every relevant knowledge to the interpretation of *T* that is not contained in *T*. For instance, other published works of the author might be helpful in interpreting the particular text *T*. However, we assume you do not possess any particular contextual knowledge here. We even ask you not to inquire into additional contextual information to ensure that every analyst invokes roughly the same background knowledge.

intentions. One and the same linguistic cue can be utilized for different purposes in different contexts. Again, you have to take the cotext into account.

*Maxim 3:* Consider the cotext of a linguistic cue to decide on what they signify!

#### MAXIM 4: JUSTIFICATORY AND EXPLANATORY REASONS

A prevalent example is the phrase ‘*x because y*’. Sometimes, it will signpost a presented justification; sometimes, it will signpost an explanation. In both cases, the author presents reasons. However, we have to distinguish between *justificatory reasons* and *explanatory reasons*. Up until now, we only considered justificatory reasons—that is, reasons that are presented as justifications for something. In contrast to justificatory reasons, explanatory reasons provide explanations. A justification for *x* is usually formulated to convince the audience that *x* is true. In this case, the author assumes that the audience needs to be convinced of *x* and provides reasons that justify the truth of *x*. An explanation for *x* is used for different purposes. It can be offered to inform the audience about the cause of *x*, or it can explain the meaning of *x*. In contrast to a justification, an explanation for *x* does not aim to establish the truth of *x*. The truth is either presupposed or simply not the issue at stake. In other words, justifications for *x* answer to the question ‘*Is x really true?*’; explanations for *x* answer to questions such as ‘*Why is x true?*’, ‘*What is the explanation for x?*’ or ‘*What is meant with x?*’

In the following examples, the because-phrase expresses a causal explanation:

*Example 3:* Meanwhile, the average white household has a net worth 13 times that of the average black household and 10 times that of the average Hispanic household, largely *because* whites inherit greater generational wealth. (J. Gordon and Wang 2019, emphasis added)

The inheritance of greater wealth in white households is offered as a causal explanation of why the average net worth of white households exceeds those of black and Hispanic households. We can think of this sentence as an answer to why white households exhibit a larger average net worth.

*Example 4:* Some heterosexual couples may hesitate to use this option [using third-party eggs or sperm] *because* they want a child who is not just spared a deleterious gene in their lineage, but is also genetically related to both of them. (Harris and Darnovsky 2018, emphasis added)

This sentence presents an explanation of behaviour in terms of personal preferences. It answers the question of why some couples prefer not to have children that result from using third-party eggs or sperm.

Understanding whether a presented reason is intended as a justification or an explanation is important. Understanding this distinction is, unfortunately, not sufficient to decide on the justificatory relevance of text segments. Even if a statement offers an explanatory reason, it can be justificatory relevant. There are several difficulties. A text segment can express both a justificatory and an explanatory reason at the same time. Another difficulty is that statements that express a causal explanation can be part of a justification for something.



Alternatively, a text segment might elaborate on the meaning of some other text segment that is meant as a justification. There are no clear-cut rules for the interpretation of such cases, as the following examples illustrate:

*Example 5:* When it comes to experimentation with human embryos, some Christian communities are opposed *because* they regard the embryo as a person from conception, whereas Jewish and Muslim traditions tend to be more accepting *because* they do not consider embryos created in vitro to be people. (Doudna and Sternberg 2017, emphasis added)

The two occurrences of the linguistic cue '*because*' signal explanatory reasons in this example. The reader is offered explanations for the attitudes of different religious communities towards germline editing on embryos for research purposes. In the example, these attitudes are explained in terms of beliefs about embryos. The first explanation explains why Christians are against such experimentation and the second why Jews and Muslims are "more accepting".

However, we can also interpret the first explanation as a presented justification. It tells us not only *why* Christians are opposed to germline editing but also how they would justify their stance. The Christian would justify the prohibition of such research with the status of embryos as persons. Such an interpretation is not possible with the second explanation. While it similarly explains the Jew's and Muslim's stance toward the status of the embryo, we do not get to know for what that would be a reason. We only get to know that their belief about the status of the embryo is *not a reason against* germline editing.

The following example provides cotext of the previous example four.

*Example 4:* Would germline gene editing be justifiable, in spite of the risks, for parents who might transmit an inherited disease? It's certainly not necessary. Parents can have children unaffected by the disease they have or carry by using third-party eggs or sperm, an increasingly common way to form families. Some heterosexual couples may hesitate to use this option *because* they want a child who is not just spared a deleterious gene in their lineage, but is also genetically related to both of them. They can do that too, with the embryo screening technique called preimplantation genetic diagnosis (PGD), a widely available procedure used in conjunction with in vitro fertilization. (Harris and Darnovsky 2018, emphasis added)

We already characterized the sentence containing the '*because*' as expressing an explanation of *why* some couples prefer not to use third-party eggs or sperm for their reproduction. But the sentence is justificatory relevant since it is part of a complex reasoning against germline editing. The authors argue that germline editing is not necessary since there are other options for parents to have children without a genetic disease: There is the option to use egg or sperm donation, and those who prefer to have a child of their genes can use PDG together with IVF.

The last example employs another linguistic cue for a causal relationship.

*Example 6:* While genetic correction poses some ethical concerns regarding which diseases and disabilities should be targeted, genetic enhancement may re-



sult in dangerous social implications. [...] For example, if parents use CRISPR to select for a specific eye color or skin tone, they will likely choose the most socially preferred traits. *As a result*, increased discrimination and prejudice may follow for those who stray from those specific traits. (Jain 2021, emphasis added)

The phrase '*as a result*' characterizes increased discrimination and prejudice as a causal effect of parents selecting socially preferred traits when they are at liberty to choose the genetic makeup of their offspring. Again, the causal explanation is justificatory relevant since the causal statement enters into a reasoning against using germline editing for enhancement purposes. According to this argumentation, we should refrain from using germline editing because of a causal relationship—namely, that enhancement would lead to increased discrimination.

These examples suggest that deciding whether explanations are justificatory relevant cannot be answered in a general way. Again, the categorization must take into account the whole cotext.

*Maxim 4:* Be aware of the difference between justificatory and explanatory reasons. Use linguistic cues and the cotext to decide whether explanations are justificatory relevant.

#### MAXIM 5: THE ROLE OF EXAMPLES

The use of examples in argumentative texts is another exemplification of the need to distinguish between justifications and mere explanations. An author might use an example as a mere illustration to explain what was meant by another text segment or as a justification. Unambiguous linguistic cues for examples include phrases such as '*for instance*,...' and '*for example*,...'

How do we decide whether a formulated example is intended as a justification? An example can only be justificatory relevant if there is a statement that is supposed to be justified with the example. For instance, propositions that claim the existence of something are typically justified by providing an example that corresponds to an instance of what is claimed to exist.

*Maxim 5:* To decide on the justificatory relevance of examples, try to identify text segments that are supposed to be justified with the example.

Take the example we've just discussed:

*Example 6:* While genetic correction poses some ethical concerns regarding which diseases and disabilities should be targeted, genetic enhancement may result in dangerous social implications. [...] *For example*, if parents use CRISPR to select for a specific eye color or skin tone, they will likely choose the most socially preferred traits. As a result, increased discrimination and prejudice may follow for those who stray from those specific traits. (Jain 2021, emphasis

added)

The first sentence claims the existence of dangerous social implications. Such an existential proposition can be justified by providing an instance of what it claims to exist. The used example, which is introduced with the linguistic cue '*for example*', is presented as an attempt to do precisely that. It describes the dangerous social consequence of increased discrimination and prejudice. The example is, therefore, presented as a justification for the existential claim.

### A.1.5 INDIVIDUATION OF REASONS

Categorizing *one* particular text segment as justificatory relevant means categorizing it as the formulation of *one* reason for or against something. In other words, identifying a reason in a text demands identifying a text segment that corresponds to the formulation of that reason. But where does the formulation of a reason exactly start and end? Should we count text segments that repeat or reformulate reasons as part of the reason's text segment itself? How do we tell different but similar reasons apart? All these questions pertain to the issue of individuating reasons. The following examples will motivate some maxims that will help you to individuate reasons.

In Figure A.4 you find an example with a suggestion to annotate justificatory relevant text segments and their accompanying support and attack relations. In the example, Doudna responds to the argument from evolutionary optimization against germline editing, which we already encountered in the second example. Let us understand the suggested interpretation of the text snippet step by step.

*Example 7:* If CRISPR can help parents conceive a disease-free child when no other options exist and it can do so safely, ought we to use it? It's a question I've asked myself again and again [...]

As National Institutes of Health director Francis Collins put it, "Evolution has been working toward optimizing the human genome for 3.85 billion years. Do we really think that some small group of human genome tinkers could do better without all sorts of unintended consequences?" [...]

While I share the general feeling of unease at the idea of humans taking control of their evolution, I wouldn't go so far as to say that nature has fine-tuned our genetic composition. Obviously, evolution didn't optimize the human genome for the present era, when modern foods, computers and high-speed transportation have completely transformed the way we live. And if we look over our shoulders at the course of evolution that has led to this moment, we'll see it's littered with organisms that didn't benefit from the mutational chaos that underpins evolution. Nature is less an engineer than a tinkerer—and a fairly sloppy one at that. Its carelessness can seem like outright cruelty for those people who have inherited genetic mutations that turned out to be suboptimal. (Doudna and Sternberg 2017)

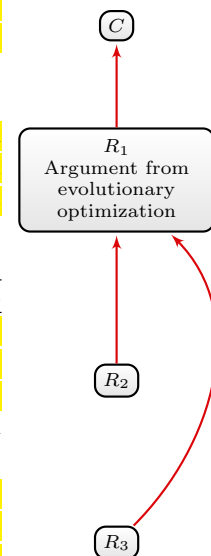


ABBILDUNG A.4

An example of a reason ( $R_3$ ) that extends over many clauses. Identified linguistic cues are underlined, and justificatory relevant text segments are highlighted.

It is not difficult to determine the exact boundaries of the justificatory relevant text segments in the first two paragraphs. The text snippet starts with a question, which indirectly formulates the main claim in question.<sup>280</sup> We already understood the rhetorical question of the second paragraph as an objection to germline editing ( $R_1$ ). According to this objection, germline editing will have unintended side effects that are to be expected when “human genome tinkers” try to do better than evolution. The point about evolutionary optimization seems to be important to understand the main idea of the argument, which motivates us to consider the sentence that precedes the question as part of the reason against germline editing.

The third paragraph is more challenging. The author starts by making explicit her stance on the argument from evolutionary optimization—namely, that she is not convinced of it. The first sentence does not formulate a reason itself but can be considered a linguistic cue that the following sentences present reasons against the argument from evolutionary optimization. She goes on to elaborate on two different ways in which our genome is not optimized. This and the used linguistic cue ‘*and*’ motivate us to interpret the paragraph as formulating two different points—that is, two different objections. The first objection ( $R_2$ ) alludes to a lack of genetic fit to our modern world and extends over one sentence. The second objection ( $R_3$ ) points to genetic diseases, which can, according to the author, hardly be described as the result of an optimization. I tend to regard all three sentences beginning with the ‘*and*’ as relevant to understanding the main idea of the objection. The main point is that evolution is based on random genetic variations—in contrast to intentional alterations—that result in “suboptimal” results on the individual level. All three sentences contribute to understanding this one point, which motivates us to consider all three sentences as formulating one objection. Admittedly, the metaphor about looking over one’s shoulder is irrelevant to the reason itself.

The example shows that one justificatory relevant text segment does not necessarily span exactly over one sentence.

*Clarification 5:* A reason can be formulated by one sentence, a subclause of a sentence, or stretch over more than one sentence.

## MAXIM 6–8: IDENTIFYING DIFFERENT PARTS OF ONE REASON

The last clarification shows that it is essential to have some criterion to determine where a reason starts and ends. The intuitive analysis of the last example motivates us to count text segments that help us understand the reason’s main point or its relevance as part of the reason. Those text segments should, therefore, be annotated as belonging to the reason. Typical examples of linguistic cues that signpost such clarifications are ‘*This argument rests on the assumption that ...*’, ‘*... presuming ...*’ and ‘*This argument is based on the grounds that. ...*’ However, often you won’t find such explicit linguistic cues. The following maxim formulates a rule of thumb and accompanying questions that can help to identify the different parts of a reason.

<sup>280</sup>You will find more information on indirectly and implicitly formulated claims on pages 273 and 275.

*Maxim 6:* If sentences or subclauses must be taken together (or are intended to be taken together) to understand a reason’s main idea or relevance, they should be considered as belonging to the formulation of that reason. In other words, if it makes sense to imagine a sentence or subclause as an answer to one of the following questions—*Why is that a reason? Why is that relevant? What is the point of that?*—they are part of the reason.

*Example 8:* [...] in theory—and eventually in practice—CRISPR could be used to modify disease-causing genes in embryos brought to term, removing the faulty script from the genetic code of that person’s future descendants as well. Proponents of such “human germline editing” argue that[...]

Let’s start with the objection that embryo modification is unnatural [...]. But diseases are natural, and humans by the millions fall ill and die prematurely—all perfectly naturally. If we protected natural creatures and natural phenomena simply because they are natural, we would not be able to use antibiotics to kill bacteria or otherwise practice medicine, or combat drought, famine, or pestilence. The health care systems maintained by every developed nation can aptly be characterized as a part of what I have previously called “a comprehensive attempt to frustrate the course of nature.” (Harris and Darnovsky 2018)



ABBILDUNG A.5

An example of applying Maxim 6.  $R_2$  corresponds to a text segment spanning over many clauses, which clarify the main idea and relevance of  $R_2$ .

In Figure A.5, you will find an example that illustrates how to apply maxim 6. The main claim (C) is attacked by an argument from naturalness ( $R_1$ ), which is itself the target of a presented refutation ( $R_2$ ). According to Maxim 6, we can regard all three sentences as part of one objection to the argument from naturalness ( $R_1$ ) since the last two sentences seem to elaborate on the objection that diseases are natural. We can reasonably imagine the three consecutive sentences to be part of the following dialogue:

A: “But diseases are natural, and humans by the millions fall ill and die prematurely—all perfectly naturally.”

B: “Why is that a reason against the argument from naturalness?”

A: “Because, if we protected natural creatures and natural phenomena simply because they are natural, we would not be able to use antibiotics to kill bacteria or otherwise practice medicine, or combat drought, famine, or pestilence.”

B: “And? Why is that relevant?”

A: “[Because we do use antibiotics to kill bacteria.] The healthcare systems maintained by every developed nation can aptly be characterized as a part of what I have previously called a comprehensive attempt to frustrate the course of nature.”

*B* uses the questions of Maxim 6 to better understand or challenge the relevance of *A*'s objection to the argument from naturalness. *A* answers by repeating the annotated sentences we find in the second paragraph. This, in turn, justifies counting them as part of *A*'s objection.

In the last two examples, we identified a reason that extends over more than one sentence. These sentences were contiguous, and we simply extended the text segment corresponding to the reason over all sentences. It can, however, also happen that sentences or clauses that explain the main point and the relevance of a reason are not contiguous to one another, as the example in Figure A.6 shows.

*Example 9:* [...] But affirmative action is still desperately needed in America's best universities to combat the racial divisions that colorblindness does nothing to resolve. Here's why. [...]

While we have made significant progress on addressing America's racist past, we still have much to do regarding the racism that our society continues to tacitly allow.

It should come as no surprise that whites continue to have a higher median income than black and Hispanic Americans (we'll get into why Asian Americans have a higher median income than even whites, as well as their role in the affirmative action debate, later). The average black American earns close to 41 percent less income than the average white, while the typical Hispanic earns 27 percent less income than their white counterparts.

Noting these statistics, some have proposed income-based affirmative action as an effective alternative to race-based affirmative action, presuming that the problems plaguing America's minority communities are caused by class, not race. But while income inequality does exacerbate racial divisions in America, it cannot fully account for those divisions. Even controlled for income, racial minorities face challenges their white peers don't have to deal with. (Gordon and Wang 2019)

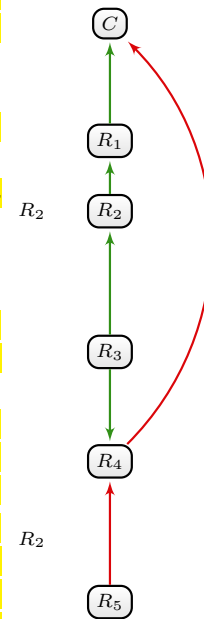


ABBILDUNG A.6

An example of applying Maxim 6. Two non-contiguous text segments belong to one reason ( $R_2$ ).

The authors argue that race-based affirmative action is still needed (the main claim  $C$ ) as a means to reduce the still existing racism in the United States ( $R_1$ ). In the third paragraph, the authors go on to elaborate on the still existing higher incomes of white Americans as compared to black and Hispanic Americans ( $R_2$ ), for which they provide further statistical evidence ( $R_3$ ).<sup>281</sup> Naturally, we can interpret these income differences as a supporting reason for contemporary racism without there being linguistic cues to corroborate this interpretation. But why are income differences relevant to racism? The authors provide one answer themselves. As the first clause in the last sentence clarifies, income differences “exacerbate racial divisions.” Therefore, we can regard this clause as explaining the relevance of  $R_2$  and should annotate it as belonging to  $R_2$ .

<sup>281</sup>Alternatively, we could regard both non-contiguous text segments (the third and fourth annotation) as one reason. On this interpretation, the fourth text segment is not a presented justification for the third text segment but rather a reformulation that adds numerical details of what is meant by the latter. See Maxim 7 below.

The example illustrates that one reason ( $R_2$ ) can spread over non-contiguous sentences and clauses. There is one practical difference in the case of non-contiguous sentences and clauses. If the sentences and clauses that belong to a reason are contiguous, you annotate these sentences as one text segment by determining the start and endpoint accordingly. However, if the sentences and clauses that belong to a reason are non-contiguous, you have to annotate *different* text segments as belonging to one reason.

According to Maxim 6, elaborations on the reason's main idea or relevance should be annotated as part of the reason. However, not every text segment that explains reasons should be annotated this way.

*Maxim 7:* Sentences or clauses that provide mere illustrative examples or contextual information and explanations that merely explain the meaning of relevant concepts should not be annotated as belonging to the reason.

*Example 10:* Genetic correction refers to the prevention and treatment of disease, while enhancement has been defined as boosting our capabilities beyond the species-typical level or statistically normal range of functioning, according to the National Science Foundation.

While genetic correction poses some ethical concerns regarding which diseases and disabilities should be targeted, genetic enhancement may result in dangerous social implications. [...]

In the case that its [CRISPR's] efficiency and precision increase enough to become integrated with healthcare on a wide scale, this biotechnological tool should be used only for disease prevention and correction as opposed to feature enhancements. (Jain 2021)



ABBILDUNG A.7

An example of applying Maxim 7. The first paragraph explains the difference between therapy and enhancement, which is mere contextual information and not part of the reason  $R_1$ .

In the example of Figure A.7, the author provides contextual knowledge that should not be annotated as justificatory relevant. The author's main claim is formulated at the end of the text. She argues for confining the use of germline editing to therapeutic uses and banning its use for enhancement. One of her reasons against genetic enhancement invokes the possibility of "dangerous social implications." While her explanations on how the treatment of diseases differs from enhancement help to better understand the reasoning, it is not necessary to understand the main point of the reason. To understand the main idea of the reason, we simply need to know that enhancement differs from therapy and that enhancement is said to have certain dangerous consequences. Since the second sentence implies both points, we do not need the first sentence to understand the main idea of the reason.

A similar but slightly different case to clarifications and elaborations of reasons are text segments that simply reformulate or repeat a reason. An author might, for instance, explain the same reason in different words to foster a better understanding or to remind the audience of an important point. Similarly, an author might begin by summarizing the main idea of a

reason and later provide further details of the same point. In line with clarifications of the reason's relevance (Maxim 6), you should consider text segments that are reformulations of the same reason as belonging to that reason:

*Maxim 8:* If sentences or clauses are (or are intended to be) reformulations or repetitions of a reason, they should be considered as belonging to the formulation of that reason. Mere references to reasons should not be annotated as justificatory relevant.

Typical linguistic cues for repetitions and reformulations include 'In other words, ...' and 'As x puts it ...'. Similar to the last examples, the repetition of reasons can extend over contiguous or non-contiguous sentences and clauses.

*Example 11:* [...] But affirmative action is still desperately needed in America's best universities to combat the racial divisions that colorblindness does nothing to resolve. Here's why. [...] [...] Affirmative action [...] [is] about deciding which qualified individuals would benefit the most from a great education. And minority students stand to gain quite a bit. Attending a highly selective college (the kind that would probably use affirmative action) produces larger increases in income and employment rates for blacks and Hispanics than for whites. Researchers have also concluded that students are more likely to graduate when they attend the most selective college that will accept them, meaning that a minority student can still succeed and graduate without having a prep-school background or a perfect SAT score. Granting more underprivileged students the opportunity to succeed through education not only could lift some individuals out of a vicious cycle of poverty, but might also... (Gordon and Wang 2019)

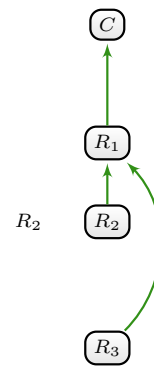


ABBILDUNG A.8

An example of applying Maxim 8 to identify reformulations of reasons. Two non-contiguous text segments formulate the main idea of one reason ( $R_2$ ) differently.

Similarly to sentences and clauses that elaborate on the relevance of reasons, reasons and their reformulations can come in contiguous or non-contiguous text segments. The example in Figure A.8 depicts an annotation of an argumentation structure with a non-contiguous reformulation of the second reason ( $R_2$ ). This reason is formulated as a justification for the claim that minority groups can benefit more from education at elite universities than other groups ( $R_1$ )—namely, that they would benefit more from it in terms of increase in income and employment. After formulating another benefit ( $R_3$ ), the authors reformulate the second reason in the third paragraph. By using the phrase “not only ..., but also ...” the authors begin to pick up an already formulated point ( $R_2$ ) and then move on to add a reason ( $R_3$ ).<sup>282</sup>

<sup>282</sup>There are, as often, alternative interpretations of the text. What I categorized as a reformulation can be, for instance, interpreted as a repetition of both points that are expressed with  $R_2$  and  $R_3$ . After all, the higher income ( $R_2$ ) presupposes that the individual who is lifted out of poverty graduates ( $R_3$ ).



Reformulations and repetitions usually have an explanatory function. They help the audience understand the reason, remind them of its basic idea, and might describe them from another perspective. In contrast to repetitions and reformulations, mere references do not contribute to the understanding of a reason and should not be annotated as justificatory relevant. A reference to a reason can be formulated, for instance, by using demonstrative determiners such as *'this'*, whose referents are context-dependent, by using proper names such as *'the slippery slope argument'*, or by using definite descriptions such as *'the first argument mentioned by Clint'*. References to reasons can be used after the author elaborates on a reason, for instance, to signpost attack and support relations, or they can be used before the author elaborates on a reason, for instance, to announce reasons. The following linguistic cues encompass some possibilities to refer to reasons:

- *'... This reason is not convincing since...'*,
- *'... The most prominent objection to the argument from x is ...'*,
- *'All of these objections can be refuted.'*, *'There are two relevant arguments: the argument from x and the argument from y ...'* and
- *'After clarifying the stance of y, I will analyze her strongest argument.'*

#### MAXIM 9–10: IDENTIFYING DIFFERENT REASONS AND THEIR RELATIONSHIPS

Maxims 6 and 8 will help you identify all text segments that belong to the same reason. Another important challenge is correctly realizing when contiguous text segments correspond to different reasons.

The example in Figure A.9 begins with an objection ( $R_1$ ) against germline editing for therapeutic uses ( $C$ ). The objection rests on the premise that there is always a safer alternative to the therapeutic use of germline editing—namely, preimplantation genetic diagnosis (PGD) in combination with in-vitro fertilization (IVF). According to the second reason ( $R_2$ ), this alternative is, in fact, not an alternative for all prospective parents. The presented refutation of  $R_1$  is supported by three reasons ( $R_{3-5}$ ). So why should we regard these nearly contiguous text segments as three different supporting reasons for  $R_2$ ? The chosen grammatical form is one hint. The authors invoke an enumeration to list points that seem to be on an equal footing with each other.

Additionally, Maxims 6 and 8 do not suggest considering the text segments as one reason: Neither of these points clarifies the relevance of the others (Maxim 6), and neither of these points is a reformulation of the others (Maxim 8). On the contrary, the points are independent. Every one of them supports, or is intended to support  $R_2$  without needing the other one; every one of them can be imagined as part of a dialogue in which a challenger successively asks for additional reasons that support the claim that PDG and IVF are not an alternative for all prospective parents ( $R_2$ ). Summarizing these considerations into another maxim:

*Maxim 9: If sentences or clauses provide or are intended to provide independent justification for some text segment, or if they are intended to provide support on their own, they should be considered as independent reasons. In other words, if it makes sense to imagine each sentence or clause as an answer to the question 'Can you give*



*me an additional reason?’, they correspond to different reasons for the same text segment.*

*Example 12:* [...] Some researchers argue against the use of germline editing for therapeutic uses that there may never be a time when genome editing in embryos will offer a benefit greater than that of existing technologies, such as preimplantation genetic diagnosis (PGD) and in-vitro fertilization (IVF). [...]

However, scientists and bioethicists acknowledge that in some cases, germline editing can address needs not met by PGD. This includes when both prospective parents are homozygous for a disease-causing variant (they both have two copies of the variant, so all of their children would be expected to have the disease); cases of polygenic disorders, which are influenced by more than one gene; and for families who object to some elements of the PGD process. (NHGRI 2017)

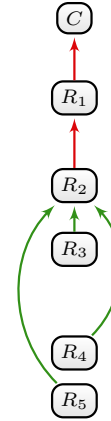


ABBILDUNG A.9

An example of applying Maxim 9. Three nearly contiguous text segments correspond to different reasons ( $R_{3-5}$ ) for one other text segment ( $R_2$ ).

Maxims 6 and 8 tell you to count reformulations of a reason and clarifications of its relevance as part of the reason itself. Sometimes, it isn't easy to distinguish a justification for a reason from a clarification of the reason. In other words, a text segment presented as a *reason* for another reason might look at first glance like a repetition or clarification of the second reason's main point. In this case, this apparent repetition should, of course, not be considered as part of the reason it supports. Rather, you should treat both text segments as independent reasons and categorize the first as supporting the second. In other words, you have to split the longer text segment into different ones and use the support relation to categorize the argumentative relation between them. Since it might be difficult to distinguish justifications for reasons from reformulations and clarifications of reasons, you should be attentive to text segments that appear to be one complex reason that stretches over many sentences.

By using the following example, we will formulate an additional maxim that will not only help you to spot the difference between reason-supporting reasons and reformulations of reasons but will be useful for analyzing the attack and support relations between text segments in general.

*Example 13:* But affirmative action is still desperately needed in America's best universities [...] Some well-intentioned people may argue that because of the academic challenges racial minorities tend to face, they may be less qualified than other students to attend elite universities and could struggle to meet the expectations of a rigorous curriculum. This is based in part on the debunked "mismatch" theory, which posits that these students would benefit more from attending less challenging schools more suited to their relative lack of educational attainment. (J. Gordon and Wang 2019)

After formulating the main claim, the remaining text snippet is introduced with a clear linguistic cue for an argumentation (*'people may argue that'*). The last sentence univocally

suggests that this argumentation is intended as an objection to affirmative action at universities. But is it adequate to interpret the paragraph as containing only two justificatory relevant text segments, one that represents the main claim ( $C$ ) and one that represents a complex objection to it ( $R_1$ ) as indicated in Figure A.10? Or is it a more complex reasoning that has to be broken down into different parts?

*Example 13:* But affirmative action is still desperately needed in America's best universities [...] Some well-intentioned people may argue that because of the academic challenges racial minorities tend to face, they may be less qualified than other students to attend elite universities and could struggle to meet the expectations of a rigorous curriculum. This is based in part on the debunked "mismatch" theory, which posits that these students would benefit more from attending less challenging schools more suited to their relative lack of educational attainment. (Gordon and Wang 2019)



ABBILDUNG A.10

First analysis of example 13. Annotating a long text segment as one objection ( $R_1$ ) to affirmative action at universities ( $C$ ).

To answer this question, we have to better understand the role of the additional linguistic cues. They pose at least two questions:

1. What does the '*because*' indicate? A presented justification, a causal explanation or both (see the Maxim 4 on page 257)?
2. How are we supposed to interpret '*This is based in part on ...*'? Does it indicate a clarification or some form of support relation between the text segments that precede and follow this phrase? If this phrase has to be interpreted as expressing a support relation, we should identify at least two reasons in the paragraph.

Our natural language is too flexible to answer such questions with simple and precise instructions. Confronted with such a complex case, you should, instead, employ the following rule of thumb:

**Maxim 10:** Reformulate the basic idea of reasons in your own words to identify support and attack relations!

- *Explicitness:* Your reformulation should be explicit about what is being justified by the reason.
- *Accuracy:* Try to stay as close as possible to the original formulation and the known or hypothesized intention of the author.
- *Charity:* If there are different and conflicting possibilities to reformulate the basic idea of a reason, choose the formulation that results in the most plausible and convincing reasoning.

The idea of the maxim is to make your understanding of the reasoning explicit. Explicitness

requires you to reflect on which text segment is being presented as a justification for which other text segment. That is why you should use a formulation that explicitly explains the relational character of reasons. For instance, you can use the following phrases for *y* being a reason that supports *x*:

- '*x because y.*'
- '*x since y.*'
- '*y is a reason for x.*'

and the following for *y* being a reason that attacks *x* (*y* being an objection to *x*):

- '*x is not true because y.*'
- '*x is implausible because y.*'
- '*y is a reason against x.*'

How can we formulate the main idea of the objection of the last example? One suggestion is the following:

Students from racial minorities (= “these students”) would benefit from attending less challenging schools *since* they could struggle to meet the expectation of a rigorous curriculum at elite universities.

This reformulation catches an important point of the reasoning and is formulated in one of the suggested forms ('*x since y*'). However, the clause following '*since*' does not directly formulate a reason against affirmative action since *x* says nothing directly about affirmative action. It is a reason in favour of why the described students should not attend elite universities. However, the context is clear enough to close this gap: Affirmative action is supposed to enable students from minority groups to attend elite universities. Since it would be better for them to not attend these universities, affirmative action is *not* needed, or so the argument goes.

This consideration suggests to separate the reasoning into two reasons:

Affirmative action is *not* needed *since* students from racial minorities would benefit from attending less challenging schools. (*C is not true since R<sub>1</sub>*)

and

Students from racial minorities (= “these students”) would benefit from attending less challenging schools, *since* they could struggle to meet the expectation of a rigorous curriculum at elite universities. (*R<sub>1</sub> since R<sub>2</sub>*)

According to this interpretation, the paragraph formulates a chain of reasons as illustrated in Figure A.11: The last annotated text segment represents a reason (*R<sub>1</sub>*) and is an objection to the main claim *C*, which is represented by the first text segment. The text segment following the '*and*' is another reason (*R<sub>2</sub>*), which is presented as a justification to support the first reasons *R<sub>1</sub>*.

The preliminary result is a much more adequate analysis of the argumentation structure than the first interpretation. The remaining task is to clarify the role of the text segment that precedes the '*and*'. It seems to be justificatory relevant. But so far, we have not used it in our reformulation of the reasoning structure. Additionally, we must still analyze the

*Example 13:* But affirmative action is still desperately needed in America's best universities [...] Some well-intentioned people may argue that because of the academic challenges racial minorities tend to face, they may be less qualified than other students to attend elite universities and could struggle to meet the expectations of a rigorous curriculum. This is based in part on the debunked "mismatch" theory, which posits that these students would benefit more from attending less challenging schools more suited to their relative lack of educational attainment. (Gordon and Wang 2019)

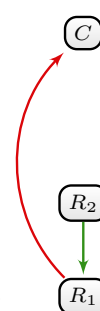


ABBILDUNG A.11

Second analysis of example 13. Applying Maxim 10 reveals a support relation between the third and fourth text segment.

role of the linguistic cue 'because'. Does it indicate a justification, a causal relationship or both? The indicator word 'and' only suggests that the clauses that are linked together by the 'and' are independent points.

Again, it might help to reformulate the presented text segment as reasons (see Maxim 10). The following reformulations continue to interpret the argumentation structure as a chain of reasoning. We could ask why these students struggle at elite universities ( $R_2$ ), and the text seems to answer by pointing to their lesser academic qualification ( $R_3$ ); we could ask why they are lesser qualified, and the text seems to answer by pointing to academic challenges minorities have to face prior to university ( $R_4$ ). In one of the suggested standard forms:

Students from racial minorities could struggle to meet the expectation of a rigorous curriculum at elite universities, *since* they may be less qualified than other students. ( $R_2$  *since*  $R_3$ )

Students from racial minorities may be less qualified than other students, *since* they tend to face academic challenges prior to university. ( $R_3$  *since*  $R_4$ )

Figure A.12 illustrates the resulting final interpretation of the argumentation structure. We started with a preliminary analysis by identifying the main claim and longer text segment, whose internal structure posed some questions (see Figure A.10). We then used the maxim to reformulate the main idea of reasons and objections. By making explicit what is presented as a justification and what is presented as being justified, we were able to analyze the argumentation structure deeper and identified four separate text segments that correspond to a chain of reasons ( $R_1 - R_4$ ).

In applying Maxim 10, you must be faithful to the text. That means that your reformulations of reasons do not deviate too much from the author's formulations and that the reformulation does not distort the author's intended meaning. Instead, your reformulations should stay as close as possible to the text. Otherwise, you might be unable to determine which text segments correspond to the reasons in your formulation. You should reassess your formulation if you cannot determine which text segments correspond to  $x$  and  $y$  in a reformulation of the form ' $y$  *since*  $x$ '. For instance, I formulated the attack between the

*Example 13:* But affirmative action is still desperately needed in America's best universities [...] Some well-intentioned people may argue that because of the academic challenges racial minorities tend to face, they may be less qualified than other students to attend elite universities and could struggle to meet the expectations of a rigorous curriculum. This is based in part on the debunked "mismatch" theory, which posits that these students would benefit more from attending less challenging schools more suited to their relative lack of educational attainment. (Gordon and Wang 2019)

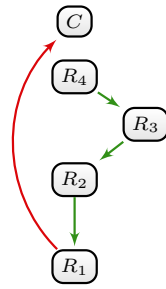


ABBILDUNG A.12

Final analysis of example 13. Applying Maxim 10 reveals a chain of support relations (from  $R_4$  to  $R_1$ ) between four text segments.

first reason and the main claim as follows:

Affirmative action is *not* needed *since* students from racial minorities would benefit from attending less challenging schools. ( $C$  is not true *since*  $R_1$ )

Now, it might make sense to insert another reason step  $R^*$  between  $C$  and  $R_1$  in the following way:

Affirmative action is *not* needed *since* it is supposed to help students from racial minorities, which it doesn't. ( $C$  is not true *since*  $R^*$ )

Affirmative action is supposed to help students from racial minorities, which it doesn't, *since* students from racial minorities would benefit from attending less challenging schools. ( $R^*$  *since*  $R_1$ )

Now, I have formulated the additional reason  $R^*$  that attacks  $C$  and is supported by  $R_1$ . However, there is no text segment to be found that corresponds to this reason. Faithfulness to the text demands the use of reformulations that make it possible—and preferably easy—to identify the text segments that correspond to reasons.

### A.1.6 IDENTIFYING THE TARGET OF SUPPORT AND ATTACK RELATIONS

In the last example, we used Maxim 10 to identify a chain of supporting reasons we did not initially spot. We saw that Maxim 10 can help to discern support relations between text segments, even if there are no obvious linguistic cues for these relations. Reformulating the basic idea of reasons in your own words can also help with other problems of identifying the argumentation structure in a text.

For instance, you might identify a text segment as a presented justification without being able to say for what it is a reason. The linguistic cues might univocally signpost the text segment as a reason but fail to clarify the exact target. Such a situation leaves the analyst to decide between different possibilities as to what is supposed to be justified by this text segment. Maxim 10 might help you to decide between these different possibilities. You should formulate the different candidate relations in some of the suggested standard forms and choose among them by applying the principle of charity and accuracy (see Maxim

10 on page 268). The main idea is to choose among these possibilities by comparing the different formulations according to their plausibility and how they capture the author's intended idea of the reasons.

*Example 14:* [...] But affirmative action is still desperately needed in America's best universities to combat the racial divisions that colorblindness does nothing to resolve. Here's why. [...] [...] Affirmative action [...] [is] about deciding which qualified individuals would benefit the most from a great education. And minority students stand to gain quite a bit. Attending a highly selective college (the kind that would probably use affirmative action) produces larger increases in income and employment rates for blacks and Hispanics than for whites. Researchers have also concluded that students are more likely to graduate when they attend the most selective college that will accept them, meaning that a minority student can still succeed and graduate without having a prep-school background or a perfect SAT score.

Granting more underprivileged students the opportunity to succeed through education not only could lift some individuals out of a vicious cycle of poverty, but might also lift entire minority groups by giving them more representation in elite institutions and positions of national leadership. (Gordon and Wang 2019)

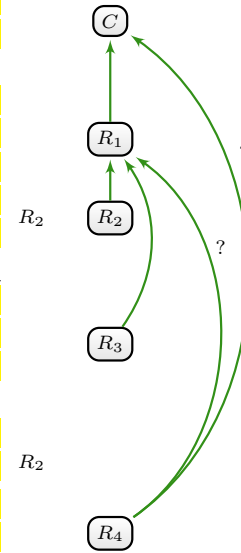


ABBILDUNG A.13

Illustrating the problem of identifying the target of a reason ( $R_4$ ).

The example in Figure A.13 continues the example from Figure A.8 and provides a concrete illustration of the problem. The main claim in favour of affirmative action is supported by one reason ( $R_1$ ), which argues that affirmative action will benefit qualified students from minority groups. The text moves on to provide a supporting reason for this claim and points to an increase in income and employment ( $R_2$ ) and the fact that these students are more likely to graduate at elite universities ( $R_3$ ). The last paragraph begins by reformulating  $R_1$  and goes on to add another point. The phrase ‘*not only . . . , but also . . .*’ designates this last point as a supporting reason. But it is less clear what exactly is being justified. Since the preceding reasons  $R_2$  and  $R_3$  support  $R_1$ , one might speculate whether  $R_4$  is another reason in support of  $R_1$ . However, since the used phrase refers to  $R_1$  in the first clause, which supports the main claim  $C$ , it is similarly possible that  $R_4$  is another reason for the main claim.

According to Maxim 10, we should first formulate both possibilities in a way that makes the reason and its target explicit (*explicitness*), which might demand reformulating some text segments.

1. *Support from  $R_4$  to  $R_1$ :* Affirmative action should help those qualified individuals who would benefit the most from an excellent education. Minority students stand to gain quite a bit *since* it might lift entire minority groups by giving them more representation in elite institutions and positions of national leadership.
2. *Support from  $R_4$  to  $C$ :* But affirmative action is still desperately needed in America's best universities *since* it might lift entire minority groups by giving them more

representation in elite institutions and positions of national leadership.

We now have to compare these different possibilities in terms of how strong or plausible the resulting support is (*charity*) and which formulations adhere best to the author's known or hypothesized intention (*accuracy*). From the viewpoint of plausibility, both possibilities seem to be fine: Since whole minority groups benefit from a better representation (first possibility), students that belong to these groups will benefit as well (second possibility). However, accuracy favours the second interpretation. The author's intended point is to justify affirmative action not only in terms of benefits for those individuals from minority groups that study at elite universities as a result of affirmative action but also in terms of how affirmative action will benefit other individuals from these minority groups. A point that is only captured in the second formulation.

While Maxim 10 will often help you identify the target of support and attack relations, some cases remain ambiguous. Different possible interpretations might be similarly charitable and accurate, or there might be a trade-off between charity and accuracy. In these cases, you can pick one of your interpretations to your liking.

However, keep in mind that the author might intend to justify two different things with the same text segment. In such a case, you should annotate all the relations.

Let me close this section with a general comment on the usefulness and scope of the ten maxims. Admittedly, these maxims are somewhat vague and imprecise. You might encounter cases where the maxims will not disambiguate between all possible interpretations, and different analysts might come up with different interpretations when using the maxims. These maxims, however, do not aspire to lead to unique and objective results. Instead, they are rules of thumb that at least narrow the degree of interpretation to some extent.

## A.1.7 FREQUENTLY ASKED QUESTIONS

### QUESTIONS CONCERNING MAIN CLAIMS

**1. WHAT ARE (MAIN) CLAIMS? HOW DO I IDENTIFY AND CATEGORIZE THEM?** There is no special category for claims—and reasons for that matter—besides them being justificatory relevant. We simply call text segments claims if they are not used as reasons for something else and if other text segments are presented as reasons for or against them. In other words, a claim is a text segment with only incoming and no outgoing attack or support relations. This is, in some sense, a technical definition since it differs from our ordinary language usage of the terminus '*claim*'. According to the normal understanding, someone can formulate a claim without presenting justifications for it. However, our usage of the terminus is a mere terminological convention, which has no further bearing on the analysis of argumentation structure.

**2. QUESTION CLAIMS: THE MAIN CLAIM IS IMPLICITLY FORMULATED AS A QUESTION. SHOULD I UNDERSTAND THE MAIN CLAIM AS A POSITIVE OR NEGATIVE ANSWER TO THE QUESTION?** This is an important question because the answer will dictate whether reasons are reasons *for* or reasons *against* the claim.

Consider the example in Figure A.14. The argumentation is about germline editing, and



*Example 15:* If CRISPR can help parents conceive a disease-free child when no other options exist and it can do so safely, ought we to use it? It's a question I've asked myself again and again [...]

For many people, the very idea [of germline editing] feels unnatural and wrong [...] Humans have been reproducing for millennia aided only by the DNA mutations that arise naturally, and for us to begin directing the process—similar to the way that plant biologists might genetically modify corn—seems almost perverse at first glance. (Doudna and Sternberg 2017)

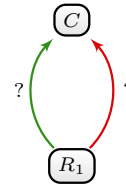


ABBILDUNG A.14

Example of a main claim implicitly formulated as a question.

the main question is formulated in the first sentence. In the second paragraph, we find an argument against germline editing, which involves the notion of naturalness. We can interpret this reason in two ways: Either as a reason *against* a claim that corresponds to an affirmative answer to the main question or as a reason *in favour of* a claim that corresponds to a negative answer to the main question. In the first case, we would categorize the relation between the reason and the main claim as an attack, and in the second case, as support.

There is no rule to decide on this issue, and either possibility is fine. You can choose the main claim according to what feels natural to you. For me, the second paragraph feels more like an argumentation against something. Therefore, I tend to understand the main claim as the positive answer to the first question and would categorize the second paragraph as formulating a reason that attacks that claim.

The only important thing is that you are consistent with that decision. Once you decide to interpret the claim as an affirmative answer or a negative answer, this decision should be regarded as fixed for the analysis of whether other text segments are reasons for or against the claim.

**3. MULTIPLE CLAIMS: CAN THERE BE MORE THAN ONE CLAIM IN A TEXT?** Yes, that can happen. Sometimes a debate is about two or more questions, which are usually related to each other in some way. Answers to these questions make up for different claims. For instance, the debate about germline editing is about at least two different issues: The question of using this technology for therapeutic uses and the question of further research into this technology by using embryos. Different positions will formulate different answers to these questions, which might represent different claims in the debate about germline editing.

**4. CONDITIONAL CLAIMS: CAN MAIN CLAIMS BE FORMULATED WITH CONDITIONS?** Yes. It is quite common to formulate claims conditionally to, for instance, express exceptions or qualify claims differently. Consider the following formulated claim:

*Example 16:* Others argue that genome editing, once proved safe and effective, should be allowed to cure genetic disease (and indeed, that it is a moral imperative). (NHGRI 2017)

The anonymous proponents do not think that germline editing should be allowed unconditionally for therapeutic uses. Instead, they formulate specific sufficient conditions under



which germline editing should be allowed as a therapy.

**5. IMPLICIT CLAIMS: IT IS MORE OR LESS CLEAR WHAT THE MAIN CLAIM IS, BUT NO TEXT SEGMENT FORMULATES THE CLAIM. HOW SHOULD I ANNOTATE AND CATEGORIZE REASONS THAT SUPPORT AND ATTACK THE MAIN CLAIM?** Such cases represent a widespread phenomenon. Often, the context is clear enough that an author does not think it necessary to explicitly formulate their main claim. In the example in Figure A.15, the author provides a reason for why the use of CRISPR as a therapy would constitute a violation of informed consent without explicitly saying that this violation is a reason against the use of germline editing for therapeutic uses ( $C_1$ ). However, the cotext—which is omitted here—makes it quite clear that she presents it as a reason against germline editing since she moves on to bring forward other reasons against germline editing.

*Example 17:*

$C_1$

$C_2$

$C_3$

When it [CRISPR] is used to edit the germline, the affected unborn fetus cannot give consent to take part in the experimental procedure or manage any complications or undesirable mutations that might arise as a result of the edited germline. Therefore CRISPR violates the fundamental tenet of medicine of providing informed consent. (Jain 2021)

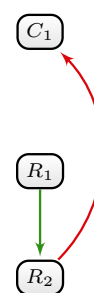


ABBILDUNG A.15  
Example of dealing with implicit claims.

If the main claim lacks a corresponding text segment, text segments presented as attacking or supporting this claim will lack a text segment as a target. In other words, attack and support relations will not point to a specific text segment. So, where should they point to? This is, fortunately, only a technical problem. *On the practical level* you will be provided with labels  $C_1$ ,  $C_2$ , ... in the text that act as placeholders for implicit claims. If there are implicit claims in the text, simply choose for every implicit claim one of these labels and annotate them as if they were the text segments formulating the claims—as indicated in Figure A.15.

## QUESTIONS CONCERNING THE NOTION OF *Justificatory Relevance*

**6. CAN A TEXT SEGMENT BE JUSTIFICATORY RELEVANT WITHOUT HAVING INCOMING OR OUTGOING SUPPORT OR ATTACK RELATIONS?** No. According to the used *definition of justificatory relevance* (see page 250), a text segment can only be justificatory relevant if it is presented as a justification for something or if some other text segment is presented as a reason for or against it. In the former case, the text segment has at least one outgoing relation and at least one ingoing relation in the latter. (However, note that in the case of implicit claims, you will encounter text segments with outgoing relations that do not point to other text segments. See the question about implicit claims on page 275.)

**7. SHOULD I ANNOTATE LINGUISTIC CUES?** Linguistic cues should usually not be annotated as justificatory relevant. In other words, we do not consider them as belonging to the formulation of reasons and claims. One exception: Linguistic cues that would tear an otherwise contiguous text segment into different non-contiguous text segments can be considered part of the reason or claim. In the example in Figure A.16, the phrase ‘*not only ... , but also ...*’ is a linguistic cue that signposts the formulation of two different reasons ( $R_1$  and  $R_2$ ) for the main claim  $C$ . However, not annotating the ‘*not only*’ as justificatory relevant would mean representing the first reason by two non-contiguous text segments.

*Example 18:* But affirmative action is still desperately needed in America’s best universities to combat the racial divisions that colorblindness does nothing to resolve. Here’s why. [...] Granting more underprivileged students the opportunity to succeed through education not only could lift some individuals out of a vicious cycle of poverty, but might also lift entire minority groups by giving them more representation in elite institutions and positions of national leadership. (Gordon and Wang 2019)

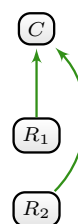


ABBILDUNG A.16

Exception to the rule of not annotating linguistic cues as justificatory relevant.

But even if we usually do not consider linguistic cues as belonging to the formulation of reasons, you should explicitly search for linguistic cues. Annotating them in some kind of way might be very helpful to keep an overview of them. For instance, linguistic cues are annotated in the examples used here by underlining them.

## QUESTIONS CONCERNING THE INDIVIDUATION OF REASONS AND CLAIMS.

**8. CAN A REASON OR CLAIM EXTEND OVER A LONG TEXT SEGMENT—LET’S SAY, MORE THAN ONE PARAGRAPH IN THE TEXT?** Yes, that can happen. You should, however, also ask yourself if everything the author writes is justificatory relevant. Mere contextual information (see Maxim 7 on page 264) or illustrative examples (see Maxim 5 on page 259) should not be annotated. Keep also in mind that claims and reasons do not necessarily correspond to exactly *one* text segment but can be scattered over many non-contiguous text segments (see the next question).

**9. HOW SHOULD I ANNOTATE A TEXT IF A REASON OR CLAIM SEEMS TO BE SCATTERED OVER NON-CONTIGUOUS TEXT SEGMENTS?** One reason or one claim does not necessarily correspond to exactly *one* text segment but can be scattered over many non-contiguous text segments. In this case, simply annotate these text segments as belonging to the same reason or claim. Maxim 6–8 (see on page 261) will help you to identify text segments that belong to the same reason. In short, they tell you that sentences and subclauses that are intended to be taken together to understand the reason’s main idea or its relevance are part of the reason (see Maxim 6 on page 262). Everything else is not.

- *On the practical level* you will be provided with **different labels** ( $R1$ ,  $R2$ , ...,  $C1$ ,  $C2$ , ...) that can be understood as names for reasons and claims. If a reason corresponds to many non-contiguous text segments, you simply have to use the same label.

**10. HOW DO I ANNOTATE A TEXT SEGMENT THAT BELONGS TO DIFFERENT REASONS?** First of all, that can happen. For instance, the same text segment can be important to understand the main idea for two different reasons. According to Maxim 6 (see on page 262), this text segment would belong to both reasons. However, one text segment can only be annotated to belong to one reason. In this case, simply do not annotate the text segment as justificatory relevant.

**11. WHAT IF LINGUISTIC CUES CONFLICT WITH THE MAXIMS ABOUT REASON INDIVIDUATION?** For instance, how do I annotate text segments that belong according to the given linguistic cues to one reason but that constitute separate reasons according to Maxim 9 (see page 267)? *The short answer:* The maxims for the individuation of reasons (Maxims 6–10) should always be prioritized over linguistic cues for the individuation of reasons.

A simple example is provided in Figure A.17. The linguistic cues ‘*the objection*’ and ‘*this argument*’ suggest that the authors think of the second paragraph as one objection. However, we can understand the paragraph as formulating two relevant points, which provide independent support for the main claim being false. The argument from naturalness is grounded in the idea that the natural is somehow better or preferable to the unnatural. The other point against the main claim seems to rely on the notion of hybris—the idea that we, as mere humans, should not claim for ourselves to act as gods. These are, admittedly, speculations of what is meant with both points. However, on the presented understanding, both points are independent. According to Maxim 9 (see page 267), we should, therefore, analyze this paragraph as presenting two objections.

*Example 19:* [...] in theory—and eventually in practice—CRISPR could be used to modify disease-causing genes in embryos brought to term, removing the faulty script from the genetic code of that person’s future descendants as well. Proponents of such “human germline editing” argue that[...] Let’s start with the objection that embryo modification is unnatural, or amounts to playing God. This argument rests on the premise that [...] But [...] (Harris and Darnovsky 2018)

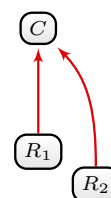


ABBILDUNG A.17

Conflict between linguistic cues and maxims for the individuation of reasons. The linguistic cues suggest that there is one objection; according to Maxim 9, there are two objections.

*Some background:* Maxims 6–10 clarify what we understand as one reason and what we understand as different reasons in the context of analyzing argumentation structure. In other words, these maxims spell out how we individuate and count reasons. The suggested understanding can, of course, conflict with what the author understands as one reason instead of many. Hence, the author may intend one particular text segment to express one reason, whereas Maxims 6–10 suggest considering it as many reasons—or the other way around. In such conflicting cases, you should always individuate reasons according to the given maxims. This rule is important because we might want to compare the number of reasons in different texts, and the result of such a comparison should be grounded in some text-independent criterion to count arguments. In other words, we should count

reasons according to clarified and transparent criteria and not necessarily by the subjective understanding of how the author would count their reasons.

### QUESTIONS CONCERNING ATTACK AND SUPPORT RELATIONS

**12. CAN A REASON BE A REASON FOR DIFFERENT THINGS? IN OTHER WORDS, CAN A TEXT SEGMENT HAVE MORE THAN ONE OUTGOING SUPPORT OR ATTACK RELATION?** Such a case can happen. A text segment can be presented as a reason for different text segments. In that case, it would have more than one outgoing support relation. The same applies to multiple outgoing attack relations. It might even happen that one text segment is intended to be both a reason for and against something. However, intended multiple outgoing relations are the exception. You should, therefore, always apply the maxims, especially Maxim 10 (see page 268), to check if there are multiple outgoing relations (see also the next question).

**13. TARGET AMBIGUITY: WHAT IF I AM UNSURE WHAT IS BEING SUPPORTED OR OBJECTED TO BY A REASON?** What should I do if it is not clear what is being justified? In other words, how should I choose the targets of outgoing support and attack relations if there are different possible interpretations?

If you are not sure what other text segment is supposed to be supported or attacked by a text segment, you should, first, use Maxim 10 (see page 268), which might help to decide between different interpretations of what the target is.

This maxim demands that you reformulate the possible attack or support relations in your own words and choose among them by using the principles of charity and accuracy. According to these principles, you should choose the formulation that maximizes argumentative strength (*charity*) and is at the same time faithful to the author's intentions of understanding the presented reasons (*accuracy*).

However, Maxim 10 will not always help to disambiguate between different interpretations of what is supposed to be justified by a text segment. If you are still unsure of what the best interpretation is, you can simply finalize the categorization with different support and attack relations. In other words, consider all remaining plausible interpretations of what the target is as a target.

**14. MULTIPLE RELATIONS: CAN TWO ARGUMENTATIVE COMPONENTS BE CONNECTED BY DIFFERENT JUSTIFICATORY RELATIONS?** Usually, there will be at most one justificatory relation between two argumentative units. There are two possible situations that can be represented by connecting two argumentative units with two justificatory relations.

First, an author might express that **two statements are contradictory to each other**. For instance, the author might suggest that two reasons are inconsistent with each other or that two claims cannot be true at the same time. In such cases, you can connect the two argumentative units with attack relations in both directions.

Second, an author might express that a **justification is a question-begging**, that it presupposes what it is supposed to justify or that a justification is circular. Support relations in both directions might represent these cases.

It is, however, conceivable that two argumentative units are connected in additional ways by combining different justificatory relations. For instance, it is conceivable that a support and an attack relation at the same time connect two argumentative units. However, such combinations will be flawed from an evaluative point of view. What can it mean, for instance, that something is a reason for *and* an objection to the same claim? It is doubtful that arguers are prone to such fundamental argumentative flaws. Accordingly, you should only connect two argumentative components in this way if the linguistic cues point unambiguously in this direction. If there are no such linguistic cues, the principle of charity recommends dispensing with such interpretations.

## A.2 MATHEMATICAL PROOFS

### A.2.1 PURE SIGNIFICANCE TESTING

Let  $\Theta_0 \subset \Theta$  be a null hypothesis represented by a subset of the parameter space  $\Theta$  of a statistical model. If  $\Theta_0$  is a singleton ( $\Theta_0 = \{\theta\}$  with  $\theta \in \Theta$ ), we say that  $\Theta_0$  is a simple hypothesis. Otherwise, we will say that  $\Theta_0$  is a complex hypothesis.

Further let  $\Omega$  be a finite sample space. The basic idea of pure significance testing is to base all relevant notions of a ranking over the sample space  $\Omega$  (Schervish 1995, 216). This ranking compares outcome  $\omega \in \Omega$  in terms of their consistency with the null hypothesis. Let  $\leq$  therefore a weak order on the sample space  $\Omega$ .

We now introduce some further definitions and results from the introduced notions.<sup>283</sup>

**Definition A.2.1** Let  $\Omega^e(\omega) := \{\omega \in \Omega | \omega \leq \nu\}$  the set of events that are at least as extreme as  $\omega$ .

**Lemma A.2.2**  $P(\Omega^e(\nu)) \leq P(\Omega^e(\omega))$  iff  $\omega \leq \nu$ .

*Proof:*

1. From A.2.1:  $\Omega^e(\nu) \subset \Omega^e(\omega)$  iff  $\omega \leq \nu$ .
2. From (1) and the fact that  $P$  is a probability:  $P(\Omega^e(\nu)) \leq P(\Omega^e(\omega))$  iff  $\omega \leq \nu$ .  $\square$

The following definitions introduce a rejection rule for simple hypotheses.

**Definition A.2.3 (critical region)** Let  $\mathcal{C}_\theta := \{\omega \in \Omega | P(\Omega^e(\omega); \theta) \leq \alpha\}$  (with  $\theta \in \Theta$ ) the set of events that lead to a rejection of a simple hypothesis  $\theta$ .

**Definition A.2.4 (p-value)**  $p(\omega; \theta) := P(\Omega^e(\omega); \theta)$

**Lemma A.2.5** There exists an  $\omega \in \mathcal{C}_\theta$  with  $\mathcal{C}_\theta = \Omega^e(\omega^*)$

*Proof:*

1. From A.2.2 it follows that: If  $P(\Omega^e(\omega); \theta) \leq \alpha$  then  $\forall \omega \leq \nu : P(\Omega^e(\nu); \theta) \leq \alpha$ .
2. From (1) and using A.2.3: If  $\omega \in \mathcal{C}_\theta$ , then  $\forall \nu \geq \omega : \nu \in \mathcal{C}_\theta$ .

<sup>283</sup>The exposition does not follow any specific textbook and is tailored to the description in Chapter 5 to provide a general formulation of using significance testing based on finite outcome spaces. However, the definitions and proofs are basic and can be found for similar contexts in introductory textbooks (e.g., DeGroot and Schervish (2012), Cox, Hinkley, and Hinkley (1974) and Casella and Berger (2002)).

3. Since  $\mathcal{C}_\theta$  is finite, there is a max element w.r.t. the ranking:  $\exists \omega^* \in \mathcal{C}_\theta : \forall \nu \in \mathcal{C}_\theta : \omega^* \leq \nu$
4. From (2) und (3):  $\forall \nu \notin \mathcal{C}_\theta : \nu < \omega^*$
5. From (3) and (4):  $\exists \omega^* \in \mathcal{C}_\theta : \mathcal{C}_\theta = \bigcup \{\nu \in \Omega | \omega^* \leq \nu\}$
6. From (5) and A.2.1:  $\exists \omega^* \in \mathcal{C}_\theta : \mathcal{C}_\theta = \Omega^e(\omega^*)$ .  $\square$

The following lemma shows that the suggested construction of the critical region (A.2.3) guarantees that the chosen significance level  $\alpha$  is an upper bound for type-I-error probabilities.

**Lemma A.2.6**  $P(\mathcal{C}_\theta; \theta) \leq \alpha$ .

*Proof:*

1. From A.2.3:  $P(\Omega^e(\omega); \theta) \leq \alpha$  for all  $\omega \in \mathcal{C}_\theta$ .
2. From (1) and A.2.5:  $P(\mathcal{C}_\theta; \theta) \leq \alpha$ .  $\square$

The following definitions introduce a rejection rule for complex hypotheses.

**Definition A.2.7 (p-value)**  $p_{\Theta_0}(\omega) := \sup\{p(\omega; \theta) | \theta \in \Theta_0\}$

**Definition A.2.8 (critical region)**  $\mathcal{C}_{\Theta_0} := \{\omega \in \Omega | p_{\Theta_0}(\omega) \leq \alpha\}$

**Lemma A.2.9**  $\mathcal{C}_{\Theta_0} \subseteq \mathcal{C}_\theta$  for all  $\theta \in \Theta_0$ .

*Proof:* Shown by reductio ad absurdum.

1. Assume that there exists a  $\theta \in \Theta_0$  with  $\mathcal{C}_{\Theta_0} \not\subseteq \mathcal{C}_\theta$ .
2. From (1):  $\exists \omega \in \Omega : \omega \in \mathcal{C}_{\Theta_0}$  and  $\omega \notin \mathcal{C}_\theta$
3. From (2), A.2.3 and A.2.8:  $\exists \omega \in \Omega : p_{\Theta_0}(\omega) \leq \alpha$  and  $p(\omega; \theta) > \alpha$ .
4. From (3) and A.2.7:  $\exists \omega \in \Omega : \sup_{\theta \in \Theta_0} \{p(\omega; \theta)\} \leq \alpha$  and  $p(\omega; \theta) > \alpha$ , which cannot be true.  $\square$

The following lemma shows that the suggested generalisation of p-values (A.2.7) and of the critical region (A.2.8) will guarantee that  $\alpha$  is an upper bound of type-I-error probabilities.

**Lemma A.2.10**  $P(\mathcal{C}_{\Theta_0}; \theta) \leq \alpha$  for all  $\theta \in \Theta_0$ .

*Proof:*

1. From A.2.9:  $P(\mathcal{C}_{\Theta_0}; \theta) \leq P(\mathcal{C}_\theta; \theta)$  for all  $\theta \in \Theta_0$ .
2. From (1), transitivity and A.2.6:  $P(\mathcal{C}_{\Theta_0}) \leq \alpha$  for all  $\theta \in \Theta_0$ .  $\square$

## A.2.2 CONSTRAINTS ON THE PARAMETER SPACE

Let  $\Omega_\tau$  an outcome space and  $\Pi$  a parameter space as defined at the end of Section 5.3.1. Let further  $R$  be a random variable  $R : \Omega_\tau \rightarrow \mathbb{R}$  and  $h_0$  a complex null hypothesis of the following form:

$$h_0 : \sum_{\tau \in \Omega_\tau} P(\tau; \pi) R(\tau) = c \quad (\text{A.1})$$

We now can understand this expression as a constraint on the parameter space that eliminates on of the free parameters  $\pi_{ij}$ . The idea is to rewrite the last expression as a constraint

for one of the subspaces in  $\Omega_\tau$ , say  $\Omega_t$ , and use normalisation to get rid of one of the  $(\pi_t)_j$ . Let's rewrite the sum:

$$\begin{aligned} \sum_{\tau \in \Omega_\tau} P(\tau; \pi) R(\tau) &= \sum_{a_{ti} \in \Omega_t} \sum_{\tau \in \Omega_\tau \setminus \{a_{ti}\}} P(\tau; \pi) R(\tau) \\ &= \sum_{a_{ti} \in \Omega_t} \sum_{\tau \in \Omega_\tau \setminus \{a_{ti}\}} P(a_{ti}) \Pi_{j \neq t}^{n_\tau} P(e_j^\tau) R(\tau) \\ &= \sum_{i=1}^{n_t} \pi_{ti} \sum_{\tau \in \Omega_\tau \setminus \{a_{ti}\}} \Pi_{j \neq t}^{n_\tau} P(e_j^\tau) R(\tau) \end{aligned}$$

Now let  $a_i(t) := \sum_{\tau \in \Omega_\tau \setminus \{a_{ti}\}} \Pi_{j \neq t}^{n_\tau} P(e_j^\tau) R(\tau)$  (or, in short,  $a_i$ ). Then, we can rewrite the constraint as:

$$c = \sum_{i=1}^{n_t} \pi_{ti} a_i \quad (\text{A.2})$$

We now use

$$\pi_{tj} = 1 - \sum_{i \neq j}^{n_t} \pi_{ti} \quad (\text{A.3})$$

for one fixed  $j$ , which leads to:

$$1 = \sum_{i \neq j}^{n_t} \frac{a_i - a_j}{c - a_j} \pi_{ti} \quad (\text{A.4})$$

The constraint has now the form

$$1 = \sum_{i \neq j}^{n_t} a_i^* \pi_{ti} \quad (\text{A.5})$$

and can be used to generate the independent free parameters in  $\pi$ . The remaining dependent parameter  $\pi_{tj}$  can then be calculated by using A.3.





## LITERATURE

- Ajjour, Yamen, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. "Unit Segmentation of Argumentative Texts." In *Proceedings of the 4th Workshop on Argument Mining*, 118–28. Copenhagen, Denmark: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-5115>.
- Amgoud, L., C. Cayrol, M. C. Lagasquie-Schiex, and P. Livet. 2008. "On Bipolarity in Argumentation Frameworks." *International Journal of Intelligent Systems* 23 (10): 1062–93. <https://doi.org/10.1002/int.20307>.
- Artstein, Ron, and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–96. <https://doi.org/10.1162/coli.07-034-R2>.
- Austin, John Langshaw. 1962. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*. Oxford: Clarendon Press.
- Baroni, Pietro, Martin Caminada, and Massimiliano Giacomin. 2011. "An Introduction to Argumentation Semantics." *Knowledge Engineering Review* 26 (4): 365.
- Bentahar, Jamal, Bernard Moulin, and Micheline Bélanger. 2010. "A Taxonomy of Argumentation Models Used for Knowledge Representation." *Artificial Intelligence Review* 33 (3): 211–59. <https://doi.org/10.1007/s10462-010-9154-1>.
- Berelson, Bernard. 1952. *Content Analysis in Communication Research*. Facs. of 1952 ed. New York: Hafner.
- Berelson, Bernard, and Paul Felix Lazarsfeld. 1948. *The Analysis of Communication Content*. University of Chicago Press.
- Betz, Gregor. 2010. *Theorie dialektischer Strukturen*. Frankfurt am Main: Klostermann.
- . 2016. "Accounting for Possibilities in Decision Making." In *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty.*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 135–71. Cham: Springer.
- Betz, Gregor, and Georg Brun. 2016. "Analysing Practical Argumentation." In *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty.*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 39–77. Cham: Springer.
- Betz, Gregor, and Sebastian Cacean. 2012. *Ethical Aspects of Climate Engineering*. Karlsruhe: KIT Scientific Publishing.
- Blair, J. Anthony. 2012. "Relevance, Acceptability and Sufficiency Today." In *Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair*, edited by J. Anthony Blair and Christopher W. Tindale, 87–100. Argumentation Library. Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-007-2363-4\\_8](https://doi.org/10.1007/978-94-007-2363-4_8).
- Blair, J. Anthony. 2015. "What Is Informal Logic?" In *Reflections on Theoretical Issues in Argumentation Theory*, edited by Frans H. van Eemeren and Bart Garssen, 27–

42. Argumentation Library. Heidelberg: Springer International Publishing. [https://doi.org/10.1007/978-3-319-21103-9\\_2](https://doi.org/10.1007/978-3-319-21103-9_2).
- Boella, Guido, Dov M. Gabbay, Leon van der Torre, and Serena Villata. 2010. "Support in Abstract Argumentation." *Proceedings of the Third International Conference on Computational Models of Argument (COMMA'10)*. <https://doi.org/10.3233/978-1-60750-619-5-111>.
- Botting, David. 2016. "The Logical Evaluation of Arguments." *Argumentation* 30 (2): 167–80. <https://doi.org/10.1007/s10503-015-9383-1>.
- Bowell, Tracy, and Gary Kemp. 2010. *Critical Thinking: A Concise Guide*. 3rd ed. New York: Routledge.
- Brun, Georg. 2003. *Die Richtige Formel: Philosophische Probleme Der Logischen Formalisierung*. Fouque London Publishing.
- . 2014. "Reconstructing Arguments. Formalization and Reflective Equilibrium." *Logical Analysis and History of Philosophy* 17: 94–129.
- . 2016. "Textstrukturanalyse und Argumentrekonstruktion." In *Neues Handbuch des Philosophie-Unterrichts*, edited by Jonas Pfister and Peter Zimmermann, 1st ed., 247–74. Stuttgart: UTB.
- . 2023. "Logical Forms: Validity and Variety of Formalizations." *Logic and Logical Philosophy* 32 (3): 341–61. <https://doi.org/10.12775/LLP.2023.016>.
- Brun, Georg, and Gertrude Hirsch Hadorn. 2009. *Textanalyse in Den Wissenschaften: Inhalte Und Argumente Analysieren Und Verstehen*. 1. Aufl. Stuttgart: UTB.
- Bucher, Theodor. [1987] 2019. *Einführung in die angewandte Logik*. Berlin Boston: De Gruyter.
- Burke, Michael B. 1985. "Unstated Premises." *Informal Logic* 7 (2).
- Cacean, Sebastian. 2012. "Ethische Aspekte von Cognitive Enhancement." In *Sport, Doping und Enhancement Ergebnisse und Denkanstöße*, edited by Giseler Spitzer and Elk Franke, 151–220. Köln: Sportverlag Strauß.
- . 2020. "Reliability of Argument Mapping." In *Proceedings of the 3rd European Conference on Argumentation*, edited by Catarina Dutilh Novaes, Henrike Jansen, Jan Albert Van Laar, and Bart Verheij, 101–24. College Publications.
- Carpini, Michael X. Delli, Fay Lomax Cook, and Lawrence R. Jacobs. 2004. "Public Deliberation, Discursive Participation, and Citizen Engagement: A Review of the Empirical Literature." *Annual Review of Political Science* 7 (1): 315–44. <https://doi.org/10.1146/annurev.polisci.7.121003.091630>.
- Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2. ed. Pacific Grove, Calif.: Duxbury.
- Cayrol, Claudette, and Marie-Christine Lagasquie-Schiex. 2013. "Bipolarity in Argumentation Graphs: Towards a Better Understanding." *International Journal of Approximate Reasoning*, Special issue: Uncertainty in Artificial Intelligence and Databases, 54 (7): 876–99. <https://doi.org/10.1016/j.ijar.2013.03.001>.
- Chambers, Simone. 2003. "Deliberative Democratic Theory." *Annual Review of Political Science* 6 (1): 307–26. <https://doi.org/10.1146/annurev.polisci.6.121901.085538>.
- Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. <https://doi.org/10.1177/001316446002000104>.
- Collins, Harry. 1985. *Changing Order Replication and Induction in Scientific Practice*.

- London [u.a.]: Sage Publ.
- Copi, Irving M., and Carl Cohen. 1990. *Introduction to Logic*. 8. ed. New York: Macmillan.
- Copi, Irving M., Carl Cohen, and Kenneth McMahon. 2014. *Introduction to Logic*. 14th edition. Upper Saddle River, NJ: Pearson Educacion.
- Cox, David R., David V. ; 123994128 Hinkley, and David Victor Hinkley. 1974. *Theoretical Statistics*. 1. publ. London: Chapman and Hall.
- de Courtenay, Nadine, and Fabien Grégis. 2017. "The Evaluation of Measurement Uncertainties and Its Epistemological Ramifications." *Studies in History and Philosophy of Science Part A, The Making of Measurement*, 65–66 (October): 21–32. <https://doi.org/10.1016/j.shpsa.2017.05.003>.
- DeGroot, Morris H, and Mark J Schervish. 2012. *Probability and Statistics*. 4th ed. Pearson Education.
- Demey, Lorenz, Barteld Kooi, and Joshua Sack. 2019. "Logic and Probability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Stanford University: Metaphysics Research Lab.
- Dickson, Michael, and Davis Baird. 2011. "Significance Testing." In *Philosophy of Statistics*, edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster, 7:199–229. Handbook of the Philosophy of Science. Oxford: Elsevier. <https://doi.org/10.1016/B978-0-444-51862-0.50006-X>.
- Dijk, Teun A. van. 1997. "What Is Political Discourse Analysis?" *Belgian Journal of Linguistics* 11 (1): 11–52. <https://doi.org/10.1075/bjl.11.03dij>.
- Doudna, Jennifer A., and Samuel H. Sternberg. 2017. "Opinion: Should We Use Gene Editing to Produce Disease-Free Babies? A Scientist Who Helped Discover CRISPR Weighs In." *Ideas.ted.com*.
- Drummond, Chris. 2009. "Replicability Is Not Reproducibility: Nor Is It Good Science." In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26-Th ICML*.
- Dung, Phan Minh. 1995. "On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games." *Artificial Intelligence* 77 (2): 321–57. [https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X).
- Duthie, Rory, John Lawrence, Katarzyna Budzynska, and Chris Reed. 2016. "The CASS Technique for Evaluating the Performance of Argument Mining." In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 40–49.
- Eckle-Kohler, Judith, Roland Kluge, and Iryna Gurevych. 2015. "On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2236–42.
- Edgeworth, F. Y. 1888. "The Statistics of Examinations." *Journal of the Royal Statistical Society* 51 (3): 599–635. <https://www.jstor.org/stable/2339898>.
- Eemeren, Frans H. van, ed. 2001. *Crucial Concepts in Argumentation Theory*. Amsterdam: Amsterdam University Press.
- Eemeren, Frans H. van, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Dordrecht: Springer.
- Eemeren, Frans H. van, and Rob Grootendorst. 1984. *Speech Acts in Argumentative*

- Discussions a Theoretical Model for the Analysis of Discussions Directed Towards Solving Conflicts of Opinion*. Dordrecht: Foris Publications.
- . 1992. *Argumentation, Communication, and Fallacies*. 1. publ. Hillsdale, NJ: Erlbaum.
- . 2004. *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach*. 1. publ. Cambridge: Cambridge University Press.
- Eemeren, Frans H. van, Rob Grootendorst, and A Francisca Sn Henkemans. 2002. *Argumentation: Analysis, Evaluation, Presentation*. London: Routledge.
- Eemeren, Frans H. van, and A Francisca Sn Henkemans. 2016. *Argumentation: Analysis and Evaluation*. New York: Routledge.
- Ennis, Robert H. 1982. "Identifying Implicit Assumptions." *Synthese* 51 (1): 61–86. <https://doi.org/10.1007/BF00413849>.
- Fairclough, Isabela, and Norman Fairclough. 2012. *Political Discourse Analysis: A Method for Advanced Students*. London: Routledge.
- Fairclough, Norman. 2012. "Critical Discourse Analysis." In *The Routledge Handbook of Discourse Analysis*, edited by James Paul Gee and Michael Handford, 9–20. London: Routledge.
- . 2013. *Critical Discourse Analysis: The Critical Study of Language*. London: Routledge.
- Feldman, Richard. 2014. *Reason and Argument*. 2nd ed. London: Pearson.
- Fidler, Fiona, and John Wilcox. 2018. "Reproducibility of Scientific Results." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University.
- Fisher, Alec. 2004. *The Logic of Real Arguments*. 2nd ed. New York: Cambridge University Press.
- Fournier, Chris, and Diana Inkpen. 2012. "Segmentation Similarity and Agreement." In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 152–61. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Freeman, James B. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. Berlin: Foris Publications.
- . 2001. "Argument Structure and Disciplinary Perspective." *Argumentation* 15 (4): 397–423. <https://doi.org/10.1023/A:1012022330148>.
- . 2011. *Argument Structure: Representation and Theory*. 2011th ed. Dordrecht: Springer.
- Friberg-Fernros, Henrik, and Johan Karlsson Schaffer. 2017. "Assessing the Epistemic Quality of Democratic Decision-Making in Terms of Adequate Support for Conclusions." *Social Epistemology* 31 (3): 251–65. <https://doi.org/10.1080/02691728.2017.1317866>.
- Früh, Werner. 2017. *Inhaltsanalyse: Theorie und Praxis*. 9th ed. Konstanz: UTB GmbH.
- Garssen, Bart. 2001. "Argument Schemes." In *Crucial Concepts in Argumentation Theory*, edited by Frans H van Eemeren, 81–100. Amsterdam: Amsterdam University Press.
- Gee, James Paul. 2011. *An Introduction to Discourse Analysis: Theory and Method*. 3rd ed. Routledge.
- Gee, James Paul, and Michael Handford. 2012. "Introduction." In *The Routledge Handbook of Discourse Analysis*, 1–6. London: Routledge.

- Gerritsen, Susanne. 2001. "Unexpressed Premises." In *Crucial Concepts in Argumentation Theory*, edited by Frans H van Eemeren, 51–80. Amsterdam: Amsterdam University Press.
- Godden, David M. 2005. "Deductivism as an Interpretive Strategy: A Reply to Groarke's Recent Defense of Reconstructive Deductivism." *Argumentation and Advocacy* 41 (3): 168–83. <https://doi.org/10.1080/00028533.2005.11821627>.
- Godden, David, and Frank Zenker. 2018. "A Probabilistic Analysis of Argument Cogency." *Synthese* 195 (4): 1715–40. <https://doi.org/10.1007/s11229-016-1299-2>.
- Goddu, G. C. 2003. "Against the 'Ordinary Summing' Test for Convergence." *Informal Logic* 23 (3): 215–36.
- Gordon, Jarom, and Miles Wang. 2019. "Why We Still Need Affirmative Action." *THE MUSE*.
- Gordon, Thomas F. 2007. "Visualizing Carneades Argument Graphs." *Law, Probability and Risk* 6 (1-4): 109–17. <https://doi.org/10.1093/lpr/mgm026>.
- Gordon, Thomas F., Henry Prakken, and Douglas Walton. 2007. "The Carneades Model of Argument and Burden of Proof." *Artificial Intelligence, Argumentation in Artificial Intelligence*, 171 (10): 875–96. <https://doi.org/10.1016/j.artint.2007.04.010>.
- Gordon, Thomas F., and Douglas Walton. 2009. "Proof Burdens and Standards." In *Argumentation in Artificial Intelligence*, edited by Guillermo Simari and Iyad Rahwan, 239–58. Boston, MA: Springer US. [https://doi.org/10.1007/978-0-387-98197-0\\_12](https://doi.org/10.1007/978-0-387-98197-0_12).
- Gough, James, and Christopher Tindale. 1985. "'Hidden' or 'Missing' Premises." *Informal Logic* 7 (2): 99.
- Govier, Trudy. [1987] 2018. *Problems in Argument Analysis and Evaluation*. 2nd ed. Vol. 6. Windsor Studies in Argumentation. University of Windsor.
- . 1992a. *A Practical Study of Argument, Enhanced Edition*. 3rd ed. Wadsworth: Cengage Learning.
- . 1992b. "What Is a Good Argument?" *Metaphilosophy* 23 (4): 393–409. <https://www.jstor.org/stable/24438930>.
- . 2013. *A Practical Study of Argument, Enhanced Edition*. 7th ed. Wadsworth: Cengage Learning.
- Gray, Judy H., and Iain L. Densten. 1998. "Integrating Quantitative and Qualitative Analysis Using Latent and Manifest Variables." *Quality and Quantity* 32 (4): 419–31. <https://doi.org/10.1023/A:1004357719066>.
- Greckhamer, Thomas, and Sebnem Cilesiz. 2014. "Rigor, Transparency, Evidence, and Representation in Discourse Analysis: Challenges and Recommendations." *International Journal of Qualitative Methods* 13 (1): 422–43. <https://doi.org/10.1177/160940691401300123>.
- Grégis, Fabien. 2015. "Can We Dispense with the Notion of 'True Value' in Metrology?" In *Standardization in Measurement: Philosophical, Historical and Sociological Issues*, edited by Oliver Schlaudt and Lara Huber, 81–93. History and Philosophy of Technoscience 7. London: Pickering & Chatto.
- Grennan, Wayne. 1994. "Are 'Gap-Fillers' Missing Premises?" *Informal Logic* 16 (3). <https://doi.org/10.22329/il.v16i3.2456>.
- Groarke, Leo. 1992. "In Defense of Deductivism: Replying to Govier." In *Argumentation*

- Illuminated*, edited by F. H. van Eemeren, R Grootendorst, J. Anthony Blair, and Charles Arthur Willard, 113–21. Amsterdam: International Society for Study of Argument.
- . 1999. “Deductivism Within Pragma-Dialectics.” *Argumentation* 13 (1): 1–16. <https://doi.org/10.1023/A:1007771101651>.
- . 2019. “Informal Logic.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University.
- Groarke, Leo, Christopher William Tindale, and Linda Fisher. 1997. *Good Reasoning Matters!: A Constructive Approach to Critical Thinking*. 2nd ed. Oxford: Oxford University Press.
- Guetzkow, Harold. 1950. “Unitizing and Categorizing Problems in Coding Qualitative Data.” *Journal of Clinical Psychology* 6: 47–58.
- Gyngell, Christopher, Thomas Douglas, and Julian Savulescu. 2017. “The Ethics of Germline Gene Editing.” *Journal of Applied Philosophy* 34 (4): 498–513. <https://doi.org/10.1111/japp.12249>.
- Habernal, Ivan. 2014. “Argumentation in User-Generated Content: Annotation Guidelines.”
- Habernal, Ivan, Judith Eckle-Kohler, and Iryna Gurevych. 2014. “Argumentation Mining on the Web from Information Seeking Perspective.” In *ArgNLP*.
- Habernal, Ivan, and Iryna Gurevych. 2016. “Argumentation Mining in User-Generated Web Discourse.” *Computational Linguistics* 43 (1): 125–79. [https://doi.org/10.1162/COLI\\_a\\_00276](https://doi.org/10.1162/COLI_a_00276).
- Hahn, Ulrike, and Jos Hornikx. 2016. “A Normative Framework for Argument Quality: Argumentation Schemes with a Bayesian Foundation.” *Synthese* 193 (6): 1833–73. <https://www.jstor.org/stable/43921200>.
- Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.
- Halpern, Joseph Y. 2005. *Reasoning about Uncertainty*. Cambridge: MIT Press.
- Hammersley, Martyn. 1987. “Some Notes on the Terms ‘Validity’ and ‘Reliability’.” *British Educational Research Journal* 13 (1): 73–82. <https://doi.org/10.1080/0141192870130107>.
- Hansen, Anders, Simon Cottle, Ralph Negrine, and Chris Newbold. 1998. *Mass Communication Research Methods*. New York: Palgrave Macmillan.
- Hansson, Sven Ove. 2016. “Evaluating the Uncertainties.” In *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty.*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 79–104. Cham: Springer.
- Hansson, Sven Ove, and Gertrude Hirsch Hadorn. 2016a. “Introducing the Argumentative Turn in Policy Analysis.” In *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty.*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 11–35. Cham: Springer.
- , eds. 2016b. *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty*. Cham: Springer.
- Hardy, Cynthia, Bill Harley, and Nelson Phillips. 2004. “Discourse Analysis and Content Analysis: Two Solitudes.” *Qualitative Methods* 2 (1): 19–22.
- Harris, John, and Marcy Darnovsky. 2018. “Pro and Con: Should Gene Editing Be Performed on Human Embryos?” *Magazine*. <https://www.nationalgeographic.com/magazine/article/human-gene-editing-pro-con-opinions>.
- Henkemans, A. Francisca Snoeck. 2000. “State-of-the-Art: The Structure of Argumentati-

- on." *Argumentation* 14 (4): 447–73. <https://doi.org/10.1023/A:1007800305762>.
- Hitchcock, David. 1985. "Enthymematic Arguments." *Informal Logic* 7: 83–97.
- . 1998. "Does the Traditional Treatment of Enthymemes Rest on a Mistake?" *Argumentation* 12 (1): 15–37. <https://doi.org/10.1023/A:1007738519694>.
- . 2002. "A Note on Implicit Premisses." *Informal Logic* 22 (2): 158–59. <https://doi.org/10.22329/il.v22i2.2581>.
- . 2011. "Inference Claims." *Informal Logic* 31 (3): 191–229. <https://doi.org/10.22329/il.v31i3.3400>.
- . 2015. "The Linked-Convergent Distinction." In *Reflections on Theoretical Issues in Argumentation Theory*, edited by Frans H. van Eemeren and Bart Garssen, 83–91. Argumentation Library. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-319-21103-9\\_6](https://doi.org/10.1007/978-3-319-21103-9_6).
- . 2019. "The Problem of Missing Premisses." In *Informal Logic: A "Canadian" Approach to Argument*, edited by Federico Puppo, 104–29. Windsor: University of Windsor.
- Hovy, Eduard H. 1995. "The Multifunctionality of Discourse Markers." In *Proceedings of the Workshop on Discourse Markers*, 1–11.
- Hux, Karen, Dixie Sanger, Robert Reid, and Amy Maschka. 1997. "Discourse Analysis Procedures: Reliability Issues." *Journal of Communication Disorders* 30 (2): 133–50. [https://doi.org/10.1016/S0021-9924\(96\)00060-3](https://doi.org/10.1016/S0021-9924(96)00060-3).
- Jain, Aliya. 2021. "Opinion: CRISPR Technology Violates Informed Consent, May Usher Era of Eugenics." *Bruin Medical Review*. <https://bruinmedicalreview.com/2021/05/07/opinion-crispr-technology-violates-informed-consent-may-usher-era-of-eugenics/>.
- Jaipal-Jamani, Kamini. 2014. "Assessing the Validity of Discourse Analysis: Transdisciplinary Convergence." *Cultural Studies of Science Education* 9 (4): 801–7. <https://doi.org/10.1007/s11422-013-9567-7>.
- Jaworski, Adam, and Nikolas Coupland. 2006. "Introduction: Perspectives on Discourse Analysis." In *The Discourse Reader*, 1–37. London: Routledge.
- JCGM. 2012. "International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (2008 Version with Minor Corrections)." Sèvres: Joint Committee for Guides in Metrology.
- Johnson, R. H., and J. A. Blair. 2002. "Informal Logic and the Reconfiguration of Logic." In *Studies in Logic and Practical Reasoning*, edited by Dov M. Gabbay, Ralph H. Johnson, Hans Jürgen Ohlbach, and John Woods, 1:339–96. Handbook of the Logic of Argument and Inference. Amsterdam: Elsevier. [https://doi.org/10.1016/S1570-2464\(02\)80010-6](https://doi.org/10.1016/S1570-2464(02)80010-6).
- Johnson, Ralph H. 2006. "Making Sense of 'Informal Logic'." *Informal Logic* 26 (3): 231–58. <https://doi.org/10.22329/il.v26i3.453>.
- Johnson, Ralph H, and J Anthony Blair. 1977. *Logical Self-Defense*. Toronto: McGraw-Hill Ryerson.
- Johnson, Ralph H., and J. Anthony Blair. 1994. *Logical Self-Defense*. 2nd Revised edition. New York: McGraw-Hill.
- Kampf, Zohar. 2015. "Political Discourse Analysis." In *The International Encyclopedia of Language and Social Interaction*, 1–17. Hoboken: John Wiley & Sons. <https://doi.org/10.1002/9781118611463.wbielsi096>.

- Kienpointner, Manfred. 1992a. *Alltagslogik Struktur Und Funktion von Argumentationsmustern*. Stuttgart- Bad Cannstatt: Frommann-Holzboog.
- . 1992b. “How to Classify Arguments.” In *Argumentation Illuminated*, edited by F. H. van Eemeren, R. Grootendorst, J. Anthony Blair, and Charles Arthur Willard, 178–88. Mahwah: Lawrence Erlbaum Associates.
- Kirschner, Christian, Judith Eckle-Kohler, and Iryna Gurevych. 2015. “Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications.” *NAACL HLT 2015*, 1.
- Koons, Robert. 2017. “Defeasible Reasoning.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2017. Metaphysics Research Lab, Stanford University.
- Kracauer, Siegfried. 1952. “The Challenge of Qualitative Content Analysis.” *The Public Opinion Quarterly* 16 (4): 631–42. <https://www.jstor.org/stable/2746123>.
- Krippendorff, Klaus. 1970. “Estimating the Reliability, Systematic Error and Random Error of Interval Data.” *Educational and Psychological Measurement* 30 (1): 61–70. <https://doi.org/10.1177/001316447003000105>.
- . 1995. “On the Reliability of Unitizing Continuous Data.” *Sociological Methodology* 25: 47–76. <https://doi.org/10.2307/271061>.
- . 2004a. *Content Analysis an Introduction to Its Methodology*. 2. ed. Thousand Oaks, Calif. [u.a.]: Sage.
- . 2004b. “Reliability in Content Analysis: Some Common Misconceptions and Recommendations.” *Human Communication Research* 30 (3): 411–33.
- . 2004c. “Measuring the Reliability of Qualitative Text Analysis Data.” *Quality and Quantity* 38 (6): 787–800. <https://doi.org/10.1007/s11135-004-8107-7>.
- . 2011. “Agreement and Information in the Reliability of Coding.” *Communication Methods and Measures* 5 (2): 93–112. <https://doi.org/10.1080/19312458.2011.568376>.
- . 2013. *Content Analysis an Introduction to Its Methodology*. 3. ed. Los Angeles, Calif. [u.a.]: Sage.
- . 2016. “Misunderstanding Reliability.” *Methodology* 12 (4): 139–44. <https://doi.org/10.1027/1614-2241/a000119>.
- Krippendorff, Klaus, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. “On the Reliability of Unitizing Textual Continua: Further Developments.” *Quality & Quantity* 50 (6): 2347–64. <https://doi.org/10.1007/s11135-015-0266-1>.
- Kriz, Jürgen. 1978. “Methodologische Grundlagen Der Inhaltsanalyse.” In *Grundlagen Und Modelle Der Inhaltsanalyse: Bestandsaufnahme Und Kritik*, 29–55. 117. Reinbek bei Hamburg: Rowohlt.
- Kuckartz, Udo. 2012. *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung*. Weinheim: Beltz-Juventa.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.
- Lakatos, Imre. 1968. “Criticism and the Methodology of Scientific Research Programmes.” *Proceedings of the Aristotelian Society* 69: 149–86. <https://www.jstor.org/stable/4544774>.
- . 1970. “Falsification and the Methodology of Scientific Research Programmes.” In *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science*, edited by Imre Lakatos and Alan Musgrave, 91–196.



- Cambridge: Cambridge University Press.
- Lawrence, John, and Chris Reed. 2019. "Argument Mining: A Survey." *Computational Linguistics*, October, 1–55. [https://doi.org/10.1162/COLI\\_a\\_00364](https://doi.org/10.1162/COLI_a_00364).
- Lawrence, John, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. "Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling." In *Proceedings of the First Workshop on Argumentation Mining*, 79–87. Baltimore, Maryland: Association for Computational Linguistics.
- Lippi, Marco, and Paolo Torroni. 2016. "Argumentation Mining: State of the Art and Emerging Trends." *ACM Trans. Internet Technol.* 16 (2): 10:1–25. <https://doi.org/10.1145/2850417>.
- Lisch, Ralf, and Jürgen Kriz. 1978. *Grundlagen Und Modelle Der Inhaltsanalyse: Bestandsaufnahme Und Kritik*. 117. Reinbek bei Hamburg: Rowohlt.
- Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28 (4): 587–604. <https://doi.org/10.1111/j.1468-2958.2002.tb00826.x>.
- Luckhardt, C. Grant, and William Bechtel. 1994. *How to Do Things with Logic*. Hillsdale, NJ [u.a.]: Erlbaum.
- Lumer, Christoph. 2011. "Argument Schemes: An Epistemological Approach." In *OSSA Conference Archive*.
- Mari, Luca. 2015. "An Overview of the Current Status of Measurement Science: From the Standpoint of the International Vocabulary of Metrology (VIM)." In *Standardization in Measurement: Philosophical, Historical and Sociological Issues*, edited by Oliver Schlaudt and Lara Huber, 69–80. History and Philosophy of Technoscience 7. London: Pickering & Chatto.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. "The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment." *Computational Linguistics* 41 (3): 437–79. [https://doi.org/10.1162/COLI\\_a\\_00227](https://doi.org/10.1162/COLI_a_00227).
- Mayring, Philipp. 1983. *Qualitative Inhaltsanalyse : Grundlagen Und Techniken*. Weinheim: Beltz.
- . 2001. "Combination and Integration of Qualitative and Quantitative Analysis." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 2 (1). <https://doi.org/10.17169/fqs-2.1.967>.
- . 2014. "Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution." Social Science Open Access Repository (SSOAR).
- . 2015a. "Qualitative Content Analysis: Theoretical Background and Procedures." In *Approaches to Qualitative Research in Mathematics Education: Examples of Methodology and Methods*, edited by Angelika Bikner-Ahsbabs, Christine Knipping, and Norma Presmeg, 365–80. Advances in Mathematics Education. Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-017-9181-6\\_13](https://doi.org/10.1007/978-94-017-9181-6_13).
- . 2015b. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 12th ed. Weinheim: Beltz.
- Merten, Klaus. 1995. *Inhaltsanalyse: Einführung in Theorie, Methode Und Praxis*. 2nd ed. Wiesbaden: Springer.
- Mochales, Raquel, and Aagje Ieven. 2009. "Creating an Argumentation Corpus: Do Theories Apply to Real Arguments? A Case Study on the Legal Argumentation of the

- ECHR.” In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 21–30. ICAIL ’09. New York: Association for Computing Machinery. <https://doi.org/10.1145/1568234.1568238>.
- Neblo, Michael A. 2011. “Family Disputes: Diversity in Defining and Measuring Deliberation.” *Swiss Political Science Review* 13 (4): 527–57. <https://doi.org/10.1002/j.1662-6370.2007.tb00088.x>.
- Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. Thousand Oaks: SAGE Publications.
- Newman, Debbie, Trevor Sather, and Ben Woolgar. 2013. *Pros and Cons: A Debaters Handbook*. 19. Edition. London: Routledge.
- NHGRI. 2017. “What Are the Ethical Concerns of Genome Editing?” *Genome.gov*. <https://www.genome.gov/about-genomics/policy-issues/Genome-Editing/ethical-concerns>.
- Paltridge, Brian. 2012. *Discourse Analysis: An Introduction*. London: Bloomsbury Publishing.
- Peirce, Charles S. 1955. “What Is a Leading Principle?” In *Philosophical Writings of Peirce*, edited by J. Buchler, 129–34. New York: Dover.
- Peldszus, Andreas, and Manfred Stede. 2013. “From Argument Diagrams to Argumentation Mining in Texts: A Survey.” *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7 (1): 1–31. <https://doi.org/10.4018/jcini.2013010101>.
- . 2015. “An Annotated Corpus of Argumentative Microtexts.” In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, 2:801–16.
- Peldszus, Andreas, Saskia Warzecha, and Manfred Stede. 2016. “Annotation Guidelines for Argumentation Structure. English Translation of Chapter ‘Argumentationsstruktur.’” In *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*. Vol. 8. Potsdam: Universitätsverlag Potsdam.
- Perelman, Chaim, and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric a Treatise on Argumentation*. Notre Dame: University of Notre Dame Press.
- Pollock, John. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge: MIT Press.
- Popper, Karl. [1959] 2002. *The Logic of Scientific Discovery*. 2nd ed. London: Routledge. <https://doi.org/10.4324/9780203994627>.
- Potter, W. James, and Deborah Levine-Donnerstein. 1999. “Rethinking Validity and Reliability in Content Analysis.” *Journal of Applied Communication Research* 27 (3): 258–84. <https://doi.org/10.1080/00909889909365539>.
- Prakken, Henry. 2010. “An Abstract Framework for Argumentation with Structured Arguments.” *Argument & Computation* 1 (2): 93–124. <https://doi.org/10.1080/19462160903564592>.
- Prakken, Henry, and Gerard Vreeswijk. 2002. “Logics for Defeasible Argumentation.” In *Handbook of Philosophical Logic*, edited by D. M. Gabbay and F. Guenther, 219–318. Handbook of Philosophical Logic. Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-017-0456-4\\_3](https://doi.org/10.1007/978-94-017-0456-4_3).
- Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. “The Penn Discourse TreeBank 2.0.” In *Proceedings*

- of the Sixth International Language Resources and Evaluation (LREC). Marrakech: European Language Resources Association (ELRA).
- Prasad, Rashmi, Bonnie Webber, and Aravind Joshi. 2014. "Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation." *Computational Linguistics* 40 (4): 921–50. [https://doi.org/10.1162/COLI\\_a\\_00204](https://doi.org/10.1162/COLI_a_00204).
- Priest, Graham, Francesco Berto, and Zach Weber. 2018. "Dialetheism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University.
- Reed, Chris, Douglas Walton, and Fabrizio Macagno. 2007. "Argument Diagramming in Logic, Law and Artificial Intelligence." *The Knowledge Engineering Review* 22 (01): 87.
- Reinmuth, Friedrich. 2014. "Hermeneutics, Logic and Reconstruction." *History of Philosophy & Logical Analysis* 17 (1): 152–90. <https://doi.org/10.30965/26664275-01701008>.
- Reiss, Julian, and Jan Sprenger. 2020. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University.
- Riffe, Daniel, Stephen Lacy, Frederick Fico, Brendan Watson, Stephen Lacy, Frederick Fico, and Brendan Watson. 2014. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. 3rd ed. Routledge. <https://doi.org/10.4324/9780429464287>.
- Ritsert, Jürgen. 1972. *Inhaltsanalyse Und Ideologiekritik Ein Versuch Über Kritische Sozialforschung*. Frankfurt am Main: Athenäum-Fischer-Taschenbuch-Verl.
- Salmon, Wesley C. 1963. *Logic*. Englewood Cliffs: Prentice Hall.
- Schervish, Mark J. 1995. *Theory of Statistics*. New York: Springer.
- Scheuer, Oliver, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. "Computer-Supported Argumentation: A Review of the State of the Art." *International Journal of Computer-Supported Collaborative Learning* 5 (1): 43–102. <https://doi.org/10.1007/s11412-009-9080-x>.
- Schmidt, Stefan. 2009. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 13 (2): 90–100. <https://doi.org/10.1037/a0015108>.
- Scholman, Merel C. J., and Vera Demberg. 2017. "Examples and Specifications That Prove a Point: Identifying Elaborative and Argumentative Discourse Relations." *Dialogue & Discourse* 8 (2): 56–83. <https://doi.org/10.5087/dad.2017.203>.
- Schreier, Margrit. 2012. *Qualitative Content Analysis in Practice*. Thousand Oaks: Sage publications.
- Schwartz, Thomas. 1981. "Logic as a Liberal Art." *Teaching Philosophy* 4 (3/4): 231–47.
- Scott, William A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding." *The Public Opinion Quarterly* 19 (3): 321–25. <https://www.jstor.org/stable/2746450>.
- Shapiro, Stewart. 2006. "Necessity, Meaning, and Rationality: The Notion of Logical Consequence." In *A Companion to Philosophical Logic*, 225–40. Hoboken: John Wiley & Sons. <https://doi.org/10.1002/9780470996751.ch15>.
- Stab, Christian Matthias Edwin. 2017. "Argumentative Writing Support by Means of Natural Language Processing." PhD thesis, Darmstadt: Technische Universität Darmstadt.

- Stab, Christian, and Iryna Gurevych. 2014. "Annotating Argument Components and Relations in Persuasive Essays." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1501–10.
- Stab, Christian, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. "Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective." In *Workshop on Frontiers and Connections Between Argumentation Theory and Natural Language Processing*.
- Stede, Manfred, and Jodi Schneider. 2018. "Argumentation Mining." *Synthesis Lectures on Human Language Technologies* 11 (2): 1–191. <https://doi.org/10.2200/S00883ED1V01Y201811HLT040>.
- Steenbergen, Marco R., André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. "Measuring Political Deliberation: A Discourse Quality Index." *Comparative European Politics* 1 (1): 21–48. <https://doi.org/10.1057/palgrave.cep.6110002>.
- Steiner, Jürg, André Bächtiger, Marco Steenbergen, and Markus Spörndli. 2004. *Deliberative Politics in Action: Analyzing Parliamentary Discourse*. Theories of Institutional Design. Cambridge: Cambridge University Press.
- Stevens, S. S. 1951. "Mathematics, Measurement, and Psychophysics." In *Handbook of Experimental Psychology*, 1–49. Oxford: Wiley.
- Strasser, Christian, and G. Aldo Antonelli. 2019. "Non-Monotonic Logic." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University.
- Sundholm, B. G. 2006. "Varieties of Consequence." In *A Companion to Philosophical Logic*, 241–55. Hoboken: John Wiley & Sons. <https://doi.org/10.1002/9780470996751.ch16>.
- Suppes, Patrick, R. Duncan Luce, Amos Tversky, and David H. Krantz. 2007. *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Vol. I. Mineola: Dover Publication.
- Tal, Eran. 2020. "Measurement in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University.
- Tannen, Deborah, Heidi E. Hamilton, and Deborah Schiffrin. 2015. "Introduction to the First Edition." In *The Handbook of Discourse Analysis*, 1–7. Hoboken: John Wiley & Sons.
- Taylor, John R. 1996. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. 2nd ed. Sausalito: University Science Books.
- Tetens, Holm. 2004. *Philosophisches Argumentieren Eine Einführung*. Orig.-Ausg. München: Beck.
- Thomas, Stephen Naylor. 1986. *Practical Reasoning in Natural Language*. 3rd ed. Englewood Cliffs N.J.: Prentice-Hall.
- Titscher, Stefan, Michael Meyer, Ruth Wodak, and Eva Vetter. 2000. *Methods of Text and Discourse Analysis: In Search of Meaning*. First edition. London ; Thousand Oaks Calif.: SAGE Publications Ltd.
- Toulmin, Stephen Edelston. 1958. *The Uses of Argument*. Cambridge: Cambridge University Press.
- Verheij, Bart. 1999. "Logic, Context and Valid Inference. Or: Can There Be a Logic of Law." *Legal Knowledge Based Systems. JURIX*, 109–21.
- . 2003. "Dialectical Argumentation with Argumentation Schemes: An Approach to

- Legal Logic.” *Artificial Intelligence and Law* 11 (2): 167–95. <https://doi.org/10.1023/B:ARTI.0000046008.49443.36>.
- Visser, Jacky, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. “Annotating Argument Schemes.” *Argumentation* 35 (1): 101–39. <https://doi.org/10.1007/s10503-020-09519-x>.
- Walton, Douglas. 1996a. *Argument Structure: A Pragmatic Theory*. Toronto: University of Toronto Press. <https://www.jstor.org/stable/10.3138/j.ctvcj2q7q>.
- . 1996b. *Argumentation Schemes for Presumptive Reasoning*. 1st ed. London: Routledge.
- . 2006. *Fundamentals of Critical Argumentation*. Cambridge: Cambridge University Press.
- . 2007. *Media Argumentation: Dialectic, Persuasion and Rhetoric*. Cambridge: Cambridge University Press.
- . 2009. “Argumentation Theory: A Very Short Introduction.” In *Argumentation in Artificial Intelligence*, edited by Guillermo Simari and Iyad Rahwan, 1–22. Boston, MA: Springer US. [https://doi.org/10.1007/978-0-387-98197-0\\_1](https://doi.org/10.1007/978-0-387-98197-0_1).
- . 2010. *Appeal to Expert Opinion: Arguments from Authority*. University Park, Pennsylvania: Pennsylvania State University Press.
- . 2012. “Using Argumentation Schemes for Argument Extraction: A Bottom-Up Method.” *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 6 (3): 33–61. <https://doi.org/10.4018/jcini.2012070103>.
- Walton, Douglas, and David M Godden. 2005. “The Nature and Status of Critical Questions in Argumentation Schemes.” *Argumentation in Artificial Intelligence and Law*, 103–11.
- Walton, Douglas, and Thomas F Gordon. 2011. “Modeling Critical Questions as Additional Premises.” In *Proceedings of the 8th International OSSA Conference*. Windsor.
- Walton, Douglas, and Thomas F. Gordon. 2015. “Formalizing Informal Logic.” *Informal Logic* 35 (4): 508–38. <https://doi.org/10.22329/il.v35i4.4335>.
- Walton, Douglas, and Chris Reed. 2003. “Diagramming, Argumentation Schemes and Critical Questions.” In *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, edited by Frans H. Van Eemeren, J. Anthony Blair, Charles A. Willard, and A. Francisca Snoeck Henkemans, 195–211. Argumentation Library. Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-007-1078-8\\_16](https://doi.org/10.1007/978-94-007-1078-8_16).
- Walton, Douglas, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge: Cambridge University Press.
- Wigmore, John Henry. 1913. *The Principles of Judicial Proof*. Boston: Little, Brown, and Co.
- Yanal, Robert J. 1988. *Basic Logic*. St. Paul: West Publishing Co.
- . 1991. “Dependent and Independent Reasons.” *Informal Logic* 13 (3).
- . 2003. “Linked and Convergent Reasons - Again.” In *5th Ontario Society for the Study of Argumentation Conference*. OSSA Conference Archive.
- Zenker, Frank, Jan Albert van Laar, Pedro Abreu, Mette Bengtsson, Paula Castro, Istvan Danka, Barbara de Cock, et al. 2020. “Goals and Functions of Public Argumentation.” SocArXiv. <https://doi.org/10.31235/osf.io/ezub8>.





