Dimensions of Register Variation in Contemporary German

Andressa Costa (KIT)
Tony Berber Sardinha (PUC-SP)

Abstract

This paper presents the first full multi-dimensional (MD) analysis of spoken and written German. While earlier MD studies have targeted English, Spanish, and Portuguese, German has remained unexplored. Our investigation uses a corpus comprising over 3,000 texts distributed across 52 varieties. It was annotated with three part-of-speech taggers, enabling the extraction of over sixty lexicogrammatical variables. Five factors were uncovered through exploratory factor analysis, labelled by interpreting a large number of texts for their underlying communicative functions as following: (1) Processuality vs. Reification, (2) Involved vs. Informational Production, (3) Overt Expression of Persuasion, (4) Evidentiality Strategy, and (5) Narrative Discourse. These dimensions explain up to 78.1% of variance between registers. Finally, we compare the German dimensions with those reported for English, Spanish, and Portuguese. Major distinctions such as oral/literate and narrative/non-narrative recur across languages, supporting their status as potential cross-linguistic dimensional universals, while German exhibits language-specific configurations in evidential discourse.

1. Introduction

Register is a fundamental aspect of human language and is central to explaining linguistic variation (Biber & Egbert 2023; Biber, 2012; Biber et al. 2006). Register variation is pervasive in communication, demonstrating how speakers adapt their language use to different situations and contexts (Biber 1988, 1995). Although many studies have focused on describing registers-specific situational and linguistic characteristics by analyzing texts in detail, Multi-Dimensional Analysis (MDA) (Biber 1988) offers a more comprehensive view of register differences. It allows researchers to investigate numerous linguistic features across a wide range of registers, providing a detailed description of linguistic variation (Berber Sardinha & Veirano Pinto 2014, 2019; Biber & Conrad 2001).

Although the MD approach has been applied to various languages (Berber Sardinha et al. 2014a; Besnier 1988; Biber 1988; Biber & Hared 1992; Biber et al. 2006; Kim & Biber 1994; Lamb 2008; Parodi 2007; Purvis 2008), its application to contemporary German remains relatively unexplored. Previous studies on German register variation have primarily focused on specific linguistic features or limited sets of registers, leaving a gap in comprehensive analyses of the German language use across registers. For instance, Koch and Österreicher (2012) developed an influential model emphasizing the roles of immediacy and distance in shaping language use across the spoken-written continuum. Ágel and Hennig (2006) refined the concepts of communicative immediacy and distance, offering a more detailed framework for understanding German register variation. Neumann (2014) combined Biber's quantitative approach with Systemic Functional Linguistics (Halliday & Matthiessen 2004) to explore register variation in German and English.

Koch and Österreicher (2012) provide a foundational model for understanding the continuum between spoken and written language, emphasizing the roles of immediacy and distance in shaping language use. In this model, they develop several significant conceptual differentiations and establish distinctions between phonic and graphic medium, spoken and written mode (cf. Koch & Österreicher 2012: 443–444), so that their model presents four possibilities: a) spoken conception + phonic realization; b) written conception + graphic realization; c) written conception + phonic realization; d) spoken conception + graphic realization.

The model situates eleven discourse types along this continuum, emphasizing that communicative constellations are shaped by immediacy and distance rather than the medium itself. Key parameters include social relationships, partner dynamics, theme fixation, public exposure, spontaneity, and contextual factors. Immediacy is characterized as dialogic, spontaneous and low-planned by the authors, leading to tentative and self-generating discourses. It is associated with the phonics code and relies on contextual clues, allowing for economical or extensive verbalization. In contrast, distance involves planned, monologic utterances that aim for a definite form, which is often realized in the graphic code. Texts in this category are compact, complex, and information-dense due to their detachment from the immediate situation and the planning involved Koch and Österreicher (2012: 448).

Ágel and Hennig (2006) refine Koch and Österreicher's model of communicative immediacy (Nähe) and distance (Distanz) by making it operational for systematic grammatical and lexical analysis across registers and genres, especially in texts from 1650 to 2000. They criticize the earlier model for lacking clear criteria to locate discourse types between the poles of immediacy and distance and instead propose a hierarchical framework, from universal axioms to specific linguistic features, aimed at practical text analysis. Initially, immediacy in language was associated with specific grammatical features at the micro-level. However, they argue for a broader perspective that includes macro-level analysis, considering the overall grammatical patterns that shape a text's profile. The methodology involves comparing texts against prototypical immediacy and distance texts to determine their immediacy language characteristics in a three-step process: micro-analysis (comparing the text to an immediacy comparison text), macro-analysis (comparing the text to both immediacy and distance comparison texts), and averaging the results. The analysis assesses the percentage of immediacy language features, which allows for comparing texts across a continuum from oral to written language.

Neumann (2014)'s work builds upon both Biber (1988) quantitative empirical approach to register variation and the principles of Systemic Functional Linguistics (Halliday & Matthiessen 2004). The objective is to explore the quantitative distribution of indicators that define language use from three perspectives: intralingual, contrastive and translations. The study focuses on comparing registers within and across English and German, as well as examining translations. It analyzes eight registers: essays, fiction, instruction manuals, popular scientific articles, business letters, political speeches, promotional texts, and website content. The corpus includes original texts in both languages and their translations, allowing for a comprehensive examination of intralingual, contrastive, and translation-specific variations. This approach enables researchers to investigate how language potential is utilized in different situations.

Building on Biber (1988) MDA, the current study seeks to extend empirical research on register variation by applying a more comprehensive method to contemporary German. It focuses on synchronic variation in modern contexts, unlike earlier work limited to historical or diachronic perspectives (Ágel & Hennig 2006). It also expands beyond Neumann's focus on written registers by incorporating spoken data, thereby addressing a broader spectrum of communicative situations. Hence, the primary aim of this study is to uncover the dimensions of linguistic variation in contemporary German through an MD approach by describing the linguistic features that characterize different registers in German, identifying the functional characteristics associated with these features, and determining the underlying dimensions of register variation. The research questions to be addressed are:

1. Which linguistic features frequently co-occur in different German registers?
2. What are the underlying dimensions of variation across these registers?

This paper is organized as follows: Section 2 describes the corpus and methodological approach, including the selection of linguistic features and statistical analyses employed. Section 3 presents the MDA results, the interpretation of the dimensions of variation identified, and their linguistic and functional characteristics. The paper concludes with a discussion of the results in section 4.

2. Methods

2.1 Corpus Collection and Annotation

The Koder corpus (Korpus Deutscher Register) is a comprehensive resource for studying register variation in contemporary German, covering written, spoken and computer-mediated communication. The need to design this corpus arose because existing German corpora did not offer sufficient diversity in terms of modes (spoken, written, and digital) or register coverage. Koder was compiled by integrating both pre-existing corpora and a substantial collection of new materials specific to this project. The aim was to capture the wide variety of communication situations encountered by present-day German speakers.

The current version of the corpus, which has been expanded from Costa (2019)'s initial compilation, Comprises 52 registers, 3,086 texts and over 14 million words. All selected texts were produced between 1990 and 2021, to reflect contemporary German language use. Table 1 shows a breakdown of the corpus by text, word, and register counts by mode.

	Texts	Words
Written	2,182	11,734,856
Spoken	886	3,070,129
Total	3,068	14,804,985
	Registers	
Written	33	
Spoken	19	
Total	52	

Table 1: Corpus design

The corpus comprises materials from two main sources: Pre-existing corpora and additional texts compiled from internet sources, publishing houses, and other public repositories, always ensuring that authorship and provenance criteria were met as rigorously as possible. Following pre-existing corpora were used in the analysis:

- The Database for Spoken German (DGD) from the Institute for the German Language (IDS), which includes material from the Forschungs- und Lehrkorpus (FOLK) and Gesprochene Wissenschaftssprache (GWISS);
- Dortmunder-Chat-Korpus (Beißwenger et al. 2016);
- German Political Speeches Corpus (Barbaresi 2012)
- Wikipedia user's talk from the Deutsches Referenzkorpus (DeReKo) (IDS).

The corpus comprises 52 registers organized by communication field that span a wide range of modern German communicative contexts, including: Academic and scientific communication, Mass media, Computer-mediated communication, Institutional and professional texts, Everyday communication, Fiction and non-fiction literature. Figure 1 presents the distribution of word counts across registers, while Figure 2 reports the corresponding text frequencies by register, thereby providing a concise overview of the corpus composition.

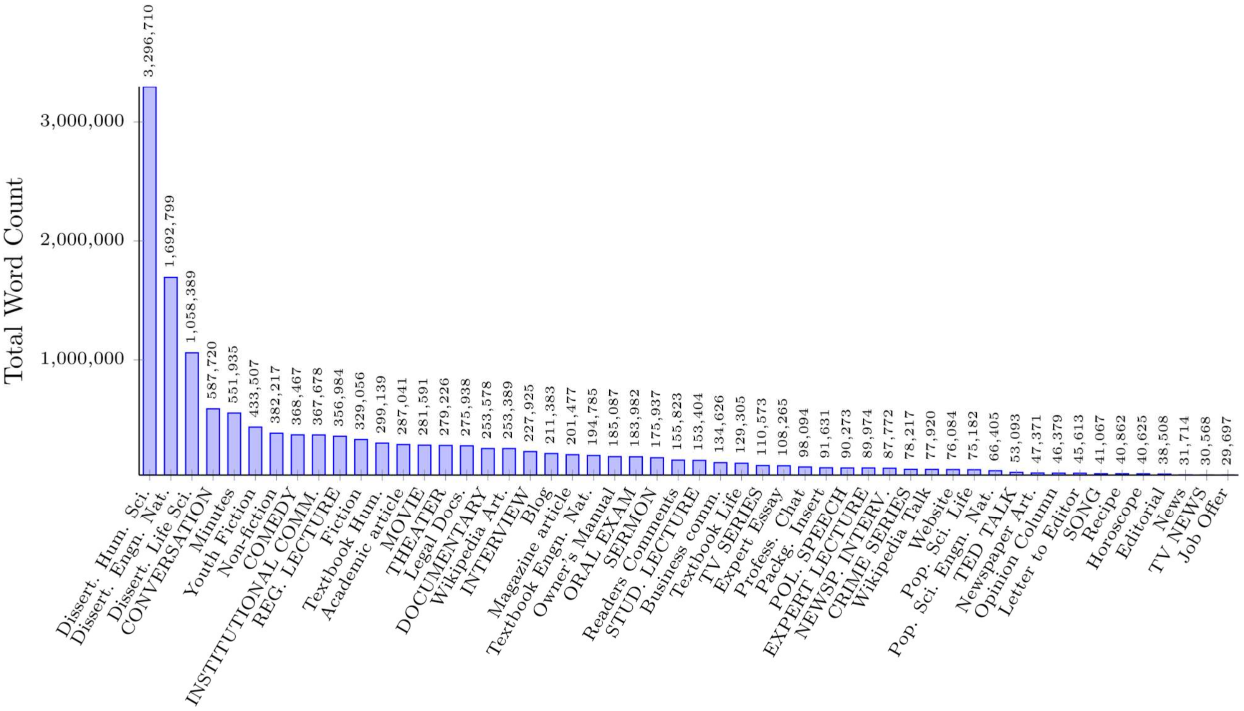

Figure 1: Word count by register

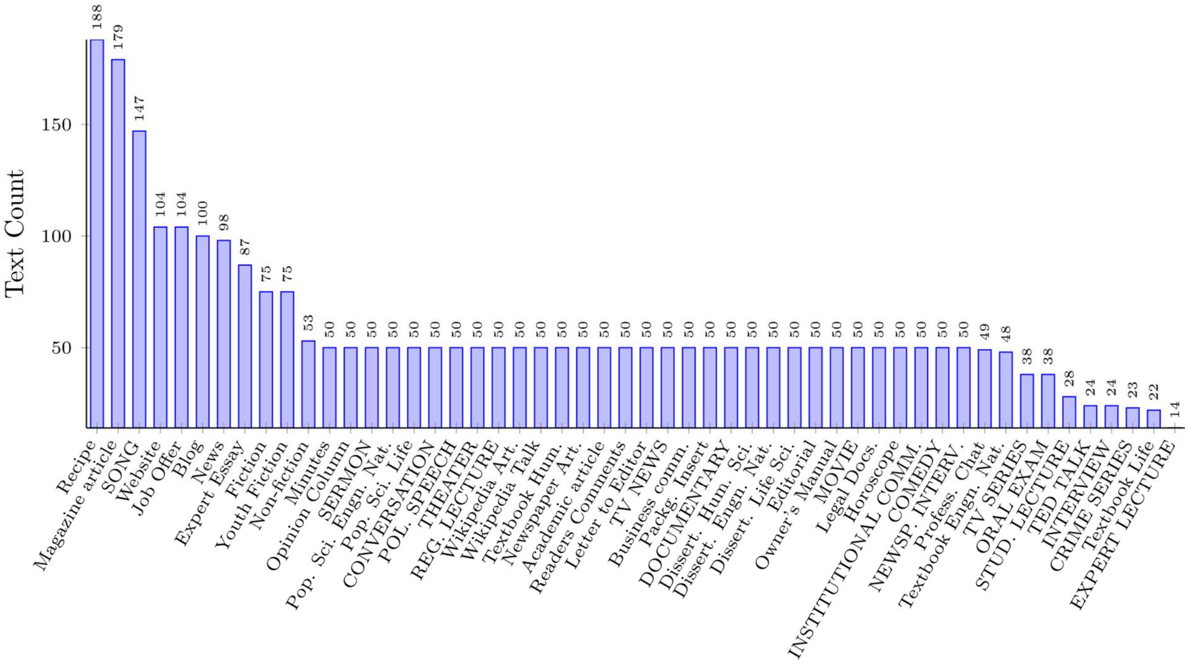

Figure 2: Text count by register

2.2 Linguistic features

Building on Biber (1988) as a starting point, a comprehensive review of existing research was conducted to identify the most relevant linguistic features of German for this study. The selected features are presented in table 2.

Lexical specificity:	1. type-token ratio, 2. average word length
Nominal forms:	3. abstract nouns, 4. regular nouns, 5. nominalization, 6. proper nouns
Pronouns:	7. 1st person pronoun, 8. 2nd person pronoun, 9. 3rd person pronoun, 10. indefinite pronoun, 11. possessive pronoun, 12. reflexive pronoun, 13. relative pronoun, 14. demonstrative pronoun, 15. interrogative pronoun
Article:	16. definite article, 17. indefinite article, 18. attributive indefinite pronoun, 19. possessive article, 20. demonstrative article, 21. interrogative article
Adjective:	22. attributive adjective, 23. predicative adjective
Adverb:	24. place adverb, 25. time adverb, 26. causal adverb, 27. modal adverb, 28. conjuncional adverb, 29. comment adverb, 30. pronominal adverb
Particles:	31. negation particle, 32. conversational particle, 33. intensity and focus particle, 34. modal particle, 35. discourse marker, 36. questions tag
Preposition:	37. preposition
Tense and verbal mood:	38. full verb indicative present, 39. auxiliary verb indicative present, 40. modal verb indicative present, 41. full verb indicative past, 42. auxiliary verb indicative past, 43. modal verb indicative past, 44. full verb *subjunctive I*, 45. auxiliary verb *subjunctive I*, 46.

	modal verb *subjunctive I*, 47. full verb *subjunctive II*, 48. auxiliary verb *subjunctive II*, 49. modal verb *subjunctive II*
Passive:	50. *werden*-passive
Modal verbs:	51. modal verbs
Verb of communication:	52. representative, 53. directive, 54. discendi, 55. comissive, 56. expressive, 57. declarative, 58. Verbs of cognition
Coordinating:	59. coordinating conjunctions
Infinitive with *zu:*	60. *zu* + Infinitive as complement
Subordinating:	61. subordinating clause with finite, 62. subordinating clause with *zu* + infinitive, 63. comparative clause, 64. Verb 2nd subordinate clause
Genitive:	65. genitive

Table 2: Linguistic Features

2.3 Corpus Analysis

The corpus was tagged with RFTagger (Schmid & Laws 2008) and TreeTagger for Spoken German (Westpfahl et al. 2017), and parsed with ParZu (Sennrich et al. 2009). As these tools could not account for all the linguistic features, the tags were combined to extract all selected features, which were then corrected to ensure accuracy.

R was used to count and normalise the frequency of each feature per 100 words, enabling comparison of texts of different lengths. This is a common process in MDA, as it provides a standardised measure. The baseline for normalisation is the shorter text in the corpus, which contains 100 words.

The data was evaluated prior to conducting the factor analysis to determine its suitability. A correlation matrix was generated to identify variables with correlation coefficients above .7. The following variables were not kept as separate variables in the factor analysis because their correlations were above this cut-off: demonstrative pronouns, the conjunction dass, the past perfect indicative, genitive, full verbs, auxiliary verbs in the present (in both the indicative and conjunctive moods), modal verbs as a single variable, modal particles, verba dicendi, verbs of cognition, and adverbs of place and time. Some individual variables were subsumed under composite variables or coded as individual variables. Full and auxiliary verbs in the present are incorporated into the variables 'indicative present' and 'subjunctive 1'; adverbs of place and time are subsumed under the variable 'adverb'; each modal verb is retained as an individual variable; and verba dicendi and verbs of cognition are included in the variable 'verbs of communication'.

After that, two tests were conducted to assess the suitability of the data for factor analysis. Bartlett's test of sphericity (Chi-square (1891) = 85007.21, $p < .001$) confirmed that there is sufficient significant correlation in the data for factor analysis. Additionally, the Kaiser-Meyer-Olkin (KMO) test yielded a value of 0.85, suggesting that the data is suitable for factor analysis.

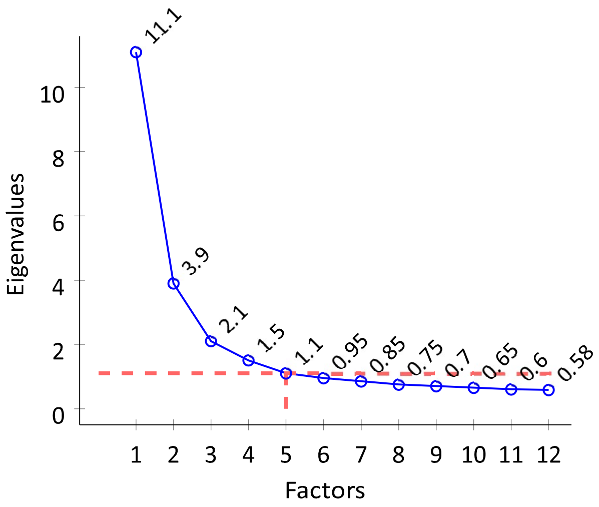

Figure 3: Scree plot

The scree plot analysis (Figure 3) suggested that a 5-factor solution was optimal. The factors were then extracted and rotated using the fa function from R psych package:

```
fa$_5$ <- fa(indices$_100$, fm = "pa", nfactors = 5, rotate = "promax")
print(fa$_5$, cut = 0)
```

Next, dimension or factor scores, which characterize each text according to each factor, were computed by summing the standardized frequencies (z-scores) of all positive features and subtracting the negative features in a factor (Biber, 1988, p.93). The average score for each dimension was then calculated for each text using the aov function from the tidyverse package and the EtaSq function from the DescTools package:

```
anova$_f1$ <- aov(f1$_score_100$ ~ register, data = df$_scores_100$)
anova$_f1$ \%>\% summary()
anova$_f1$ \%>\% EtaSq()
```

3. Dimensions of variation

To interpret the factors as textual dimensions, it is necessary to assess the shared function of the co-occurring features underlying each factor. As Gray and Egbert (2019: 1) observe, “texts that differ in situational features tend to exhibit distinct linguistic patterns, whereas texts with similar situational features tend to exhibit linguistic similarities.” Each dimension of register variation reflects a specific way in which language is used in different contexts. In this study, the linguistic co-occurrence patterns were interpreted by analyzing the positive and negative features of each dimension, focusing on registers marked by positive or negative features, considering dimension interpretations from previous MD analyses and reviewing pertinent literature on text linguistics, stylistics and spoken and written language in German.

3.1 Dimension 1: Processuality vs. Reification

The largest set of linguistic characteristics among the factors is contained in Factor 1, as Table 3 shows. This set comprises 17 positive and 8 negative features, accounting for 78% of the observed variation. The positive characteristics associated with this factor emphasize processual discourse, characterized by direct, face-to-face negotiation between speakers and listeners. In this type of discourse, production and reception occur simultaneously, allowing for immediate feedback, clarification, and adjustments. Shared action and contextual cues facilitate spontaneous interaction and shared knowledge. In contrast, the negative features represent a style characterized by compactness, complexity, and information density aiming at a definite form and reification. This style is typical of more distant forms of communication, being explained by relative detachment from the situation and the increased degree of planning required (cf. Koch & Österreicher 2012: 448 f.).

Positive Loadings	
Conversational particle (gespptk)	.83
Discourse marker (diskmark)	.82
Intensity particle (ptkifg)	.76
Adverb (adv)	.73
Predicative adjective (adjd)	.73
Indefinite pronouns (pindef)	.60
Question tag (qutag)	.51
Subordinating with finite verb (nebfin)	.49
Causal clause (kaus)	.46
Subjunctive 2 (vspast)	.43
Negation particle (ptkneg)	.37
Modal verb (vmod)	.36
Comment adverb (advkom)	.31
(Verb of communication (kommv)	.43)
(1st Person Pronoun (pper1)	.36)
(Indicative present (vipres)	.32)
(Interrogative pronoun (pinter)	.32)
Negative Loadings	
Preposition (prep)	-.64
Regular nouns (nreg)	-.62
Attributive adjective (adja)	-.50
Nominalization (noml)	-.44
Average word length (awl)	-.32
(Definite article (artdef)	-.35)
(Abstract nouns (nabs)	-.34)
(Proper noun (eignam)	-.30)

Table 3: Dim. 1 Loadings: Processuality vs. Reification

The positive features of Dimension 1—adverbs, conversational particles, discourse markers, question tags, and intensity/focus particles—jointly manage the dynamic demands of face-to-face communication, where planning and production occur simultaneously and multiple metacommunicative functions are required. They guide comprehension, attract attention, provide pre-utterance comments, and structure interaction through opening, continuation, and closing signals, thereby coordinating participants and maintaining intersubjectivity in ongoing discourse (cf. Schwitalla 2012: 157; Dudenredaktion 2022: 546). Beyond interaction management, these features also encode evaluative and affective stance, as intensity and focus particles together with comment adverbs enable speakers to express attitudes, assessments, and emotions, making utterances more persuasive and contextually precise (cf. Dudenredaktion 2022: 798).

Predicative adjectives and subordinating clauses characterize a verbal style that yields the accessible expressions typical of spoken discourse, in contrast to the dense nominal constructions of written language, by distributing information across explicit clauses and directly expressing actions and states to meet real-time processing demands. Indefinite pronouns function as strategic resources for managing referential imprecision in immediate, process-oriented discourse, indicating that referents cannot be specified more precisely; this vagueness constitutes a pragmatic adaptation to real-time interaction rather than a communicative deficit (cf. Schwitalla 2012: 161, Dudenredaktion 2022: 756).

Modal verbs and the subjunctive 2 jointly shape how speakers express intentions, desires, politeness, and degrees of certainty, allowing stance-taking, management of social relations, and calibration of illocutionary force, especially in requests, questions, and commands. Subjunctive 2 in particular softens requests and proposals and, in indirect speech with communication verbs, signals that what is reported is someone's claim or supposition rather than established fact (Schwitalla 2012: 137; Dudenredaktion 2022: 207, 238). Negation particles allow speakers to disagree or reposition themselves relative to prior utterances, rejecting or correcting earlier assertions or expectations and serving as key devices for discursive positioning and refutation within the evolving common ground (Apothtloz et al. 1993: 28; Lüdtke: 2008: 8, 42).

These features form a processual, immediate form of communication: language produced under real-time pressures, where speakers plan, produce, coordinate and evaluate simultaneously, while maintaining conversational flow and mutual understanding. They allow speakers to produce responsive discourse while being sensitive to interpersonal dynamics and processing constraints.

The negative pole of Dimension 1 is characterized by features that form nominal phrases, such as attributive adjectives, nouns, nominalizations, long words, definite articles and prepositions. This nominal style enables concise and information-dense expression through precise lexical choices and complex phrase structures. Prepositions and regular nouns provide informational support, while attributive adjectives elaborate the nominal elements more than clausal development. Nominalizations are used instead of adverbial subordinate clauses to describe actions and processes, reducing personal involvement and enhancing objectivity (Dudenredaktion 2022: 546). The greater the complexity of nominal phrases, the higher the information density, aligning with a formal, literate style (ebd.). Longer words, which indicate high information density, facilitate precise and exact content presentation (Biber 1988: 104). This style is predominantly written rather than spoken, as it requires the time and opportunity for careful crafting and revision. Consequently, texts characterized by the negative pole of Dimension 1 rely heavily on nouns and intricate noun phrases to pack dense informational content into relatively few words, resulting in a precise and formal linguistic style (Biber et al. 2006: 14).

This set of features supports a formal style of communication that focuses on carefully planned content rather than process-orientation and tentativeness. This organisational principle is typical of distance communication, which is characterised by formality, public access and officialness. In such contexts, spontaneity and emotional expression are minimised, and language use is characterised by distance, objectivity and precision (Schwitalla 2012: 21).

This plot shows the register distribution across this dimension. Registers such as conversation, interviews and institutional communication score highest on the processuality pole, indicating prototypical spontaneous, face-to-face spoken interaction marked by a high degree of participant involvement, a shared situational context and reliance on non-verbal cues. Academic spoken registers, including oral exams and student lectures, also display high means, demonstrating a strong association with this dimension despite their more formal nature. All registers scoring high on this pole are medially spoken, and vary between spontaneous or semi-planned dialogic interactions and non-dialogic formats. These registers share extensive common knowledge enabled by communicative parameters including co-presence in time and space, multimodality, turn-taking opportunities, and varying degrees of spontaneity and involvement (Koch & Österreicher 2012).

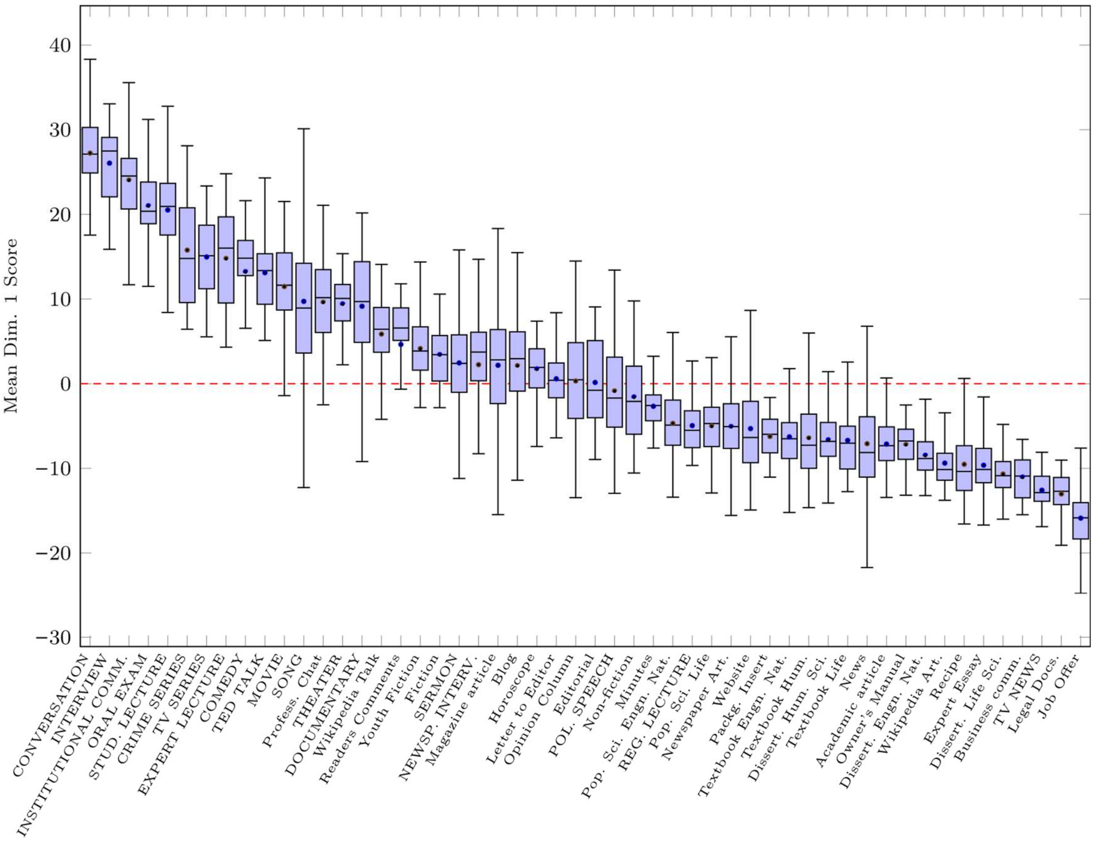

Figure 4: Mean Dim. 1 Scores by Register (R^2 =.781)

In contrast, registers such as legal documents, TV news, and job advertisement score higher on the reification pole (negative), consisting almost exclusively of written texts from diverse communicative fields. An exception is spoken news, which, despite being spoken, is produced under conditions similar to those of written communication. It is highly planned and lacks the spontaneity of real-time speech. These distance-oriented registers emerge from contexts of extreme literacy characterized by monologic production, an absence of speaker alternation, a lack of familiarity between participants, separation in space and time, fixed thematic content, orientation towards the public sphere, high levels of reflection, low levels of involvement and situational detachment (cf. Koch & Österreicher 2012: 447).

This excerpt illustrates the use of the positive features in everyday conversation, where language is highly interactive, context-bound, and produced in real time.

Text Sample 1: Positive Dim. 1

[Features in the text sample: **Conversational particle**, *Adverb, Comment adverb*, Subordination with finite verb, ***Interrogative pronoun***, *1st Person Pronoun*, **Modal verb, Indicative present**]

Conversation
AM: da **können** *wir* auch mit dem Linienbus hinfahren
PB: **ja** das ist der Linienbus quasi

AM: °h **ja** aber (.) du *vielleicht* **gibt** es da °h am Flughafen auch andere Unternehmen die einen **hinfahren** oder einen Taxifahrer wenn *wir* den **runterhandeln** oder **so**
PB: ***wo*** *wir* (rein) **müssen**
AM: irgendwie
PB: **ja** das **kriegen** *wir* schon hin

Translation
AM: *we* **can** also go there by the local bus
PB: **yeah** that's basically the local bus
AM: °h **yeah** but (.) you *maybe* there **are** °h at the airport also other companies that **go** there or a taxi driver if *we* **bargain** him down or something
PB: ***where*** *we* **must** go in
AM: somehow
PB: **yeah** *we*'ll **get** it sorted

Together, these features shape the characteristic texture of processual discourse: fragmented, collaborative, and responsive to the emergent demands of real-time communication where speakers negotiate meaning through continuous interactive adjustment. In contrast to the conversational excerpt, this legal text exemplifies the reified pole through its densely packed nominalizations, extensive use of prepositions, and prevalence of abstract nouns. References are formal, explicit, and detached, prioritizing precise content specification over interpersonal engagement or interactive negotiation. The result is discourse designed for autonomous comprehension by unknown, potentially adversarial readers across time and space.

Text Sample 2: Negative Dim. 1

[Features in the text sample: **Preposition**, *Regular nouns*, *Nominalisation*, *Abstract nouns*, Attributive adjective, Average word length, ***Definite article***, *Proper noun*]

Legal Document:
Studien- und *Prüfungsleistungen* sowie *Studienzeiten*, ***die*** **in** *Studiengängen* **an** staatlichen oder staatlich anerkannten *Hochschulen* **im** In- und *Ausland* sowie **an** *Berufsakademien* ***der*** *Bundesrepublik Deutschland* erbracht worden sind, werden anerkannt, sofern **hinsichtlich** ***der*** erworbenen *Kompetenzen* kein wesentlicher *Unterschied* **zu** ***den*** *Leistungen* besteht, die ersetzt werden.

Translation:
Study and examination achievements as well as periods of study, completed in programmes at public or state-recognized universities in Germany and abroad or at vocational academies of the Federal Republic of Germany, are recognized provided that with respect to the acquired competences there is no substantial difference from the achievements they replace.

Both processuality and reification are aspects of immediate and distant communication. Immediacy is a style of discourse shaped by the demands of spontaneous, face-to-face interaction, where language production and planning occur in real time. In contrast, distance communication is characterized by detached discourse that prioritizes informational precision and density. The fundamental contrast lies in whether discourse prioritizes managing real-time social interaction or conveying planned, elaborated, finished content — a distinction that transcends the simple spoken/written distinction. Building on this framework, this study argues that the first dimension captures these particular aspects of the immediacy–distance axis, while the second dimension emphasizes other aspects of this continuum.

3.2 Dimension 2: Involved vs. Informational Production

Factor 2 (Table 4) comprises a total of 15 linguistic characteristics, with eight positively loaded and seven negatively loaded features. Positively loaded features primarily include personal pronouns, imperative verbs, possessive articles, interrogative pronouns, negation particles, and indicative present-tense verbs and second subordinate clauses. Overall, these features highlight an interpersonal and interactive orientation.

Positive Loadings	
2nd Person Pronoun (pper2)	.62
1st Person Pronoun (pper1)	.59
Imperative (vimp)	.58
Possessive article (artposs)	.52
Interrogative pronoun (pinter)	.42
Indicative present (vipres)	.38)
Verb 2nd subordinate clause (v2s)	.30)
Negation particle (ptkneg)	.32)
Negative Loadings	
Definite article (artdef)	-.52
Indefinite article (artindef)	-.34
Conjunctional adverb (advkonj)	-.40
Demonstrative article (artdem)	-.36)
Preposition (prep)	-.34)
Attributive adjective (adja)	-.47)
Discourse marker (diskmark)	-.34)

Table 3: Dim. 4 Loadings: Involved vs. Informational Production

On the positive pole, the most prominent group includes person markers: first and second person pronouns, which are typically associated with direct address and self-reference. Imperatives and interrogative pronouns define the dimension as dialogic, reflecting speaker's intent to prompt action and elicit responses. Possessive determiners contribute to the establishment of shared reference between speaker and listener. The use of the indicative present tense suggests temporal immediacy and congruence with ongoing events, while verb-second subordinate clauses are characteristic of real-time discourse planning. Negation particles signal dissent, correction, or evaluation, and are commonly found in argumentation or informal rejection.

Functionally, the positive pole encodes a discourse style based on interpersonal engagement and dialogic exchange. These features support speaker–addressee interaction and real-time coordination of meaning. While they are common in spontaneous exchanges, they also define the language of edited media discourse, particularly formats that simulate or dramatize interaction.

Registers such as comedy, scripted series, theater, and popular music frequently draw on these features to simulate involvement with an imagined audience (Figure 5). In documentaries and TED talks, these features are used to maintain connection with the viewer or listener. Across these registers, language is a resource for engaging audiences through directness and interpersonal involvement.

In contrast, the negative pole is associated with informationally oriented, detached, and structured discourse, represented by features such as definite, indefinite, and demonstrative articles, attributive adjectives, conjunctional adverbs, prepositions, and discourse markers. These features contribute to dense nominal structures, logical organization, and textual coherence. The reliance on prepositions and attributive adjectives supports abstract description and precise specification, while conjunctional adverbs and discourse markers indicate explicit logical relationships across clauses and paragraphs. Articles and demonstratives signal referential tracking in contexts where shared knowledge cannot be assumed.

This discourse configuration is characteristic of registers that prioritize content over interpersonal engagement (Figure 5). The R^2 value for this dimension shows the register distinctions capture a sizable portion of the variation (73.8%).

The most distinctive registers are dissertations, legal documents, textbooks, and academic articles, where communication is oriented toward the transmission of knowledge. In these texts, the language is shaped by the need for accuracy and objectivity rather than interaction. The negative pole thus marks a discourse type based on abstraction and extended informational elaboration.

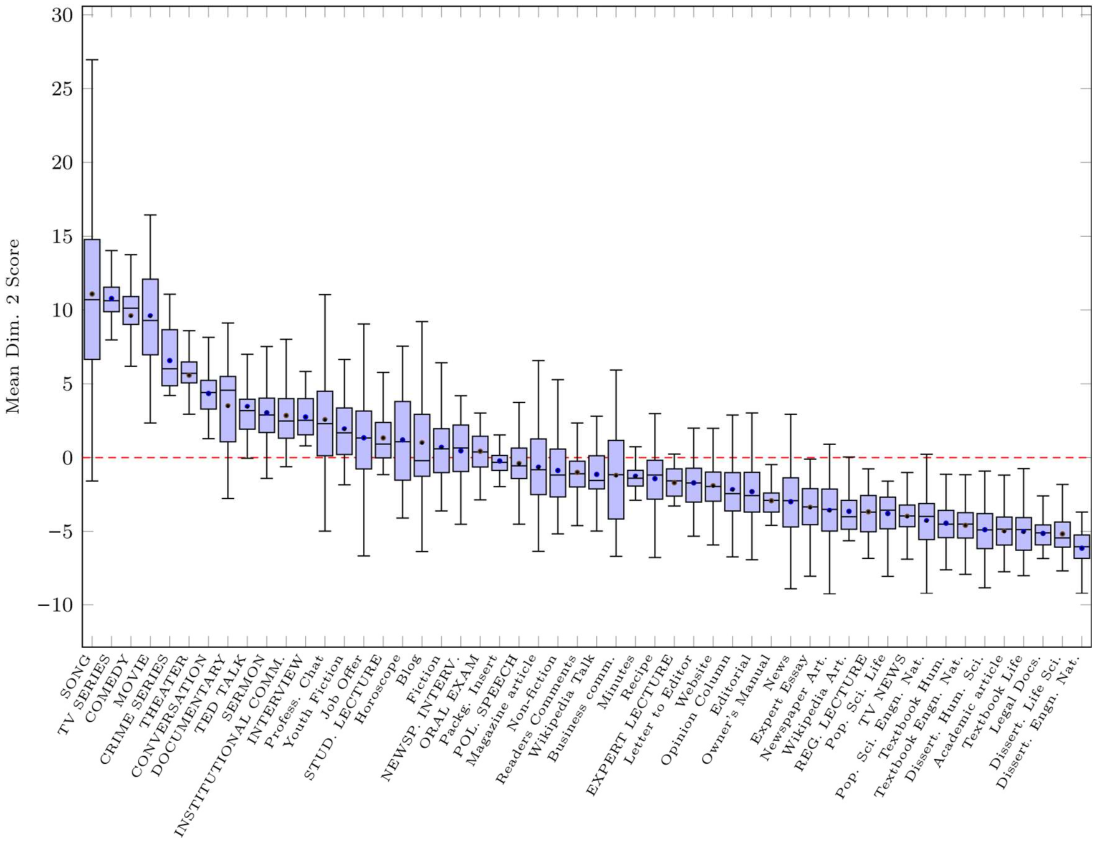

Figure 5: Mean Dim. 2 Scores by Register (*R* ^{2} =.738)

Functionally, the opposition between involved and interaction-oriented discourse from informational and carefully edited texts represents a fundamental communicative contrast frequently noted in linguistic analyses of register variation (Biber, 1988). The involved pole emphasizes interpersonal relationships and speaker-addressee oriented expression, whereas the informational pole prioritizes precise communication of content and detailed information transfer.

Text Sample 3: Positive Dim. 2

[Features in the text sample: **1st Person Pronoun**, **2nd Person Pronoun**, *Imperative*, *Possessive article*, Indicative present]

Song Lyrics:
Yeah, Baby *nimm’* *meine* Hand **Ich**
hab alles schon gepackt *Komm* **wir**
beide gehen weg von hier
Sieh der Jet **ist** getankt
Ich hab Geld auf der Bank
Und noch jede Menge Plätze hier
Und immer wenn **Du** einsam **bist**
Komm’ **ich** rum’ **Du** musst nie wieder alleine sein
Denn immer wenn **ich** **Dich** seh’
Macht es in **mir** Tick Tick Boom
So wie Dynamite

Translation:
Yeah, baby take my hand
I have already packed everything
Come on we both go away from here
See, the jet is fueled
I have money in the bank
And still plenty of spots here
And whenever you feel lonely are
Come I around you must never be alone again
Whenever I see you
It makes me go Tick Tick Boom
Just like dynamite

The example in Text Sample 3, taken from song lyrics, illustrates the linguistic and functional properties of the positive pole of Dimension 2. The discourse is explicitly interpersonal, built around direct address and emotional engagement with the addressee. First and second person pronouns (e.g., *ich*, *du*, *mir*, *meine*) construct a dyadic relationship by foregrounding personal reference and mutual involvement. Imperative forms such as *nimm’*, *komm*, *sieh* establish an interactional structure in which the speaker prompts action and positions the listener as the one expected to respond. These directives presuppose a conversational frame in which meaning is jointly negotiated in real time, even when the dialog is fictionalized or stylized.

Present-tense indicative verbs (e.g., *hab*, *gehen*, *ist*, *musst*) convey immediacy by placing the discourse in an unfolding moment. This concentration on the present gives the utterance a sense of spontaneity and personal relevance. Possessive determiners such as *meine* reinforce referential cohesion by tying the utterance to the speaker’s perspective. Together, these linguistic choices present a communicative scene in which the speaker appears to address another person directly, even if the interaction is imagined. Although the utterances are carefully constructed and embedded in a particular mass media format, they replicate the interactional patterns typical of conversation and dialogic exchange. In this way, song lyrics show how media genres use interpersonal features to simulate involvement and create a sense of immediacy and audience-oriented expression (Delfino et al., 2023).

In contrast, the excerpt in Text Sample 4, taken from an engineering dissertation, illustrates the linguistic profile associated with the negative pole of this dimension. The discourse is informationally dense and oriented toward exposition rather than interaction. Referential precision is achieved through frequent use of definite and indefinite articles (*die*, *der*, *des*, *einem*), which introduce and track entities. Attributive adjectives such as *konventionelle*, *iterativen*, *virtuellen*, and *eingestellten* contribute to compact, noun-centered constructions. These modifiers allow the encoding of conceptual distinctions in single nominal groups. Prepositions (*aus*, *mit*, *von*, *zu*) organize spatial, logical, and procedural relationships, supporting linear exposition of technical content. Conjunctional adverbs (*anschließend*, *also*) mark discourse structure explicitly by signaling logical progression and explanatory transitions across clauses.

Rather than addressing a listener, the author focuses on presenting procedures and defining relationships among concepts. The absence of person markers or extended verbal constructions reinforces the detachment of the discourse. Verb phrases are minimal, although nominal structures carry most of the informational content. This passage illustrates the type of discourse found in academic and technical writing, in which language serves the transmission of knowledge through abstraction and careful elaboration. It reflects the properties of the negative pole of Dimension 2, which keeps its distance from the interpersonal involvement that characterizes the positive pole

Text Sample 4: Negative Dim. 2

[Features in the text sample: **Definite article**, ***Indefinite article***, *Conjunctional adverb*, Preposition, *Attributive adjective*]

Engineering Dissertation:
Anschließend folgt **die** Therapieplanung. **Die** *konventionelle* Vorwärtsplanung besteht aus ***einem*** *iterativen* Prozess, **der** mit **der** *virtuellen* Therapiesimulation, *also* **der** Festlegung von Bestrahlungsparametern, beginnt. Zu **den** Bestrahlungsparametern gehören **die** Geometrie **des** zu *bestrahlenden* Volumens, **die** Strahlenergie, **die** *unterschiedlichen* Einstrahlrichtungen, **die** Anzahl **der** Bestrahlungsfelder und **die** *jeweiligen* Feldformen. Aus **den** *eingestellten* Parametern lässt sich **die** daraus *folgende* Dosisverteilung berechnen.

Translation:
Subsequently, the therapy planning follows. The conventional forward planning consists of an iterative process which, with the virtual therapy simulation — that is, the definition of irradiation parameters — begins. Among the parameters are the geometry of the volume to be irradiated, the beam energy, the different beam directions, the number of beam fields, and the respective field shapes. From the set parameters, the resulting dose distribution can be calculated.

3.3 Dimension 3: Overt Expression of Persuasion

This dimension has only one interpretable pole because the negative pole contains too few features for coherent interpretation. The positive pole is characterized by constructions such as *zu*-infinitives, reflexive pronouns, relative clauses, non-finite subordinating clauses (for example *um zu* or *ohne zu*), demonstrative articles, abstract nouns, attributive indefinites, and conditional clauses (Table 5). These features fall into two functional groups. One group consists of referential resources, including abstract nouns together with determiner forms. A second group consists of cohesive devices that connect stretches of text through logical–semantic relations. These include conjunctions that introduce subordinate clauses and infinitive constructions with *zu*. Conjunctions and pronominal adverbs also contribute to cohesion by linking segments of discourse and indicating relations such as cause or time sequence or conditional meaning.

Positive Loadings	
Zu infinitive (vinfzu)	.68
Reflexive pronoun (prefl)	.54
Relative pronoun (prel)	.54
Subordinating with zu Infinitive (nebzuinf)	.42
Demonstrative article (artdem)	.40
Abstract nouns (nabs)	.38
Attributive indefinita (attindef)	.36
Conditional clause (kond)	.33
Pronominal adverb (proadv)	.30
(Possessive article (artposs)	.41)
(Modal verb (vmod)	.39)
(Subordinating with finite verb (nebfin)	.47)
Negative Loadings	
(Conversational particle (gespptk)	-.40)

Table 5: Dim. 3 Loadings: Overt Expression of Persuasion

Abstract nouns refer to general concepts such as actions, processes, states, properties, relationships, and measurements (Dudenredaktion 2022: 698), facilitating an impersonal tone suited to abstract reasoning and concise conveyance of complex concepts. Demonstrative articles single out particular referents, serving a determining function (Berber Sardinha et al. 2014: 48; Dudenredaktion 2022: 749). Possessive articles specify ownership or association (Berber Sardinha et al. 2014b: 44), personalizing arguments or connecting abstract ideas to the audience. Attributive indefinite articles refer to indefinite quantities (Dudenredaktion 2022: 731), promoting generalization or ambiguity through reference to unspecific groups.

Reflexive pronouns refer to occurrences within the same sentence or clause (Dudenredaktion 2022: 730), while pronominal adverbs operate at a broader textual level, pointing deictically to specific elements or referring anaphorically to previously mentioned content (Dudenredaktion 2022: 804). This anaphoric function ensures discourse cohesion and textual continuity without repetitive nominal phrases.

The high frequency of zu-infinitives reflects their systematic co-occurrence with subordination structures, abstract nouns with haben (have), and pronominal adverbs serving as correlates (Buscha et al. 2017: 191; Dudenredaktion 2022: 119). Combined with subordinating conjunctions like um (in order), ohne (without), or (an)statt (instead), they express goals, intentions, alternatives, or unfulfilled expectations (Buscha et al. 2017: 183).

Subordinating clauses, especially conditional clauses, establish relationships between main and subordinate clause content marked by adverbial conjunctions expressing temporal, causal, conditional, or concessive relationships (Dudenredaktion 2022: 166), structuring reasoning or explanations. Modal verbs communicate possibility, necessity, or obligation, expressing speaker stance.

These features collectively characterize public-facing, persuasive monologic discourse, most prominently horoscope and political speeches. They make it possible for the writer to present complex arguments and hypotheses, along with generalizations that characterize the communicative aims of particular registers. By combining abstract nouns, intricate subordination, cohesive referential devices, and modals, such texts achieve both explicitness and independent comprehensibility, addressing an anonymous and heterogeneous audience.

This dimension overlaps significantly with previous MD studies, sharing features such as relative pronouns, abstract nouns, demonstrative articles, and subordinating clauses with Dimension 2

(Argumentation) of the Portuguese MDA (Berber Sardinha et al. 2014), where horoscopes and political speeches also score highest. It also resembles Biber's (1988) English MDA Dimension 4 (Overt Expression of Persuasion) in features like infinitives, modal verbs, and conditional subordination. Despite differing registers across studies, Biber's functional characterization effectively captures this dimension's communicative essence.

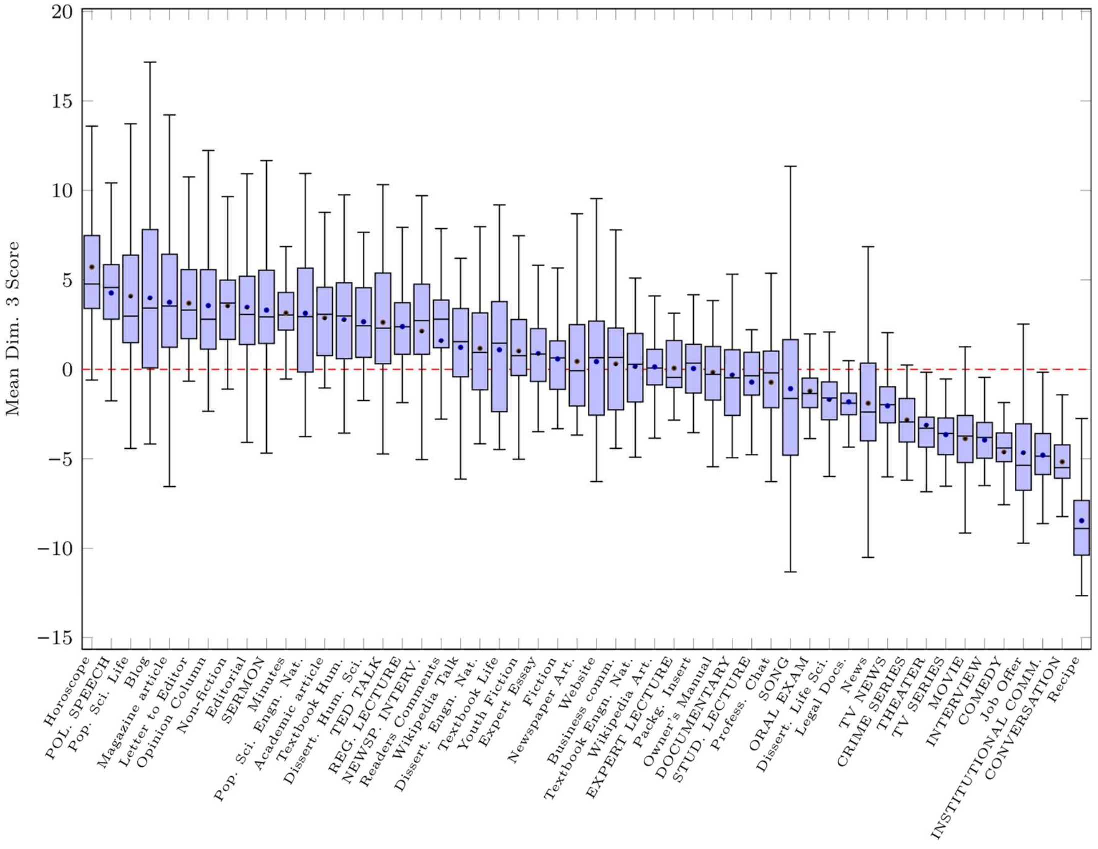

Figure 6: Mean Dim. 3 Scores by Register (R^2 =.544)

These registers are characterized by their shared goal of influencing audiences and persuading readers toward particular perspectives, beliefs, or actions. Political speeches shape public opinion through reasoned persuasion, while horoscopes influence readers' interpretations and guide decision-making through abstract, generalized advice. Accordingly, Dimension 3 is labeled 'Overt Expression of Persuasion,' encompassing both the argumentative structure of political discourse and the influence-oriented abstraction of horoscopes, the highest-scoring registers (see Figure 6), reflecting their public-oriented, persuasive nature. Conversely, recipe, conversation, and institutional communication rank lowest, highlighting their limited use of reference tracking and clause linking features. This excerpt shows how persuasive discourse appears in horoscopes:

Text Sample 5: Positive Dim. 3

[Features in the text sample: **Zu infinitive**, ***Subordination with zu-infinitive***, *Relative pronoun*, Abstract nouns, **Attributive indefinite pronoun**]

Horoscope:
[...] Genauso hat **jeder** bestimmte Schwächen, *die* es gilt **zu kennen**, ***um an ihnen zu arbeiten***. ***Um dich besser kennenzulernen und dir so die Möglichkeit zu geben***, dein Schicksal selbst in die Hand **zu nehmen**, findest du im Anschluss eine ausführliche Beschreibung mit deinen typischen Sternzeichen Eigenschaften. Ein Partnerhoroskop findest du hier.

Translation:
[...] Everyone also has certain weaknesses that must be recognized in order to work on them. To get to know yourself better and give you the opportunity to take your destiny into your own hands, you will find a detailed description of your typical zodiac sign traits below. You can find a partner horoscope here.

Dimension 3 accounts for 54% of the variance, distinguishing between different types of text based on their use of referencing and cohesive devices. High-scoring registers, such as horoscopes and political speeches, are characterized by their public orientation and persuasive intent, which are attributes made possible by planning and revision, as well as the separation between production and reception. Conversely, registers that score low, such as conversation, institutional communication and comedy, exhibit limited use of these linguistic features, reflecting their more spontaneous and interactive nature, as well as their reduced emphasis on elaborate persuasive strategies.

3.4 Dimension 4: Evidentiality strategy

Dimension 4, summarized in Table 6, focuses on evidentiality strategies (Aikhenvald 2006) and comprises seven features, predominantly with positive loadings (six out of seven). This dimension captures how speakers convey the source of information, primarily through markers of reported speech. Features on the positive pole include indirect speech forms such as Subjunctive 1 and verbs of communication, which indicate that a proposition is being reported or inferred rather than directly asserted. Subjunctive 1 enables speakers to express claims indirectly and avoid full commitment to the statement. The result is epistemic distance: speakers do not commit to the truth value of the proposition they are reporting. Verbs of communication and proper nouns link claims to external sources.

Positive Loadings	
Subjunctive 1 (vspres)	.62
Proper noun (eignam)	.53
Modal verb Subjunctive 1 (vmspres)	.47
Verb of communication (kommv)	.45
Verb 2nd subordinate clause (v2s)	.42
Comissiva (komsv)	.32
Negative Loadings	
Coordinating conjunction (kokonj)	-.50

Table 6: Dim. 4 Loadings: Evidentiality Strategy

The grammatical encoding of evidentiality in German reported speech involves distinctive interplay between clause structure and mood selection. Subordinate clauses introduced by dass (that) can be realized as verb-second clauses without explicit subordinating conjunctions when the main verb expresses attitudes, reporting, or preferences (Dudenredaktion 2022: 164). This structural flexibility reflects Aikhenvald's (2006: 111) "de-subordination," where complement clauses function as main clauses with reported speech meaning.

This structural variation depends on verb semantics. Non-factive verbs of speech or thought do not commit to reported information's truth, creating grammatical space for both indicative and subjunctive moods in embedded clauses. Verb-second clauses, being less overtly marked as dependent than verb-final clauses, more readily permit subjunctive mood (Dudenredaktion 2022: 242).

Indirect speech and subjunctive moods are particularly prevalent in journalistic genres like TV news and newspaper articles, reinforcing neutrality and clearly attributing information to sources. The subjunctive marks reported statements, distinguishing them from the reporter's original utterances and delegating responsibility (Aikhenvald 2006: 108). The indicative mood signals the author's or society's approval of the reported viewpoint, while the subjunctive creates distance, indicating the journalist does not guarantee the information's truthfulness (ebd.). Figure 7 confirms this pattern, showing TV news and newspapers consistently score highest on this dimension due to their systematic use of indirect reporting.

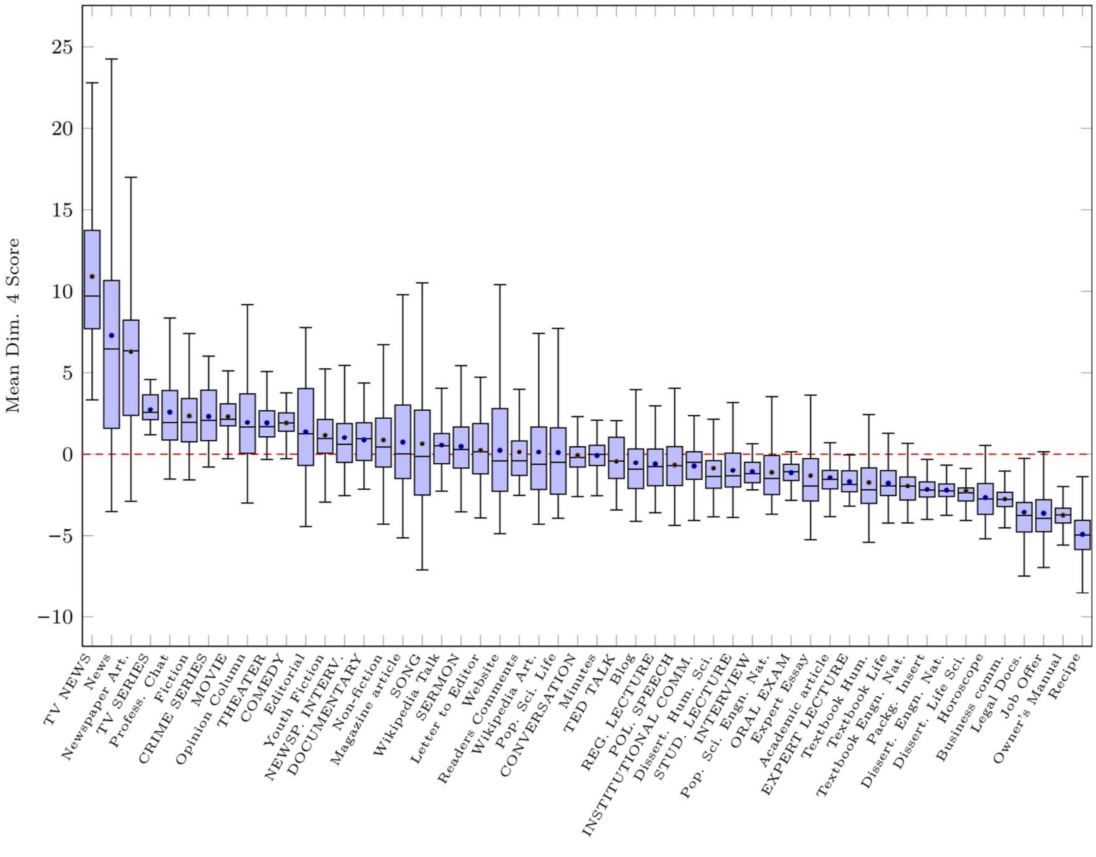

Figure 7: Mean Dim. 4 Score by Register (R^2 =.53)

Conversely, genres such as recipes, legal documents and instruction manuals tend to rank lower as they favor direct statements with minimal source marking. The R^2 value for this dimension corresponds to 53% of the explained variation, which is lower than earlier dimensions, but it still shows a clear difference between the evidential and non-evidential registers. This highlights the functional and stylistic distinction present in various communicative contexts. Example (6) shows the use of these features in spoken news:

Text Sample 6: Positive Dim. 4

[Features in the text sample: *Proper noun*, **Communication verb**, Verb 2nd subordinate clause, ***Subjunctive 1***]

TV News
Der deutsche Außenminister *Frank-Walter Steinmeier* **erklärte**, der Sicherheitsrat ***habe*** noch einmal bestätigt, was schon lange die Position der Bundesregierung ***sei***: Der Siedlungsbau in den besetzten Gebieten ***behindere*** die Möglichkeit eines Friedensprozesses.
[...]

Translation
German Foreign Minister Frank-Walter Steinmeier stated that the Security Council had once again confirmed what had long been the position of the German government: settlement construction in the occupied territories was hindering the possibility of a peace process. [...]

In summary, Dimension 4 focuses on evidentiality strategies and features, particularly those that indicate reported speech through indirect forms. Indirect speech and the subjunctive mood are particularly prevalent in press genres, where they fulfil the essential journalistic functions of maintaining neutrality and clearly attributing sources. Because press registers are designed for a broad and heterogeneous audience with no shared time or place of production and reception, careful revision and editing are necessary to ensure accuracy and clarity.

3.5 Dimension 5: Narrative Discourse

Dimension 5 has three features on the positive side and only one on the negative, which could not be interpreted because it lacked enough defining features. The positive pole captures narrative discourse, primarily through the indicative past (Table 7), which establishes a past temporal reference and allows for the sequential unfolding of actions. As a marker of narrative action, the past tense enables the construction of storylines and temporal progression. Modal verbs in the past reinforce this pattern by expressing what was possible, necessary, or intended at a given moment in the past. Temporal clauses specify the timing of events and establish relationships between them, supporting a coherent flow of narrated action. The R^2 value for this dimension corresponds to 49% of the explained variation, which suggests a lower though reasonably stable distinction among the registers.

Positive Loadings	
Indicative past (vipast)	.91
Modal verb Indicative past (vmipast)	.50
Temporal clause (temp)	.43
Negative Loadings	
(Indicative present (vipres)	-.50)

Table 7: Dim. 5 Loadings: Narrative Discourse

As a result, the dimension encodes narrative discourse centered on event sequencing and temporal structure. These features work together to mark actions as part of a developing scenario, typical of genres such as fiction, youth fiction, and narrative essays, which appear at the top of the register distribution (Figure 8). Example 7 shows a series of past-tense verbs (*stand*, *trug*, *hielt*, *gehörte*) that create a step-by-step narration, while the temporal clause introduced by *während* binds simultaneous actions into a single scene.

In contrast, registers on the negative pole such as institutional communication, oral exams, legal documents, and conversation focus on real-time interaction or exposition, where the absence of past-time reference results in discourse that is more descriptive or procedural.

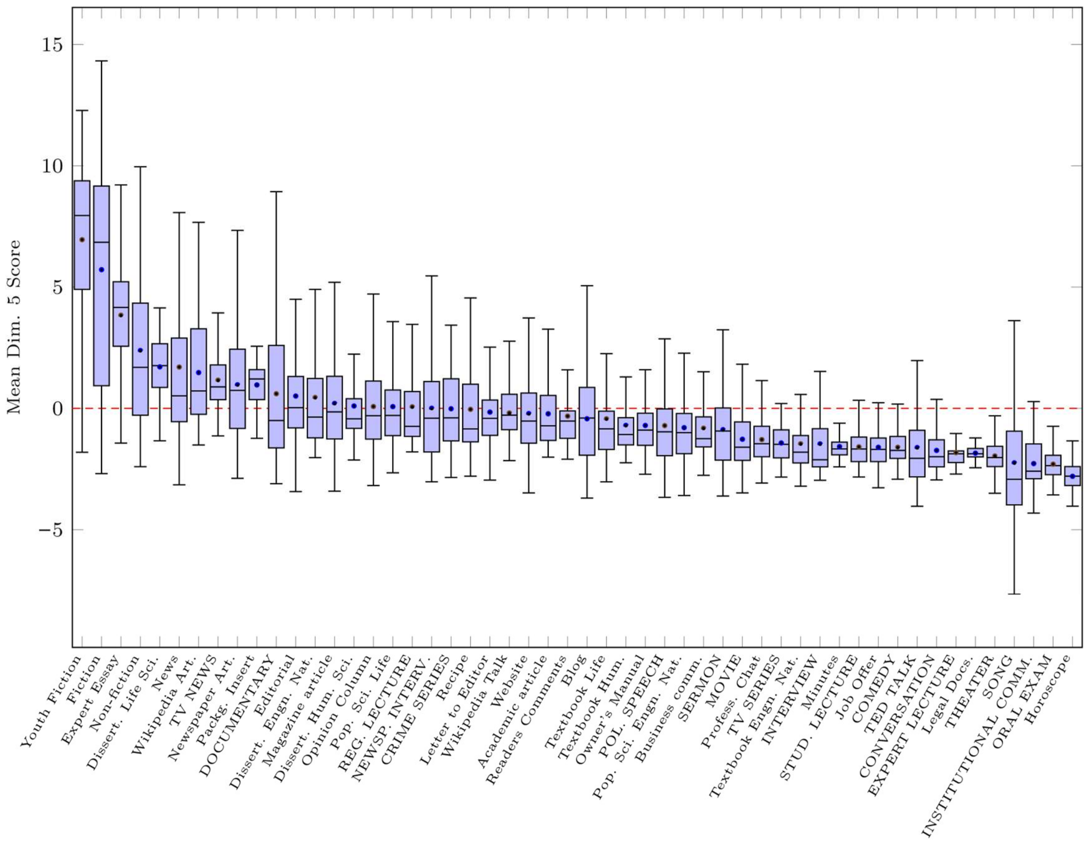

Figure 8: Mean Dim. 5 Score by Register (R^2 =.49)

Text Sample 7: Positive Dim. 5

[Features in the text sample: **Indicative past**, *Temporal conjunction*]

Youth Fiction:
Hier **stand** ich also völlig erschöpft im Schnee, *während* vom Ballsaal Violinenklänge zu uns **hinüberwehten**. Um meinen Hals **trug** ich einen Diamanten von fünfunddreißig Karat, der mir nicht **gehörte**, und in meinen Armen **hielt** ich ein schlafendes Kleinkind, das mir ebenfalls nicht **gehörte**. Irgendwo unterwegs **hatte** ich einen Schuh verloren.

Translation
There I stood, completely exhausted in the snow, while violin music drifted over to us from the ballroom. Around my neck I wore a diamond of thirty-five carats, which did not belong to me, and in my arms I held a sleeping infant, which also did not belong to me. Somewhere along the way, I had lost a shoe.

Thus, Dimension 5 captures discourse organized around the chronological structuring of events, characteristic of narrative discourse. The positive pole is marked by features which together establish a temporal framework that supports the unfolding of actions over time. This enables the construction of coherent event sequences situated in the past, allowing the speaker to structure experience retrospectively. In contrast, the negative pole includes only a single weakly loading feature and could not be meaningfully interpreted as a distinct pole. The overall contrast in this dimension is between texts that reconstruct past realities through time-based linkage and those that lack overt temporal structuring, such as present-tense procedural or expository texts that present information as general or ongoing.

4. Discussion

We compare the five German dimensions with those previously described for major European languages, namely English (Biber, 1988), Portuguese (Berber Sardinha et al., 2014a, 2014b), and Spanish (Biber et al., 2006; Parodi, 2007), considering both linguistic features and communicative functions.

Both German Dimension 1 (Processuality vs. Reification) and English Dimension 1 (Involved vs. Informational Production) capture a fundamental contrast, separating highly involved, spontaneous, and processoriented communication from detached, planned, and informationally dense discourse. In English, this dimension reflects two underlying communicative parameters: whether the primary purpose of discourse is informational or interactional, and whether the production circumstances allow careful editing or require real-time planning. German expresses a parallel contrast, but this is interpreted through the lens of immediacy and distance, reflecting whether discourse is shaped by the situational demands of face-to-face interaction or by the pressures of planning, revision, and informational precision.

The positive poles of both dimensions mark discourse produced under real-time constraints. In English, high-loading features such as private verbs, contractions, THAT-deletion, present-tense verbs, second-person pronouns, and analytic negation signal an interactional, affective, and fragmented style typical of spontaneous communication. German shows a comparable orientation toward immediacy, but the defining linguistic devices differ: conversational particles, discourse markers, intensity particles, adverbs, and predicative adjectives are the most prominent features. Whereas English relies on syntactic reduction and personal reference to encode involvement, German emphasizes meta-communicative management and interpersonal coordination. The German features regulate attention, facilitate turn-taking, and express evaluation, underscoring the real-time negotiation of meaning central to processual discourse.

The negative poles of the two dimensions are even more closely aligned. Both represent discourse that is nominal, dense, and informationally elaborated. In English, high-loading negative features include nouns, word length, prepositions, type–token ratio, and attributive adjectives, i.e. elements that support precise lexical specification and tightly integrated structures. German shows nearly identical patterns: prepositions, regular nouns, attributive adjectives, nominalizations, and longer word forms all contribute to a compact, elaborated style. These features enable the reification of content, concentrating information into dense nominal structures and reducing the visibility of the speaker. In both languages, this pole is associated with planned, distanced, and informationally focused discourse, where linguistic choices prioritize explicitness, precision, and autonomy over interpersonal involvement.

German Dimension 2 and English Dimension 1 are labeled similarly, as both distinguish involvement from information conveyance, although they realize this contrast through different sets of linguistic features. In English, involvement is signaled primarily through features associated with spontaneous or conversational production: private verbs, contractions, THAT-deletion, present-tense verbs, second-person pronouns, analytic negation, and related markers of syntactic reduction. These features reduce surface form, foreground speaker stance, and capture the fragmented character of real-time discourse. By contrast, German Dimension 2 constructs involvement around features that manage the speaker-addressee relationship more explicitly, through high loadings for first- and second-person pronouns, imperatives, possessive determiners, interrogative pronouns, and present-tense verbs, which define the dimension as dialogic and immediately oriented toward a concrete interlocutor. Whereas English involvement foregrounds production constraints and affective expression, German involvement foregrounds interpersonal coordination, directive force, and mutual orientation between interactional participants.

The negative poles of English Dimension 1 and German Dimension 2 also converge functionally, both representing informationally focused and detached discourse. English exhibits this through nouns, word length, prepositions, type-token ratio, and attributive adjectives, i.e. features that contribute to high information density and precise lexical choice. German Dimension 2 relies on prepositions, attributive adjectives, definite and indefinite articles, demonstrative articles, and conjunctional adverbs to enable detailed referential specification, logical structuring, and explicit discourse organization. In both languages, the negative pole represents discourse where the primary communicative goal is the careful integration and transmission of information rather than the management of interpersonal relations.

German Dimension 3 and English Dimension 4 also share a similar interpretive label (Overt Expression of Persuasion), as both identify the linguistic resources that signal argumentative intent, express stance, and guide an audience toward particular interpretations or actions, even though the specific grammatical differ. In English, Dimension 4 reflects the explicit marking of a speaker's point of view through modal verbs, suasive predicates, infinitival constructions, and conditional subordination, all of which help articulate obligation, possibility, or the desirability of certain states of affairs. The German dimension targets the same communicative purpose but operates on a different set of linguistic characteristics, such as *zu*-infinitives, abstract nouns, reflexive and relative pronouns, and multiple forms of syntactic subordination, which support the construction of complex arguments and generalized reasoning.

Although the linguistic inventories diverge, both dimensions correspond to features that contribute to stance expression and logical-semantic relations. English relies on modal systems (prediction modals, necessity modals, and suasive verbs) to encode assessments of likelihood, obligation, or advisability. Infinitives and conditional clauses reinforce these assessments by expressing expectations or hypothetical conditions. German, by contrast, draws on a combination of verbal and nominal strategies: infinitives, conditional clauses, and various forms of subordination create hierarchical argument structures, and abstract nouns contribute to an impersonal, generalized tone suited to public persuasion. Reflexive and relative pronouns help maintain cohesion across complex, multi-clause arguments, thereby supporting the sustained development of claims and justifications.

German Dimension 5 (Narrative Discourse) and English Dimension 2 (Narrative vs. Non-Narrative Concerns) correspond to a functional opposition between narrative and non-narrative discourse that is based on temporal reference. English Dimension 2 distinguishes narrative texts by their reference to past actions, realized through past tense verbs, perfect aspect, third person pronouns, and public verbs marking reported speech. This indicates an emphasis on event structure as well as participant tracking and attribution. In contrast, German Dimension 5 centers the narrative pole on the use of past indicative and modal past verbs, as well as temporal clauses, all of which supports a temporally sequenced account of actions but do not include reference to participant types or speech reporting. Hence, although both dimensions capture timebound discourse, the English dimension constructs narrative through a combination of temporal anchoring and participant orientation.

In comparison with Portuguese, German Dimension 1 reflects the Oral vs. Literate dimension, reflecting functions of immediacy and informality in spoken discourse, indicated by discourse markers, adverbs, and indefinite pronouns. Dimension 2 closely matches Portuguese Dimension 3 (Involved vs. Informational Production), as both perform the communicative function of involvement versus detached information reporting, marked linguistically by personal pronouns, modal verbs, negation, and shorter clauses. Dimension 3 corresponds closely to Portuguese Argumentation, sharing communicative purposes such as explicit reasoning and stance-taking, indicated by causal conjunctions, stance adverbs, and extensive subordination. Dimension 4 partially overlaps Portuguese Directive Discourse, expressing authority and directive functions through communication and modal verbs. Dimension 5 partly aligns with Portuguese

Future vs. Past Time Orientation, sharing narrative and temporal sequencing functions marked by verbs of motion, third-person references, and temporal adverbs.

Compared with Spanish, German Dimensions 1 and 2 both correspond to Spanish Dimension 1 (Oral vs. Literate Discourse), reflecting functions of oral interactivity versus literate detachment, marked linguistically by discourse markers, demonstratives, personal pronouns, communication verbs, and temporal-locative adverbs. German Dimension 2 also partially matches Spanish Dimension 5 (Informational Reports of Past Events), sharing the communicative purpose of providing informationally dense accounts through third person pronouns, preterite tense, and prepositions. German Dimension 3 partly corresponds to Spanish Dimension 1 through explanatory functions marked by causal subordination, indicative mood, and mental verbs. Dimension 4 partially matches Spanish Dimension 2 (Spoken Irrealis Discourse), both dimensions expressing hypothetical and *irrealis* meanings through subjunctive verbs, complement clauses, and conditionals. Lastly, German Dimension 5 strongly corresponds to Spanish Dimension 3 (Narrative Discourse), fulfilling a narrative communicative function through past-tense forms, clitics, possessives, and third-person pronouns.

Overall, the dimensions identified for German consistently reproduce communicative functions established cross-linguistically, notably distinctions involving interactivity, informational content, explanatory clarity, evidential attribution, and narrativity. These functional similarities underscore commonalities in register variation across the languages examined, although dimensions related to evidentiality and explanation display partial rather than complete cross-linguistic correspondence.

The dimensions identified for German show correspondence with the cross-linguistic dimensions described as ‘universal’ by Biber (2014). Specifically, German Dimension 1 (Processuality vs. Reification) and Dimension 2 (Involved vs. Informational Production) both reflect the oral-literate continuum that has been documented across various languages. These dimensions share key linguistic indicators, such as pronouns, discourse markers, verbs of communication, modal verbs, negation, and interactional features typical of spoken registers.

Similarly, the German Narrative Discourse dimension (Dimension 5) mirrors the Narrative vs. Non-Narrative dimension reported in previous MD studies. This narrative dimension consistently involves linguistic markers such as past-tense verbs, third-person pronouns, temporal adverbials, and aspectual constructions, suggesting a common communicative function involving the recounting of events. Thus, the German dimensions support Biber’s proposal that certain fundamental communicative oppositions, particularly those related to interactional versus informational language and narrative versus non-narrative discourse, are commonly found across languages.

5. Conclusion

This study applied MDA to investigate register variation in contemporary German, addressing a significant gap in the empirical research on German linguistic variability. By examining 52 registers across more than 14 million words from the Koder corpus (*Korpus deutscher Register*), this research provides new insights into the systematic patterns of language use that characterize different communicative situations in modern German, extending previous studies that were limited to specific registers, historical periods, or written texts.

A factor analysis revealed systematic patterns of linguistic co-occurrence across 65 features, which were grouped into five dimensions. Dimension 1 showed a strong co-occurrence of conversational particles, discourse markers, adverbs, modal verbs, subjunctive 2, indefinite pronouns and predicative adjectives at the positive end of the scale. In contrast, prepositions, nouns, nominalizations, attributive adjectives, definite articles and long words were found at the negative end of the scale. Dimension 2 showed a positive pole of first- and second-person pronouns, imperative forms, and interrogative pronouns, and a negative pole of definite and indefinite articles, attributive adjectives, prepositions, and conjunctional adverbs. Dimension 3 showed a cluster of *zu*-infinitives, reflexive pronouns, relative clauses, abstract nouns, demonstrative articles and conditional clauses primarily on a single interpretable pole. Dimension 4 showed co-occurrence of evidential markers, including subjunctive 1, verbs of communication, proper nouns and verb-second subordinate clauses. Finally, dimension 5 displayed co-occurrence of the indicative past tense and temporal conjunctions.

The first dimension (Processuality vs. Reification) distinguishes between discourse produced under real time, interactive pressures and discourse characterized by careful planning and informational density. This dimension captures the immediacy–distance continuum theorized by Koch and Österreicher (2012). Spoken registers such as everyday conversation, interviews, and institutional communication score highest on the processual pole, while legal documents, academic texts, and news texts score highest on the reified pole.

The second dimension (Involvement vs. Information Production) emphasizes the interactive and informative aspects of language use, distinguishing between discourse that prioritizes interpersonal engagement through direct addressee reference, and discourse that emphasizes precise specification and the presentation of abstract information. Scripted spoken registers, such as song lyrics, TV series and comedy, scored highest on the involvement pole, while PhD dissertations, scientific textbooks and legal texts scored highest on the informational pole. The immediacy-distance continuum, as in Dimension 1, is also reflected in this dimension.

The third dimension (Overt Expression of Persuasion) is characterized by public, persuasive monological discourse supported by features that facilitate complex argumentation, hypothetical reasoning and generalization. Horoscopes and political speeches scored highest on this dimension, reflecting their shared objective of influencing audiences through abstract, elaborate discourse.

The fourth dimension (Evidentiality Strategy) captures how speakers convey the source of information and their epistemic stance, particularly through indirect reported speech marked by subjunctive I and verbs of communication. TV news and newspaper articles consistently scored highest, reflecting the journalistic practices of source attribution and epistemic distancing typical of German journalism. Dimension 5 (Narrative Discourse) distinguishes narrative from non-narrative registers through the contrast between past tense with temporal conjunctions and present tense. Youth literature and fiction scored highest, reflecting their fundamental narrative structure.

These findings contribute to register theory by demonstrating that, while German register variation exhibits language-specific characteristics, it also follows systematic patterns comparable to those found in other languages. Firstly, the study provides a comprehensive MDA of contemporary German, incorporating both spoken and written registers. This expands beyond Neumann's (2014) focus on written texts and Ágel and Hennig (2006) historical perspective. Secondly, the findings support a continuum-based rather than binary conceptualization of register variation, aligning with Koch and Österreicher (2012)'s immediacy–distance framework rather than simple spoken–written dichotomies. Thirdly, identifying dimensions specific to the linguistic structure of German (particularly the evidentiality dimension centered on subjunctive I) shows how grammatical resources specific to a language can shape patterns of register variation.

It is important to acknowledge the limitations of this study. Firstly, the interpretation of the dimensions identified through MDA is tentative rather than definitive. The interaction between linguistic features and their functional interpretations allows for multiple valid perspectives. This flexibility reflects the complexity of linguistic variation, and alternative interpretations of statistical patterns are possible. The corpus is extensive and diverse, but some registers may be underrepresented or absent. The 52-register selection cannot cover all contemporary German communicative situations. Research could benefit from expanding the corpus to include emerging forms of digital communication and specialized professional registers. Social network analysis could reveal how register variation relates to social structures and community practices. Thirdly, while the feature set analyzed was extensive, it may not capture all aspects of register variation. Due to technical limitations of the available tools, certain characteristics typical of spoken German were not included in the analysis. Fourthly, methodological constraints affected feature selection and analysis. None of the tools used could account for all the relevant linguistic features. This meant that multiple tools had to be used, and semi-automatic corrections were necessary. Although this process ensured greater accuracy than relying on a single tool, it may have introduced inconsistencies. This limitation highlights an important area for future work in German NLP: the need for a single, unified toolset capable of handling a wider range of linguistic features with improved accuracy for both spoken and written German. Despite these limitations, this study shows that MDA provides a solid empirical basis for describing register variation in modern German. The identification of five dimensions of variation offers a foundation for understanding how German speakers adapt their language use across contexts. We hope the current study serves as a baseline for future research.

Acknowledgments

This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), under PNPD Grant #88887.352835/2019-00; the São Paulo Research Foundation (FAPESP), under Grant #2022/05848-7; and the National Council for Scientific and Technological Development (CNPq), under Grants #420180/2022-2, #403130/2024-7, #444019/2024-3, #310140/2021-8, and #403834/2022-8.

References

Ágel, V., & Hennig, M. (2006). *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650–2000*. Max Niemeyer.

Aikhenvald, A. Y. (2006). *Evidentiality* (Repr). Oxford Univ. Press.

Apothtloz, D., Brandt, P.-Y., & Quiroz, G. (1993). The function of negation in argumentation. *Journal of Pragmatics*, (19), 23–38. Retrieved October 3, 2025, from https://perso.atilf.fr/apotheloz/wpcontent/uploads/sites/59/2015/06/JPragmArgum.pdf

Barbaresi, A. (2012). German political speeches, corpus and visualization. Retrieved November 15, 2016, from http://purl.org/corpus/german-speeches

Beißwenger, M., Herold, A., Lüngen, H., & Storrer, A. (2016). Das Dortmunder Chat-Korpus in CLARIN-D: Modellierung und Mehrwerte [Meeting Name: DHd]. In E. Burr & Digital Humanities im deutschsprachigen Raum (Eds.), *DHd 2016: Modellierung - vernetzung - visualisierung: Die digital humanities als fächerübergreifendes forschungsparadigma: Konferenzabstracts: Universität leipzig, 7. bis 12. märz 2016*. nisaba verlag. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/5578

Berber Sardinha, T., Kauffmann, C., & Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portuguese. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (pp. 35–80). John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). *Multi-Dimensional analysis, 25 years on: A tribute to Douglas Biber*. John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research methods and current issues*. Bloomsbury Academic.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language*, *64*, 707–736. https://doi.org/10.2307/414565

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/ cbo9780511621024

Biber, D. (1995). *Dimensions of register variation – A cross-linguistic comparison*. Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9–37.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation.

Languages in Contrast, *14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., & Conrad, S. (Eds.). (2001). *Variation in English: Multi-dimensional studies*. Longman.

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, *1*(1), 1–37. https://doi.org/10.3366/cor.2006.1.1.1

Biber, D., & Egbert, J. (2023). What is a register?: Accounting for linguistic and situational variation within– and outside of– textual varieties. *Register Studies*, *5*(1), 1–22. https://doi.org/10.1075/rs.00004.bib

Biber, D., & Hared, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change*, *4*, 41–75.

Buscha, A., Szita, S., & Raven, S. (2017). *C-Grammatik: Übungsgrammatik Deutsch als Fremdsprache* (5. Aufl.,) [Num Pages: 266]. Schubert-Verlag.

Costa, A. (2019). Koder – a multi-register corpus for investigating register variation in contemporary german. *Research in Corpus Linguistics*, *7*, 69–83. https://doi.org/10.32714/ricl.07.04

Delfino, M. C. N., Berber Sardinha, T., & Collentine, J. G. (2023). Dimensões de variação lexical e acústica na música popular em inglês: Um estudo baseado em corpus [Lexical and acoustic dimensions of variation in popular music in English: A corpus-based study]. *Cadernos de Estudos Linguísticos*, *65*, e023025. https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8671801/33021

Dudenredaktion, A. W. (Ed.). (2022). *Duden – Die Grammatik* (10., völlig neu verfasste Auflage, Vol. 4). Dudenverlag.

Gray, B., & Egbert, J. (2019). Editorial: Register and register variation. *Register Studies*, *1*(1), 1–9. https: //doi.org/10.1075/rs.00001.edi

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Arnold.

Kim, Y.-J., & Biber, D. (1994). A corpus-based analysis of register variation in korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 157–181). Oxford University Press.

Koch, P., & Österreicher, W. (2012). Language of immediacy—language of distance: Orality and literacy from the perspective of language theory and linguistic history.

Lamb, W. (2008). *Scottish Gaelic speech and writing: Register variation in an endangered language*. Cló Ollscoil na Banríona.

Lüdtke, J. (2008). *Sprachpragmatische Aspekte der Negationsverarbeitung: Bestätigen und Zurückweisen mit negativen Sätzen* [Doctoral dissertation, Technische Universität Berlin]. Retrieved October 3, 2025, from https://depositonce.tu-berlin.de/items/a70216f5-dfda-4b0b-9370-d0d2926ae523/full

Neumann, S. (2014). *Contrastive register variation: A quantitative approach to the comparison of English and German*. De Gruyter. https://doi.org/10.1515/9783110238594

Parodi, G. (2007). Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 11–53). Continuum.

Purvis, T. (2008). *A linguistic and discursive analysis of register variation in Dagbani* [PhD dissertation]. Indiana University.

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. https://www.cis.uni-muenchen.de/~schmid/papers/Schmid-Laws.pdf

Schwitalla, J. (2012). *Gesprochenes Deutsch: Eine Einführung* (4., neu bearb. und erw.). Erich Schmidt Verlag.

Sennrich, R., Schneider, G., Volk, M., & Warin, M. (2009). A new hybrid dependency parser for German. https://files.ifi.uzh.ch/cl/volk/papers/Sennrich_Schneider_Volk_Warin_Pro3Gres_GSCL.pdf

Westpfahl, S., Schmidt, T., Jonietz, J., & Borlinghaus, A. (2017). *STTS 2.0. Guidelines für die Annotation von POS – Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS)* (Edition: Version 1.1, März 2017). Retrieved December 28, 2021, from https://idspub.bsz-bw.de/frontdoor/index/index/docId/6063