The FinderApp WiTTFind for Wittgenstein's Nachlass

Maximilian Hadersbeck, Alois Pichler, Sabine Ullrich, Ines Röhrer and Nivedita Gangopadhyay

Abstract Since 2010, the Wittgenstein Archives at the University Bergen (WAB, Alois Pichler) and the Centre for Information and Language Processing at the Ludwig-Maximilians University Munich (CIS, Max Hadersbeck et al.) cooperate in the research group the "Wittgenstein Advanced Search Tools" (WAST). The WAST research group develops the web-frontend FinderApp WiTTFind together with specialized search tools for scholars in the humanities to investigate WAB's transcriptions of the Nachlass of Ludwig Wittgenstein with advanced computational search tools. Their FinderApp WiTTFind (http://wittfind.cis.lmu.de) displays facsimile-extracts on the hit-page and allows double-sided paging through the facsimile with its WiTTReader Application. In our paper, we want to present the research work around the FinderApp

Dr. Maximilian Hadersbeck · Sabine Ullrich · Ines Röhrer
The Center for Information and Language Processing, Oettingenstraße 67, 80538 München, Germany
✉ Maximilian.Hadersbeck@lmu.de
✉ Sabine.Ullrich@campus.lmu.de
✉ I.Roehrer@campus.lmu.de

Prof. Alois Pichler · Dr. Nivedita Gangopadhyay
Wittgenstein Archives at the University of Bergen, Snydesplass 12-13, 5020 Bergen, Norway
✉ Alois.Pichler@uib.no
✉ Nivedita.Gangopadhyay@uib.no

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/14
ISSN 2363-9881

WiTTFind, the WiTTReader, and our latest developments within WAST, the synonym-lexicon and the similarity search tools.

1 Introduction

During his lifetime, Ludwig Wittgenstein (1889-1951) published only one philosophical book, the *Logisch-philosophische Abhandlung / Tractatus logico-philosophicus*. However, he became one of the most important philosophers of the last century. This is mainly due to the various publications from his philosophical Nachlass (von Wright, 1969) which comprises about 20 000 pages of manuscripts and typescripts. Titles like *Philosophical Investigations* (1953), *Remarks on the Foundations of Mathematics* (1956) or *On Certainty* (1969), to mention only a few, are all posthumous publications that have contributed to Wittgenstein's fame as a philosopher of the first rank. But Wittgenstein never saw these publications himself and his literary heirs and trustees G.H. von Wright, Elizabeth Anscombe and Rush Rhees edited them from the Nachlass. Recognizing the importance of digitizing the Nachlass and providing digital access to it, the Wittgenstein Archives at the University Bergen (WAB) produced in the years 1990-2000 machine-readable transcriptions and facsimile of the Nachlass and published in 2000 on CD-ROM the Bergen Electronic Edition. Since 2016, WAB offers open online access to its entire set of transcriptions, in their continuously revised and improved form, on the "Nachlass transcriptions" site `http://wab.uib.no/transform/wab.php?modus=opsjoner` (Wittgenstein, 2016).

Obviously, digital editions can offer significant advantages over print editions in that they allow unprecedented dynamic and user-tailored access to the material, unmatched in print. One great advantage of digital editions over print editions is text search. Luckily, today we are in the position to also perform lemmatized and sentence context driven searches in the Nachlass – online, grounded in highly reliable scholarship and programming, and entirely for free. This is made possible thanks to a cooperation between WAB and the Centre for Information and Language Processing (CIS) at the Ludwig-Maximilians University Munich (LMU) which resulted in the web-frontend FinderApp WiTTFind (see Figure 1; Hadersbeck et al, 2012; Hadersbeck et al, 2018). Since the annual International Wittgenstein Symposium in Kirchberg am Wechsel (Austria) 2018, the cooperation project was finally able to present a first full version of the tool WiTTFind as it applies to the *entire* Wittgenstein Nachlass and thus offer online search in WAB's transcriptions of the entire Nachlass.

Figure 1: WiTTFind: Search for the word "Übereinstimmung" and display hit in Ts-213.

WiTTFind displays each sentence containing any grammatical form of the word searched for within the context of the larger remark and additionally highlights the hit in the corresponding facsimile of the remark. This is possible thanks to dynamic web-page interaction between the user and the more comprehensive WAST tools. In strong contrast to search methods from Google Books or the Open Library project, where standard technologies with almost no domain-specific or sentence context knowledge are used, WAST uses both specialized statistical search-methods as well as specialized rule-based syntactic text-corpus algorithms which are semi automatically optimized and adapted to their specific research questions. A digital lexicon coupled with a symmetrical autosuggestion tool enables symmetrical lexicon-based search-queries (Bruder, 2012). A double-sided facsimile-reader, WiTTReader, permits to skim comfortably through the query results in Wittgenstein's Nachlass (Lindinger, 2015).

2 WiTTFind and WiTTReader

The web-frontend WiTTFind uses several tools from WAST and offers an easy to use interface to query Wittgenstein's Nachlass. All search results are displayed as HTML-transformation of the transcribed edition-texts together with an excerpt from the original facsimile of Wittgenstein's remark (see Figure 1). Alternative readings of a remark are presented within additional tabs on the hit-page. Every query result of the Nachlass is linked to Wittgenstein Source and the WiTTReader Web-Application (see Figure 2), which offers the user a double-sided facsimile page-view of the found hits and enables easy exploration of surrounding areas of the document.

For organizing the search hits, WiTTFind utilizes an established distinction between parts of the Wittgenstein Nachlass. This distinction has been present in Wittgenstein research for a long time and actually goes back to Georg Henrik von Wright's original division and classification of the Nachlass in his seminal paper (von Wright, 1969). Georg Henrik von Wright organizes the Nachlass into three numbered groups:

1. numbers 101–182 refer to *manuscripts* (texts in Wittgenstein's hand, primarily notebooks and bound volumes),
2. numbers 201–245 refer to *typescripts* (either dictated by Wittgenstein directly from his manuscripts or typed by a typist on Wittgenstein's order), and
3. numbers 301–311 refer to *dictations* (to friends and students, e.g. in connection with lectures and seminars).

In Pichler (1994; see especially pages 232–235) further subcategorization, however not without basis in von Wright's own descriptions of the Nachlass materials of the manuscripts, was suggested:

- *Loose sheets* ("MS1"),
- *Notebooks* ("MS2"), and
- and *Bände* (engl. *volume*; MS3: large bound volumes, often ledger books).

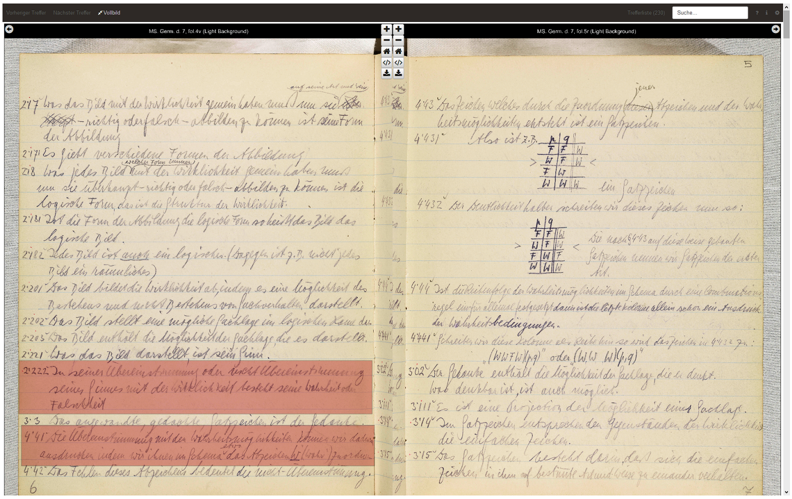

Figure 2: WiTTReader: Browsing through Ms-104.

Pichler (1994) also suggested further division of the typescripts: *Typescripts* ("TS") and *Typescript cuttings* ("TSC"). The WiTTFind organization of the search hits now mirrors parts of this more detailed structuring. Thus, the search results can be filtered according to four categories: Notebooks, Volumes (i.e. "Bände"), Typescripts and Typescript cuttings.

3 The WAST Infrastructure

The core of the WAST landscape is the web-frontend FinderApp WiTTFind which allows lemmatized, syntactic and semantic search in the Nachlass (see Figure 3). The data basis for WAST are the transcriptions and facsimiles of the twenty thousand pages of Nachlass which are provided by WAB. WAB produces XML TEI-P5 transcriptions (WAB XML) from which it releases the simplified open access XML OA. The XML edition from Bergen uses a reference system, which does not stop at page level; it continues down to Bemerkungen-level. The Bemerkung siglum is composed of a sequence of "subnames":

- the *name of the overarching Nachlass item* in which the Bemerkung (engl. *remark*) is found, prefixed by "Ms-" or "Ts-",

- the *name of the page*(*s*) on which it occurs, and
- the *name*(*s*) *for the segments* of which the Bemerkung consists.

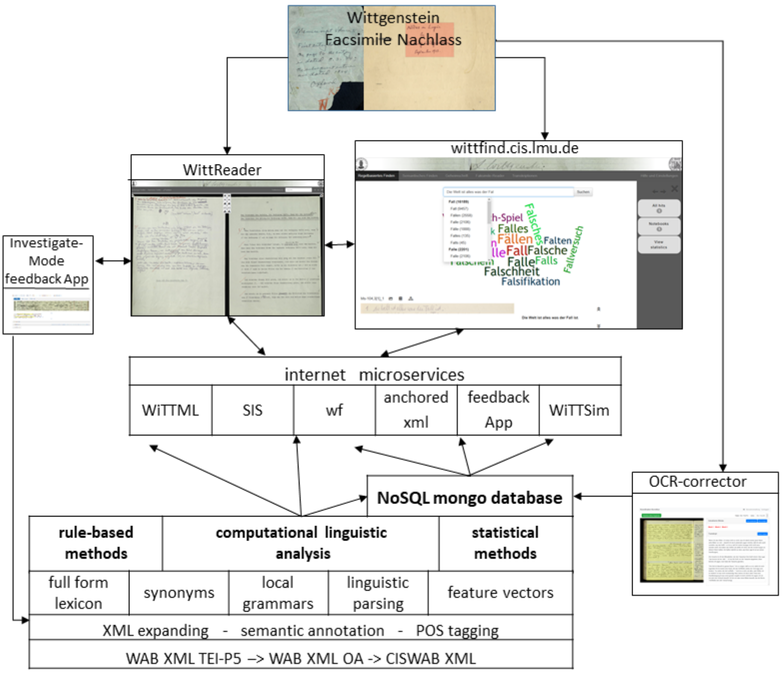

Figure 3: The infrastructure around our FinderApp WiTTFind (`http://wittfind.cis.lmu.de`).

"Ms-101" for example refers to the Nachlass item, which in the catalogue has the number 101, belonging to the class of manuscripts. "Ms-101,1r[1]" is then the siglum for a specific single Bemerkung referring to the first block of text on page 1r in Ms-101 and thus to the remark: "Vorgestern bei der Assentierung genommen worden ...", dated by Wittgenstein to August 9, 1914. A Bemerkung can extend across page breaks and sometimes continues over several pages; this too will be mirrored in the siglum. In the end, each of the more than fifty thousand Bemerkungen in the Wittgenstein Nachlass is identified through such a unique siglum. The WAB XML OA edition format is further simplified to CISWAB XML which is then used at CIS as transcription input for WAST. On the ground level of our WAST infrastructure we start the computational-analysis of the CISWAB XML data:

- fullform lexicon,
- synonyms,
- tokenizing,
- parsing,
- POS-tagging,
- syntactic expanding, and
- semantic annotation.

The first technical challenge was tokenizing and POS-tagging of the XML data. Wittgenstein's writings are far from so called "linear texts", his writings are highly dynamic: textinsertions, -deletions, -substitutions and -repositioning on character, word, sentence, remark and even on page level characterize the Nachlass-pages of Wittgenstein. These text-variations are highly-accurate XML-annotated according to the TEI-P5 suggestions at the Wittgenstein archive in Bergen. Stripping XML-tags or TEI-P5 xslt-Parsers did not produce grammatically correct sentences, ready for linguistic analysis. Together with the author Schmid, the author of the TreeTagger (Schmid, 1995), we developed a Wittgenstein-Korpus variant of the TreeTagger.

On the next level of our WAST infrastructure we implemented tools that cover functionalities ranging from semantic categorization (WiTTML) similarity search (WiTTSim), symmetric autosuggestion (SIS), xml-annotationserver (Anchored XML), search-engine (wf) and user feedback-App.

The core of our WAST infrastructure is its search-engine wf. To solve the problem of word-ambiguity we did not use an index-based search engine (i.e. lucene). As Wittgenstein states in his early works "Nur der Satz hat Sinn; nur im Zusammenhange des Satzes hat ein Name Bedeutung." (engl. "Only propositions have sense; only in the nexus of a proposition does a name have meaning."; TLP, 3.3, Wittgenstein and Ogden, 1990) , we wanted a search-engine, which operates on sentence level and has direct access to lexicon, syntax and semantic word-context information. The linguistic theory, which stands behind this kind of context driven linguistic analysis, is called "local grammar technique" (Gross, 1997) and is implemented as softwaresystem UNITEX at the Laboratoire d'Automatique Documentaire et Linguistique (Paumier, 2003). We developed our search-engine wf (Hadersbeck et al, 2014) as a high-performance, multithreaded C++ application, which lingustically analyses all sentences via preprocessing and uses local grammar techniques to find hits in the corpus.

Long programming-experience with directed acyclic word graphs (DAWG; Bruder, 2012) gave us the possibility to implement another C++ program for our infrastructure: a symmetric-DAWG. With the help of our full form lexicon we can expand entered letters during input with appropriate words from the Nachlass containing the letters as infix. The full form lexicon gives us also the possibility to enrich the presented words with its morphological variants. With the help of frequency lists from the Nachlass, we can additionally enrich the autosuggestion with frequency information. Users who are not familiar with German morphology get additional ideas for the "lemmatized" search within our FinderApp. To analyze Wittgenstein's Nachlass on the semantic level we use NLP-tools (WiTTML and WiTTSim), which are described in detail in Section 5.

4 WiTTLex and Semantic Search

Successful computational linguistic work relies crucially on the use of an electronic full-form lexicon. For the work with the Nachlass we constructed a special lexicon called WiTTLex (Röhrer, 2017). WiTTLex includes all words from the normalized BNE edition. Each word-entry in WiTTLex is formatted according to the DELA Format, defined at the Laboratoire d'Automatique Documentaire et Linguistique (Paumier, 2003). The lexicon entries contain the word's full form, lemma, and lexicographical word form, together with flexion and semantic notations for frequent words. With the help of WiTTLex search queries to WiTTFind can be processed with respect to lemmatization and grammatical category.

With WiTTLex in the background, our FinderApp can find all text passages where morphological variants of the queried word occur: The search engine accesses the lemma for the specified word, and from this lemma it obtains knowledge of all of its morphological variations. If the user specifies for example the word "sagte" (engl. *said*, singular), it finds in the lexicon the lemma "sagen" (engl. *to say*, infinitive) and from here it finds all the other morphological variants like: "sagten" (engl. *said*, first-person or second-person plural), "sagst" (engl. *say*, second-person singular) and so on. With the additional help of local grammar techniques, even separated particle-verb constructions like "fällt ... heraus" (engl. *to fall or drop out*; from "herausfallen") can be disambiguated and will be found, which is impossible for usual search engines.

To enable users to access the Nachlass conceptually, i.e. through content rather than string, word or phrase search only, WiTTFind has started to provide facilities for semantic access. Special attention was given to the word field of colors (Krey, 2014) and music (Röhrer, 2017). In the field of colors the subcategories *Grundfarbe* (engl. *basic color*), *Zwischenfarbe* (engl. *intermediate color*), *Transparenz* (engl. *transparency*), *Glanz* (engl. *gloss*) and *Farbigkeit* (engl. *colorfulness*) were applied. In the field of music the subcategories *Komponisten* (engl. *componists*), *Instrumente* (engl. *instruments*), *Gattungen* (engl. *genres*), *Intervalle* (engl. *intervals*), *Bezug zu Komponisten* (engl. *reference to composers*) and *sonstige musikalische Begriffe* (engl. *other musical concepts*) were defined (see Figure 4). A lot of interesting conclusions can be drawn from this work: beside the problem of general classification in specific domains, we showed that, for example, the category "sonstige musikalische Begriffe" is filled with a lot of ambiguous words that are hard to disambiguate. For example, the German word "Ton" can be understood as both "sound" or "clay". In such cases, simplistic positive / negative lists are of very limited use only. Ontology research and the use of local grammars could be a way to improve the results.

As one of the essential parts of the WiTTFind project, our digital lexicon WiTTLex is being continuously improved. Improvements are, for example, correction of spelling or formatting errors, incorporation of additional information in existing entries or creation and insertion of entries for newly found words. Both the Nachlass with about 20 000 pages and the lexicon with about 60 000 entries are so large that constant maintenance is necessary. The special feature of WiTTLex is that it is tailored to Wittgenstein's language and contains only words that occur in his Nachlass. Because of these characteristics, the lexicon offers a unique base for linguistic investigations in Ludwig Wittgenstein's works. To enable more detailed textual research on semantic issues, we are currently working on implementing a synonymy lexicon, called "WiTTLex Synonym", within the scope of a student research project. The lexicon is based on the word database created by WiTTLex and on the synonyms extracted from GermaNet (Hamp and Feldweg, 1997) and WordNet (Fellbaum, 1998). GermaNet is – equivalent to WordNet – a lexical-semantic word network developed at the University of Tübingen. Subsequently, this base is formatted into a structure similar to the DELA system of WiTTLex, and it is manually tested and supplemented. The gradual construction and extension of this lexicon allows an evaluation of GermaNet in the linguistic context of philosophy, where it can show for how many words synonyms could be found automatically, and the quality of the synonyms. In a second evaluation process, we can compare whether or not our final WiTTLex Synonym shows an improvement in contrast to a purely automatic system based on GermaNet and WordNet in our similarity search WiTTSim on Ludwig Wittgenstein's Nachlass.

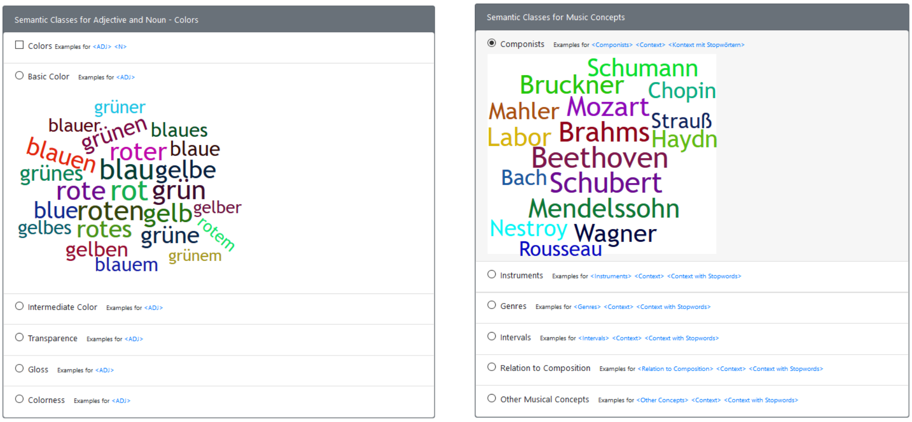

Figure 4: Semantic search for color and music.

In addition to this evaluation, the new WiTTLex Synonym will provide other interesting possibilities for digital work on the Nachlass. It will be possible to find shared synonyms between words that are not synonyms. For example, "violin" and "trumpet" are not synonyms, but "instrument" would be assigned to both. By this method, we will be able to link words together that are likely to occur in a similar context. This opens up an entirely new set of possibilities such as providing a new search option to find similar texts, providing new arrangements for texts to certain topics or improving the WiTTSim similarity search.

5 WiTTSim and WiTTML: Semantic Similarity Detection and Machine-Learning Methods

Our current project, to be integrated into WiTTFind, is WiTTSim (Ullrich et al, 2018), an NLP-based similarity search tool. The reason for developing this tool is that Ludwig Wittgenstein takes up topics and ideas from various different sources. Since he assumed a general knowledge and familiarity of the reader with common literature (like Goethe, Lichtenberg etc.), he did not bother to provide appropriate citations in his works. Apart from these external ideas, he revisited several topics from his own former works. Without indications of source, it is inherently challenging to reveal these external and internal influences, which occur in all manuscripts and typescripts. Therefore, in order to allow for similarity-based browsing, we developed a method to detect these similarities across Wittgenstein's Nachlass.

WiTTSim transforms each input text into a vector and calculates the cosine distances between the vectors using vector space models (VSMs). The program outputs a ranked list of similar remarks to a given remark. This technique (adapted from plagiarism detection) will enable users to find similar remarks to a given remark.

It is important to note that each of Wittgenstein's remarks is presented as a Bag-of-Words (BOW), which implies that it is also important that the documents to be compared (here: remarks) remain at a relatively low extent since the end of sentences will not be captured by this method and results will yield very low precision for large documents. The BOW representation helps, however, to determine the similarity between remarks where similar statements are made across sentences.

For the implementation of WiTTSim, we define linguistic features that play a major role in textual similarity. These hand-crafted features are determined for each word in the BOW representation – in other words, let $D = d_1, d_2, \ldots, d_M$ be the collection of M documents (remarks) where each document d is translated into an n-dimensional vector in the VSM. Each position of the vector represents different characteristics $C = c_1, c_2, \ldots, c_n$ of the text. These characteristics comprise the words themselves, their respective lemmata, their category (i.e. Part-of-Speech (POS) tag), as well as associated synonyms. The extracted vectors of $|D| = 54\,930$ documents comprise $|C| = 115\,601$ characteristics, in more detail 45 337 words, 22 739 lemmas, 38 POS-tags, and 47 487 synonyms.

The number of occurence of these features, i.e. their frequency $f_d(c)$, is also denoted in the vector for each remark. For extracting the features, the lemmata and POS tags are determined by means of the TreeTagger which has been chosen since it yields very fast and accurate results: the accuracy reached for German and English texts reaches 97.53 % and 96.81 %, respectively (Schmid, 1995). The utilized English and German synonyms are extracted from the ontology collections WordNet (Fellbaum, 1998; Miller, 1995) and GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), respectively. In order to extract the appropriate synonyms, a language detection is done beforehand. Moreover, the synonyms have been enriched and adapted to the research context, which has been shown in Section 4. Likewise, synonym pairs not listed in the used ontologies, such as *präsentieren* (engl. *to present*) and *darstellen* (engl. *to depict*), can be classified correctly.

After computing the features for each word and for the entire remark, the features are binarized and form the feature vector for the remark. That is, each remark is presented as a large vector in an 115 601-dimensional vector space. These computations are all done in advance since a run-time computation would be too costly and would imply long waiting periods. WiTTSim provides two functionalities:

1. *Browsing* similar remarks in the entire Nachlass, and
2. *Comparison* of two selected remarks.

However, because of the large number of dimensionalities and documents, a reduction to the most essential features from $\mathbb{R}^n$ to $\mathbb{R}^k$ where $n = 115\,601$ and $k << n$ – while keeping as much information as necessary – can be helpful. Furthermore, using document clustering methods (MacQueen, 1967; Ball and Hall, 1965; Lloyd, 1982; Steinhaus, 1955; Ester et al, 1996; Redner and Walker, 1984; Ward Jr, 1963) is planned for future work to improve efficiency and further reduce the search time.

For defining similarities, a classification of remarks beforehand can be inherently helpful. Thereby, the remarks are grouped by topic in a (semi-)supervised way. In our Wittgenstein machine learning studies (or short WiTTML), the underlying data base are 7 000 remarks that are grouped into approximately 100 pre-defined categories. The categories and the labelling have been realised by domain experts such that high quality results can be achieved. For classifying the unseen data, 90 % of the labelled data has been used for training, 10 % for development, and 10 % for testing. The tests have been performed using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Results show that due to the long siglum length and the sparse labelled data, CNNs yield better results. An integration of the document classification is planned in the near future.

Browsing similar remarks can help reveal hidden or unknown influences across the entire Nachlass. It can also simplify the comparison of similar remarks over time. The second functionality (remark comparison) allows specifying two remarks and computing their similarity value. This feature especially addresses domain experts who are familiar with the topic and deliberately want to compare two texts.

At present, WiTTSim solely allows browsing for similar remarks across the Nachlass. Also, the similarity detection does not yet allow cross-language application and thus even direct translations from German to English will not be revealed. Tests with neural machine translation systems showed that the translated results are unusable for presenting them to philosophical users, though they could be used to enrich the feature vectors and would significantly boost similarity detection across languages. Translated ideas will be included in the similarity-browsing functionality.

6 Digital Humanities: Computational Linguists, Editorial Philologists and Philosophers Joining Forces

The development of the Wittgenstein Advanced Search Tools can be regarded as an exemplary cooperation project of open access Digital Humanities where complementary resources, skills, expertise and partnerships are joined together for mutual development and benefit of the community (see Figure 5).

The work on a digital, searchable representation of analog documents leaves us with many difficulties at the junction of media and research, starting with analog documents in the archives, producing facsimile, using OCR-techniques, developing a digital-edition, finding the appropriate TEI-P5 annotation etc. We are continuing the challenging work by condensing the annotation to a degree suitable for search-engines, developing digital, multilingual lexica and performing syntax and semantic analyses with the highest accuracy. For web-presentation, we offer an easy-to-use, browser-compatible web-frontend for searching and exploring the digital data.

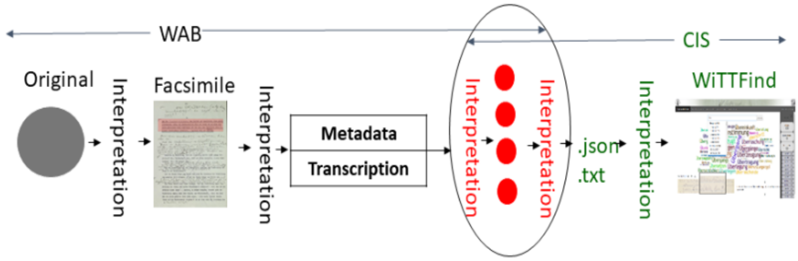

Figure 5: From the originals to the FinderApp.

Intensive and fruitful discussions with Alois Pichler, his team, the user community around WAB, Josef Rothhaupt (Faculty of Philosophy, Ludwig-Maximilians University Munich) and researchers from the CIS help us maintain the degree of quality required by the technicians and the end users. Presentations of WiTTFind at research events and discussions within Digital Humanity projects play a crucial role in providing the necessary impetus to improve the quality of our tools.

Given the unavoidable idiosyncrasies of data, which reflect the genetic process of Wittgenstein's writings (Bruder and Teufel, 2018), our first conclusion was that any search-engine for such types of editions will – unavoidably and unstoppably – always remain "work-in-progress", especially if the users and scholars demand the highest precision and the best recall for the search hits. For any such edition beyond linear text, it will remain impossible to accurately represent such data and guaranteeing the fidelity of the representation.

Thus, apart from questions on the sustainability of such approaches and the long-term storage of digitized cultural heritage, the user must always be given the chance to study not only the HTML-texts of transcriptions but also the corresponding facsimile of the original document. In WiTTReader's "Investigate-Mode" (see Figure 6) – accessible by simply clicking the facsimile – the user is given the possibility to investigate, study, compare, check at all times all underlying layers of data in full detail and can give feedback to WiTTFind developers with one click. At the same time, "Investigate-Mode" also offers benefits to users wanting to learn about transcription techniques and XML TEI-P5 markup.

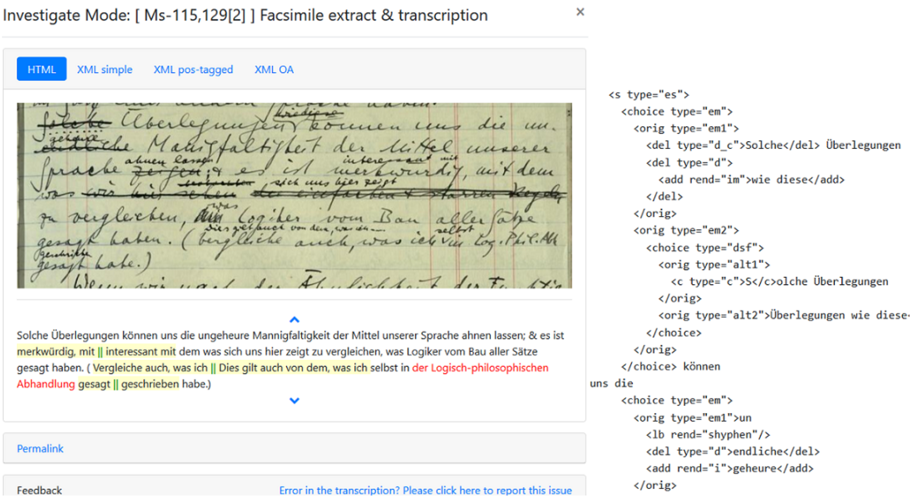

Figure 6: WiTTReader: "Investigate Mode and connection to Feedback-App".

A further conclusion that we may draw from our cooperative research efforts is an imperative for the search engine to guarantee absolute precision in the search results. In essence, the statistical – i.e. approximate – search methods like those commonly used by search engines like Google are not enough and this makes precise rule-based search technologies an absolute necessity for any meaningful philosophical work with the material: The expert users of WiTTFind know what they are searching for, they want to (WiTT)find what they are looking for.

7 Conclusion

For almost ten years now the research group "Wittgenstein Advanced Search Tools" (WAST) under the leadership of Max Hadersbeck from CIS and Alois Pichler from WAB has worked on the development of the web-frontend Finder-App WiTTFind for the possibility of open, advanced search of the Wittgenstein

Nachlass. Today we can firmly state that, thanks to the intense cooperation between philosophy, philology and computational linguistics, we have been able to implement a very successful infrastructure around our FinderApp WiTTFind that covers all aspects of interdisciplinary cooperation regarding facsimile, transcriptions, computational linguistic analyses, web-programming, database, search-machine technology and git-management. We presented our tools and projects at several summer schools, meetings and congresses. We view our cooperation as a very good example of how the research field "Digital Humanities" can offer a new open-access of Wittgenstein's Nachlass. The importance of our project was also highlighted when we won the EU-Open Humanities Award in 2014, Wittgenstein's Nachlass was incorporated into the list of world cultural heritage by the UNESCO's "Memory of the World" program in 2017 and when we were elected to take part in a Wittgenstein exhibition in Vienna (Hadersbeck, 2018).

We hope that our project continues to contribute to the scientific community directly involved in research on Wittgenstein's philosophy and open the path for other communities to investigate Ludwig Wittgenstein's works.

References

Ball GH, Hall DJ (1965) ISODATA, a novel method of data analysis and pattern classification. Tech. Rep., Stanford research inst Menlo Park CA.

Bruder D (2012) Symmetric Index Structures – 'Highly Efficient Symmetric Text Indexing'. Master thesis, Ludwig-Maximilians University Munich, Centre for Information and Language Processing (CIS).

Bruder D, Teufel S (2018) Data Models for Digital Editions: "Complex XML Versus Graph Structures". In: 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., Cologne University, Cologne (Germany), pp. 158–162. ISBN: 978-3-946275-02-2, URL: `http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf`.

Ester M, Kriegel HP, Sander J, Xu X, et al (1996) A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Association for the Advancement of Artificial Intelligence (AAAI) Press, Vol. 96, pp. 226–231. URL: `https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf`.

Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge (USA).

Gross M (1997) The Construction of Local Grammars. Finite-State Language Processing, pp. 329–354, MIT Press, Cambridge (USA). URL: https://halshs.archives-ouvertes.fr/halshs-00278316/document.

Hadersbeck M (2018) Investigating Wittgenstein's Nachlass: WiTTFind, WiTTReader, OdysseeReader and Wittgenstein Advanced Search Tools. In: Ludwig Wittgenstein: Die Tractatus Odyssee-Ausstellung, Schweitzer R, Laufersweiler B (eds), Wittgenstein Initiative, Vienna, pp. 226–237.

Hadersbeck M, Pichler A, Fink F, Seebauer P, Strutynska O (2012) New (Re)Search Possibilities for Wittgenstein's Nachlass. In: Contributions to the 35th International Wittgenstein Symposium 2012, Greis H, Weiss MG (eds), Austrian Ludwig Wittgenstein Society (ALWS), Kirchberg am Wechsel (Austria), Publications of the Austrian Ludwig Wittgenstein Society – New Series 20, pp. 102–105. URL: https://www.degruyter.com/view/product/211268.

Hadersbeck M, Pichler A, Fink F, Gjesdal Ø (2014) Wittgenstein's Nachlass: WiTTFind and Wittgenstein Advanced Search Tools (WAST). In: Proceedings of the 1st International Conference on Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014), Association for Computing Machinery (ACM), New York (USA), pp. 91–96.

Hamp B, Feldweg H (1997) GermaNet – A Lexical-Semantic Net for German. Association for Computational Linguistics (ACL).

Henrich V, Hinrichs E (2010) GernEdiT: A Graphical Tool for GermaNet Development. In: Proceedings of the ACL 2010 System Demonstrations, Association for Computational Linguistics (ACL), pp. 19–24. URL: https://www.aclweb.org/anthology/P10-4004.

Krey A (2014) Semantische Annotation von Adjektiven im Big Typescript von Ludwig Wittgenstein. Bachelor thesis, Ludwig-Maximilians University Munich, Centre for Information and Language Processing (CIS).

Lindinger M (2015) Entwicklung eines WEB-basierten Faksimileviewers mit Highlighting von Suchmaschinen-Treffern und Anzeige der zugehörigen Texte in unterschiedlichen Editionsformaten. Master thesis, Ludwig-Maximilians University Munich, Centre for Information and Language Processing (CIS).

Lloyd S (1982) Least Squares Quantization in PCM. IEEE Transactions on Information Theory 28(2):129–137, Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/TIT.1982.1056489.

MacQueen J (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley (USA), Vol. 1, pp. 281–297.

Miller GA (1995) WordNet: A Lexical Database for English. Communications of the ACM 38(11):39–41, Association for Computing Machinery (ACM), New York (USA). DOI: 10.1145/219717.219748.

Paumier S (2003) Unitex 3.1 User Manual. URL: https://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf. Originally published by Université Paris-Est Marne-la-Vallée, revision and English translation by the Ludwig-Maximilians University Munich, Centre for Information and Language Processing (CIS).

Pichler A (1994) Untersuchungen zu Wittgenstein's Nachlass. In: Working Papers from the Wittgenstein Archives at the University of Bergen, Huitfeldt C, Johannessen KS, Nordenstam T (eds), Wittgenstein Archives of the University Bergen (WAB), Bergen (Norway). ISBN: 82-9107-107-1, URL: http://wab.uib.no/wp-no8.pdf.

Redner RA, Walker HF (1984) Mixture Densities, Maximum Likelihood and the EM Algorithm. SIAM review 26(2):195–239, Society for Industrial and Applied Mathematics (SIAM).

Röhrer I (2017) Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass. Master thesis, Ludwig-Maximilians University Munich, Centre for Information and Language Processing (CIS).

Schmid H (1995) Improvements in Part-of-Speech Tagging with an Application to German. In: Natural Language Processing Using Very Large Corpora, Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D (eds), Springer, Dordrecht (The Netherlands), Text, Speech and Language Technology (TLTB), Vol. 11, pp. 13–25. DOI: 10.1007/978-94-017-2390-9_2.

Steinhaus H (1955) Quelques Applications des Principes Topologiques à la Géométrie des Corps Convexes. Fundamenta Mathematicae 41(1955):284–290. DOI: 10.4064/fm-41-2-284-290.

Ullrich S, Bruder D, Hadersbeck M (2018) Aufdecken von „versteckten“ Einflüssen: teil-automatisierte textgenetische Prozesse mit Methoden der Computerlinguistik und des Machine Learning. In: 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., Cologne University, Cologne (Germany), pp. 355–357. ISBN: 978-3-946275-02-2, URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

Ward Jr JH (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association (ASA) 58(301):236–244. DOI: 10.1080/01621459.1963.10500845.

Wittgenstein L (2016) Interactive Dynamic Presentation (IDP) of L. Wittgenstein's Philosophical Nachlass. URL: http://wab.uib.no/transform/wab.php?modus=opsjoner. Edited by Wittgenstein Archives of the University of Bergen under the direction of Alois Pichler. Bergen (Norway): Wittgenstein Archives of the University Bergen (WAB).

Wittgenstein L, Ogden C (1990) Tractatus Logico-philosophicus. International Library of Psychology, Philosophy, and Scientific Method, Routledge. ISBN: 978-0-415051-86-6.

von Wright G (1969) The Wittgenstein Papers. In: The Philosophical Review, Vol. 78, pp. 483–503.