Wittgenstein's Nachlass: WiTTFind and Wittgenstein Advanced Search Tools (WAST)

Max Hadersbeck
Centrum für Informations- und Sprachverarbeitung (CIS), University of Munich
Maximilian.Hadersbeck@lmu.de

Alois Pichler
Wittgenstein Archives at the University of Bergen (WAB)
Alois.Pichler@fof.uib.no

Florian Fink
Centrum für Informations- und Sprachverarbeitung (CIS), University of Munich
finkf@cis.uni-muenchen.de

Øyvind Liland Gjesdal
Wittgenstein Archives at the University of Bergen (WAB)
Oyvind.Gjesdal@ub.uib.no

ABSTRACT

In the current paper, we present the web-based finder frontend WiTTFind, together with the Wittgenstein Advanced Search Tools (WAST), which offer new possibilities for exploring and researching Ludwig Wittgenstein's philosophy and work. Unlike the search capabilities of Google books and the Open Library project, our tools are rule-based together with local grammar search technology and TEI-P5 XML version of Wittgenstein's Nachlass, in combination with an electronic lexicon and various computational tools to enable lemmatized, semantic and syntactic queries to the texts. The query results are displayed in a web browser as HTML transformations of the transcribed texts, together with a facsimile of the matching segment in the original.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital libraries**]: Collection—*Wittgenstein's Nachlass*; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Application, Editorial Philology

Keywords

digital humanities, applied computational linguistics, local grammars, electronic lexicon with semantic annotation, OCR, Ludwig Wittgenstein, manuscripts, full text search, lemmatized search, XML, TEI, corpus linguistics, editorial philology, digital library, philosophy

DATeCH '14 Madrid, Spain
 ACM 978-1-4503-2588-2/14/05. http://dx.doi.org/10.1145/2595188.2595202 ...$15.00.

1. INTRODUCTION

The Austrian philosopher Ludwig Wittgenstein (1889-1951) left at his death behind 20,000 pages of philosophical manuscripts and typescripts. This is called Wittgenstein's Nachlass. In 2000 the Wittgenstein Archives at the University Bergen (WAB) published the Nachlass in a CD-ROM edition, the so-called Bergen Electronic Edition [21].

In 2009, WAB made additionally 5000 pages from the Nachlass freely available on the web at Wittgenstein Source[1] [12]. Though both the BEE and Wittgenstein Source are equipped with search tools, none of them includes lemmatized, grammatical or semantic search. Since 2010, WAB (Dr Alois Pichler) and the Centre for Information and Language Processing, Ludwig-Maximilians University of Munich, Germany, CIS (Dr Maximilian Hadersbeck) cooperate in the research group "Wittgenstein in Co-Text"[2]. In this research group they develop the web-frontend WiTTFind [6] and the Wittgenstein Advanced Search Tools (WAST), which provide lemmatized and inverse lemmatized search and allow queries to the Nachlass which include word forms, semantic and sentence structured specifications. Syntactic disambiguation is done with Part-of-Speech tagging and local grammar techniques [2]. WiTTFind, together with WAST provide the possibility of searching for words and phrases in the context of sentences, the only meaningful units, as Wittgenstein writes in the Tractatus logico-philosophicus [22, 3.3]:

> Nur der Satz hat Sinn; nur im Zusammenhang des Satzes hat ein Name Bedeutung[3]

This rule-based approach is central to provide researchers with fine-grained perspectives on the Nachlass.

[1] `http://www.wittgensteinsource.org/`
[2] `http://www.cis.uni-muenchen.de/forschung/ehumanities/research-group-co/index.html`
[3] "Only propositions have sense; only in the nexus of a proposition does a name have meaning"

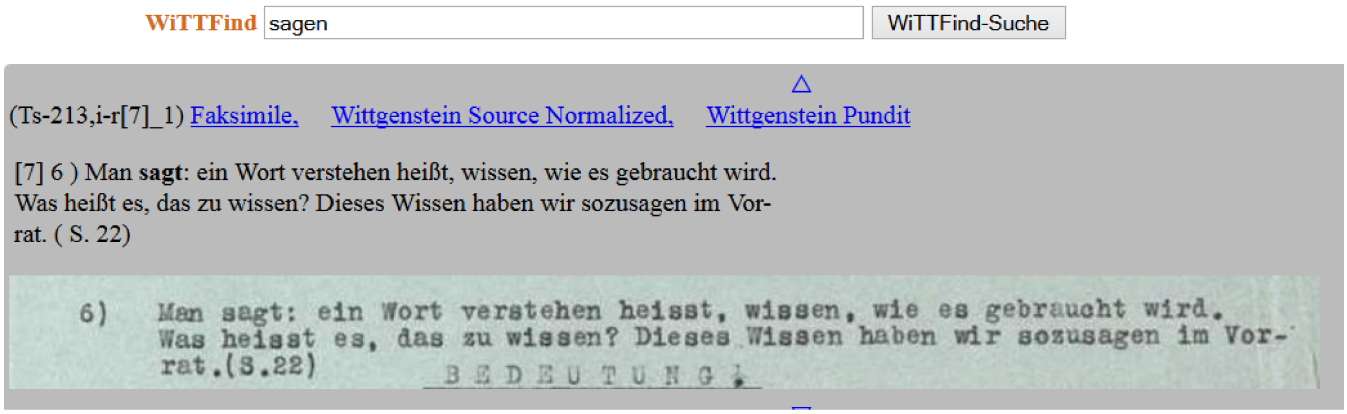

Figure 1: web-frontend WiTTFind, see http://wittfind.cis.uni-muenchen.de

2. THE WEB-BASED FRONTEND WITTFIND AND RESOURCES

2.1 WiTTFind

The web-based frontend WiTTFind consists of several web programs which offer an easy to use interface to query Wittgenstein's Nachlass and presents all search results as HTML-transformation of the edited text together with an excerpt from the original facsimile of Wittgenstein's remark. For the texts available on Wittgenstein Source[4], every result is linked to Wittgenstein Source as well as to the Semantic web annotation tool Pundit [4][5] (see figure 1)

2.2 Resource Text: TEI P5 conformant XML

The XML-Transcription of Wittgenstein's Nachlass [11], which was developed at WAB in Bergen, annotates the texts in greatest detail. All deletions and substitutions etc. of Wittgenstein are annotated in the XML-texts. The latter is too detailed in order to be used by WAST and thus a new TEI P5 [19] compatible and reduced XML-format CISWAB was defined. This text format is an optimal base for the cooperation between the Wittgenstein researchers and programmers in the field of computational linguistic. The CISWAB texts are extracted via XSLT-transformation from WAB's XML-transcriptions of Wittgenstein's Nachlass [12].

2.3 Resource Lexicon: Electronic full-form Lexicon

Successful text searches crucially rely on the availability of an electronic full-form lexicon. For the work on Wittgenstein's Nachlass, we used CISLEX [5], one of the biggest electronic German full-form lexica, which has been developed at CIS over the last 20 years. We extracted all words from the texts available on Wittgenstein Source, the word information from CISLEX and extended it to the special lexicon, called WiTTLex. Each wordentry in WiTTLex is formatted according to the DELA format, defined at the Laboratoire d'Automatique Documentaire et Linguistique (LADL), Paris [2]. The lexicon entries contain the word's full-form, lemma, lexicographical word form, inflection and semantic information. Search queries to WiTTFind can be grammatically processed with the help of WiTTLex. Figure 2 shows a short excerpt from the lexicon.

```
Advokat, Advokat.N+HUM:neM
gesagt, sagen.V:0Z
gesamte, gesamt.ADJ+NUM:aeFxp:aeFyp:aeFzp:aeNyp:\
 amUxp:neFxp:neFyp:neFzp:neMyp:neNyp:nmUxp
dagestanden, dastehen.V+#2
glänzende, glänzend.ADJ+ER+COL+Glanz
Fermat, .EN
Fermatsche, Fermat'sch.ADJ+EN
rot, rot.ADJ+COL+Grundfarbe
rötlicher, rötlich.ADJ+COL+Zwischenfarbe+KOMP:\
 deFxp:geFxp:gmUxp:neMxp:neMzp
```

Figure 2: Sample dictionary entries in DELA-Format

2.4 Applicability to other document collections

As far as the applicability to other document collections is concerned, our tools are ready to host other authors and their respective cultural heritage. In order to put our search technologies to use, TEI-P5 XML edited texts of the authors and a full-form lexicon of the corpus are necessary. Tokenization, Part-of-Speech-Tagging and End-of-sentence detection are provided in our institute's lab.

3. WITTGENSTEIN ADVANCED SEARCH TOOLS (WAST)

3.1 The finder Application suite wf

In order to provide extended linguistic search capabilities for explorations from non-technical and non-linguistic backgrounds, a suite of tools was developed at CIS [17, 7, 1, 20]. With the help of local grammar techniques we are able

[4] http:/www.wittgensteinsource.org
[5] http://feed.thepund.it

to offer advanced linguistic search capabilities to researchers from different fields of the digital humanities. The next section gives a bottom up overview on the program suite wf, one section out of WAST, that is used to find text, i.e. to find utterances of Wittgenstein.

Every query to wf is internally transformed to a local-grammar-search-graph of the search terms [10]. In its simplest form these graphs are a chain of the search terms. But in general search-graphs can contain an arbitrary number of loops and branches to express complex search patterns [1]. The system uses various strategies to match tokens in the text. It is able to use e.g. string-based matching, regular expressions, semantic and morphological information from dictionaries and arbitrary annotations from the searched text itself.

In order to present a simple interface to the user and hide much of the complexities involved, queries from the user are automatically converted to local-grammar-search-graphs. A simple query language enables the user to search for sequences of arbitrary length or combine tokens with boolean operators [1]. Furthermore, the system applies implicit conversions on the queries to generate more natural search results.

3.2 Implementation aspects of wf

The wf tool is implemented in a client-server architecture. There are three main tools involved in the processing of queries: A server-, a client- and a displaytool. The client merely takes queries from the user and applies some very basic transformation to them before sending the actual query to the server. These transformations mostly convert convenient wildcard expressions to valid more complex regular expressions.

On the server side each query is parsed and then transformed to a local-grammar, that is applied asynchronously to the different documents the query wishes to search. One running server instance is able to handle parallel searches over a various set of different documents. All results are sent back to the client.

The third tool is a simple displaying tool, that is able to output the result of the queries in different formats[6]. These three tools combined, are the backend to present query results, that are further used by the WiTTFind front end to present the final results of queries.

Table 1: Benchmarks of the wf server

Query type	Median (seconds)
strict	0.082
word phrase	0.080
lemma-based	0.088
inverted lemma-based	0.088
sentence	0.169
local grammar (particle verbs)	0.214

Table 1 lists a view benchmarking results of different query types to a wf server instance running on a modern desktop PC[7]. The numbers each show the median of 10 different queries to the part of Wittgenstein's Nachlass, which is called the "Big Typescript" (Ts-213). All queries where limited to a maximum of 100 hits[8]. It should be noted, that the last two lines show the results of complex queries, that generate very few hits. These queries take a lot more time to finish (more than twice as long), since the whole document has to be processed and local grammars are used to disambiguate.

[6]At the time being, there are only two formats supported; one format for the web-service, and another format for the terminal.

[7]Intel 2.90GHz Quadcore

3.3 Lemmatized and inverted lemmatized word-search

The electronic full-form lexicon WiTTLex allows to perform lemmatized queries to the Nachlass. For example, the search for the word *denken* (think) returns all sentences which contain morphological variants of the word, like *dachte*, *gedacht*. We also implemented with the help of WiTTLex an inverted lemmatized search where we used the lemma of the queried word to produce all morphological variants of it. The word *dachte* leads to the lemma *denken* again, from which all word variants are derived.

3.4 Word-Form Search and Part-of-Speech Tagging

In the full-form lexicon, WiTTLex, every word is stored together with its lexical word form as it is defined in the German CISLEX (see table 2).

Table 2: Wordform Tags

Name	Tag	Translation
Nomen	<N>	noun
Adjektiv	<ADJ>	adjective
Verb	<V>	verb
Determinativ	<DET>	determiner
Adverb	<ADV>	adverb
Partikel	<PART>	particle
Präposition	<PREP>	preposition
Präposition + Artikel	<PDET>	preposition + article
Konjunktion	<KONJ>	conjunction
Interjektion	<INTJ>	interjection
Verbpartikel	<VPART>	verb particle
Eigenname	<EN>	proper name

The finder allows the use of word forms as placeholders for words in a query. For example, the query *Die <ADJ> Farbe* finds all sentences that contain the nominative feminine definite article *Die*, followed by an adjective, which precedes the noun *Farbe* (colour) as in sentence (1).

(1) Q: Die <ADJ> Farbe

A: Ich könnte Dir **die genaue Farbe** der Tapete zeigen, wenn hier etwas wäre was diese Farbe hat.[9]

To reduce the syntactic ambiguity of the word forms in the text, we additionally tagged the data with the Part-of-Speech (POS) treetagger [16]. The finder permits POS-tags to be used as placeholders for the selection of syntactic word forms, as defined in the Stuttgart-Tübingen-Tagset [15]. The user can decide to search for a lexical word form (e.g. <ADJ>) or the syntactic word-form within the sentence (e.g. [ADJ*]).

[8]On the web the limit is set to 25 for unregistered users.

[9]I could show you the exact colour of the wallpaper if there was anything around that has this colour.

Table 3: Semantic labels for nouns

Name	Tag	Translation	Occurrences
Menschen	<HUM>	humans	140
Tiere	<T>	animals	96
Pflanzen	<PF>	plants	26
Objekte	<OBJ>	objects	1402
Ereignisse	<ER>	events	589
Zustände	<ZU>	states	51
Eigenschaften	<EIG> p	properties	236
Temporalia	<TEMP>	time	49
Eigennamen	<EN>	proper names	60
Numeralia	<NUM>	number	47
Diversa	<SONST>	other	713

Table 4: Semantic labels for adjectives

Name	Tag	Translation	Occurrences
Farben	<COL>	colour	974
Numeralia	<NUM>	number	1258
Relation	<REL>	relation	2517
Eigennamen	<EN>	proper names	17
Temporalia	<TEMP>	time	619
Evaluation	<EVAL>	evaluation	1732
Zustände	<ZU>	states	6629
Komparativa	<KOMP>	compariative	2080
Stilistika	<STIL>	style	1917
Eigenschaft	<EIG>	property	382
Ereignisse	<ER>	proppperty	187

3.5 Semantic Lexical Classification

In WiTTLex we classified nouns [18] and adjectives [7] semantically. According to the work by Langer and Schnorbusch [8] we defined eleven classes for nouns (see table 3) and eleven classes for adjectives (see table 4).

In our investigations in Wittgenstein's public available part of the Nachlass[10], we found that there are about 1800 nouns out of all 46000 words in the data (see tables 3 and 4). All the tags from the tables can be used in WiTTFind. For example, the query *<EN> und <EN>* returns the sentence (2) in which the two proper names *Fregeschen* and *Russellschen* are joined by the coordinating conjunction *and*.

(2) Q: <EN> und <EN>

A: Unzulänglichkeit der **Fregeschen und Russellschen** Allgemeinheitsbezeichnung.[11]

In cooperation with the Faculty of Philosophy, Philosophy of Science and the Study of Religion, at the University of Munich (Rothhaupt [13]), we implemented a first semantic classification of Wittgenstein's colour vocabulary together with a special HTML-interface for querying colours in the Nachlass. We found, that five different categories for colours are optimal for Wittgenstein's Nachlass: Grundfarbe (primary colour), Zwischenfarbe (intermediate colour), Transparenz (transparency), Glanz (gloss) and Farbigkeit (colourness) [7]. Table 5 shows the different labels for colours and the number of their occurrences in the text.

In the web-frontend WiTTFind, the user can select between different colour categories (see table 5), view statistics and query Wittgenstein's Nachlass.

3.6 Sentence structure and Wildcards

For the Wittgenstein researchers it is very important to take the sentence structure into account. To enable users to specify the sentence structure within queries, we introduced sentence structuring tags in queries, such as <BOS> (the beginning of a senctence), <EOS> (the end of a sentence) and <PUNCT> for punctuation characters. The wildcard operator * can be used as placeholder for arbitrary character sequences. The query: *<BOS> Ich meine nur <PUNCT> * * * <PUNCT> <EOS>* would return all sentences that consist of six words starting with the sequence of three tokens `Ich meine nur` followed by a punctuation character as in example (3)

(3) Q: <BOS> Ich meine nur <PUNCT> * * * <PUNCT> <EOS>

A: **Ich meine nur, was ich sage.**[12]

Table 5: Semantic labels for colours

Name	Tag	Translation	Occurrences
Grundfarbe	<Grundfarbe>	basic colour	454
Zwischenfarbe	<Zwischenfarbe>	intermediate colour	301
Transparenz	<Transparenz>	transparency	105
Glanz	<Glanz>	gloss	2
Farbigkeit	<Farbigkeit>	colourness	29

3.7 Rule-based linguistic search with Part-of-Speech Tagging (POS-tagging)

To enrich the number of found verbs concerning a search query, we implemented an automatic particle-verbs detection and distinction. Particle-verbs are marked in their lexical entry and divided into verb and particle. In the Big Typescript we found almost 750 verbs with separated particles. To disambiguate the separated particles from prepositions, we use Part-of-Speech tagging and local grammars. For example, the query with the particle verb *dastehen* would extract instances as: *steht klar da ...* in which the verb particle *da*, which may occur in German separately from the verb *stehen*, is recognized as such and not as a preposition (see figure 3).

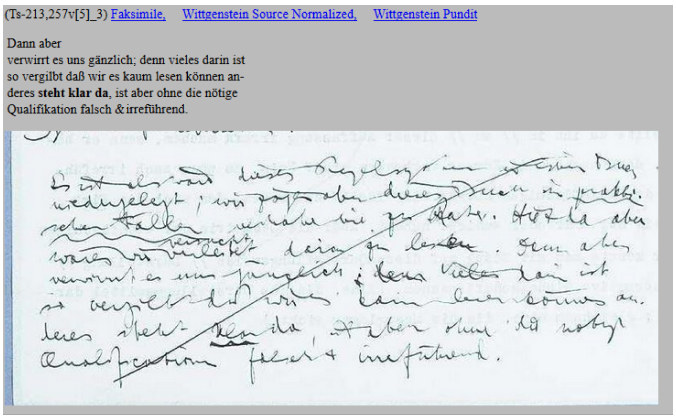

Figure 3: Distinction of particle verbs : "dastehen"

[10] `http://www.wittgensteinsource.org/`

[11] Shortcomings of the denomination of universality made by Frege and Russell.

[12] I only mean, what I say.

3.8 Search without Alternatives

One characteristic feature of Wittgenstein's Nachlass is, that Wittgenstein changed his texts very often and as a result the Nachlass offers a lot of different readings. To enable the finding of remarks in all of the latter, in a background process of the finder application, we generate all different readings, which are processed simultaneously [17].

4. SEARCH RESULT LAYOUT WITH THE FACSIMILE

Ludwig Wittgenstein's Nachlass is highly heterogeneous. It consists of a large number of texts, some of them handwritten, with numerous passages edited from the author himself. The researchers can only appreciate the found remarks to the fullest with all its editions, errors and overall form, if they see not only the HTML-transformation of the edition, but as well their original facsimile which should be displayed simultaneously [14]. Only this combined view provides the possibility for comparison and analysis of the original material, which carries the complete set of information. In order to implement this layout, it was necessary to OCR all facsimile of Wittgenstein's Nachlass, extract the coordinates of his remarks and link them to the search result according to their siglum [3]. All the work was done with ABBYY FineReader and additionally developed PERL programs. The highlighting in the display of the facsimile is done with CSS techniques within the HTML Page [9].

5. CONCLUSION

In this paper we demonstrated that the web-based frontend WiTTFind together with the Wittgenstein Advanced Search Tools (WAST) offer a new and very efficient application to the Nachlass of Wittgenstein for researchers on an interdisciplinary level. The computational linguists had to realize that searching for utterances is not a statistical process. It must be well defined, configurable and easy to use, because every word is important. Therefore, we used rule-based systems, developed HTML front ends and help-pages specifically designed for these users and this task. Our computational enrichment like lemmatized, semantic, syntactic offer the Wittgenstein community a new web-based access to Wittgenstein's Nachlass to specify highly sophisticated queries. Especially users, who are not native German speakers, are for example very fond of our linguistic extensions, like lemmatized search. With our web-based finder application WiTTFind we cover all texts from the freely available part of the Ludwig Wittgenstein's Nachlass and offer an efficient search in a short time. With the display of the edited text together with facsimile, researchers are able to overcome edition errors and can explore in their specific "aura" the original handwritten Nachlass-texts which are otherwise stored in access-restricted archives (see figure 3).

Our finder application WiTTFind, together with the WAST-tools, can easily used in conjunction with different document selections, as long as the researchers can offer TEI-P5 XML annotated texts and a full form lexicon of the used words.

6. REFERENCES

[1] F. Fink. Programming of the rule-based wf. http://www.cis.uni-muenchen.de/kurse/max/scholarship/finkwf.pdf (1. Wittgenstein Scholarship 2013), 2013.

[2] G. Gaston. La form d'un dictionnaire electronique. *LADL-Report. Laboratoire d'Automatique Documentaire et Linguistique*, 1991.

[3] A. Gotscharek, U. Reffle, C. Ringlstetter, K. U. Schulz, and A. Neumann. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International Journal on Document Analysis and Recognition*, 14(2):159–171, 2011.

[4] M. Grassi, C. Morbidoni, M. Nucci, S. Fonda, and F. Piazza. Pundit: Augmenting WEB Contents With Semantics. *Literary and Linguistic Computing. Special Issue 'Digital Humanities 2012: Digital Diversity: Cultures, Languages and Methods'. Edited by Paul Spence, Susan Brown and Jan Christoph Meister*, 28(4), December 2013. siehe http://llc.oxfordjournals.org/content/current.

[5] F. Guenthner and P. Maier. Das CISLEX Wörterbuchsystem. *CIS-Bericht-94-76*, 1994.

[6] M. Hadersbeck, A. Pichler, F. Fink, P. Seebauer, and O. Strutynska. New (re)search possibilities for Wittgenstein's Nachlass. *35th International Wittgenstein Symposium 2012, Kirchberg am Wechsel*, 2012.

[7] A. Krey. Semantische Annotation von Adjektiven im Big Typescript von Ludwig Wittgenstein. Bachelorarbeit am Centrum für Informations- und Sprachverarbeitung , Ludwig-Maximilians-Universität, München, 2013.

[8] S. Langer and D. S. (Hrsg.). *Semantik im Lexikon.* Tübingen: Narr, 2005.

[9] M. Lindinger. Highlighting von Treffern des Tools WiTTFind im zugehörigen Faksimile. Bachelorarbeit am Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2013.

[10] G. Maurice. The Construction of Local Grammars. *Finite-State LanguageProcessing*, page 329–354, 1997.

[11] A. Pichler. „Towards the New Bergen Electronic Edition". *Wittgenstein After His Nachlass. Edited by Nuno Venturinha*, pages 157–172, 2010.

[12] A. Pichler, H. Krüger, D. Smith, T. Bruvik, A. Lindebjerg, and V. Olstad. Wittgenstein Source Bergen Facsimile Edition (BTE). *Wittgenstein Source. Bergen: WAB, Wittgenstein Source*, 2009. `http://www.wittgensteinsource.org`.

[13] J. G. F. Rothhaupt. *Farbthemen in Wittgensteins Gesamtnachlaß. Philologisch-philosophische, Untersuchungen im Längsschnitt und in Querschnitten.* PhD thesis, Weinheim, 1996. Beltz Athenaum.

[14] J. G. F. Rothhaupt. Zur dringend notwendigen Revision des „standard view" der Genese der "Philosophischen Untersuchungen". *Gasser, Georg / Kanzian, Christian / Runggaldier, Edmund (Hg.): Cultures: Conflict - Analysis - Dialogue. Papers of the 29th International Wittgenstein Symposium 2006,*

Kirchberg 2006, pages S. 278–280, 2006.

[15] A. Schiller, C. Thielen, and S. Teufel. Stuttgart Tübinger Tagset (STTS). Technical report, http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html, 1999.

[16] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*, 1994.

[17] P. Seebauer. Verbesserung der Suche im Wittgenstein-Nachlass. Suche in alternierenden Texten. Magisterarbeit am Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2012.

[18] O. Strutynska. Nachlass von Ludwig Wittgenstein: Optimierung eines digitalen Lexikons und semantische Kodierung der Nomen. Evaluation des Lexikons mit Hilfe von Konkordanzanalysen. Bachelorarbeit am Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2012.

[19] TEI Consortium. Guidelines for Electronic Text Encoding and Interchange. Technical report, TEI Consortium, July 2009. http://www.tei-c.org/Guidelines/P5/ (September 2nd 2009).

[20] L. Volos. Disambiguierung von Partikelverb – Konstruktionen und Verbpräpositional – Konstruktionen im Big Typescript von Ludwig Wittgenstein. Bachelorarbeit am Centrum für Informations- und Sprachverarbeitung Ludwig-Maximilians-Universität, München, 2013.

[21] L. Wittgenstein. *Wittgenstein's Nachlass: The Bergen Electronic Edition*. Oxford: OUP, 2000.

[22] L. Wittgenstein and C. Ogden. *Tractatus Logico-philosophicus*. International Library of Psychology, Philosophy, and Scientific Method. Routledge, 1990.