New (Re)search Possibilities for Wittgenstein's Nachlass

Max Hadersbeck, Alois Pichler, Florian Fink, Patrick Seebauer, and Olga Strutynska*

Munich, Germany | Maximilian.Hadersbeck@lmu.de

1. Introduction

„Nur der Satz hat Sinn; nur im Zusammenhang des Satzes hat ein Name Bedeutung“ Wittgenstein writes in the *Tractatus logico-philosophicus* (3.3). But how does Wittgenstein use words himself, what context does he choose in his own sentences?

In this paper we present to the Wittgenstein research community a new Web-based computational linguistic access to the Big Typescript Ts-213 (BT). We developed a special electronic full-form lexicon WiTTLex and the tool WiTTFind to search within the BT for special words, sequences of words, parts of sentences and special sentence structures using methods of computational linguistics. Like usual search machines the user communicates via an internet browser with WiTTFind. However the query possibilities of our tool exceed the possibilities of search engines by far. Our user queries are not limited to word based queries, they can be lemmatized and grammatically structured.

The WWW-address of the tool is: http://wittfind.cis.uni-muenchen.de

Figure 1 shows an example of a lemmatized search for the word “sagen”. The tool finds more than 1600 inflected word forms in the BT, which are displayed with the corresponding “Satzsiglum” and the sentence context in the browser.

By clicking on the “Satzsiglum” the user can view our triptych-display showing scans of Wittgenstein’s original double-sided typescript and annotations. (Figure 2)

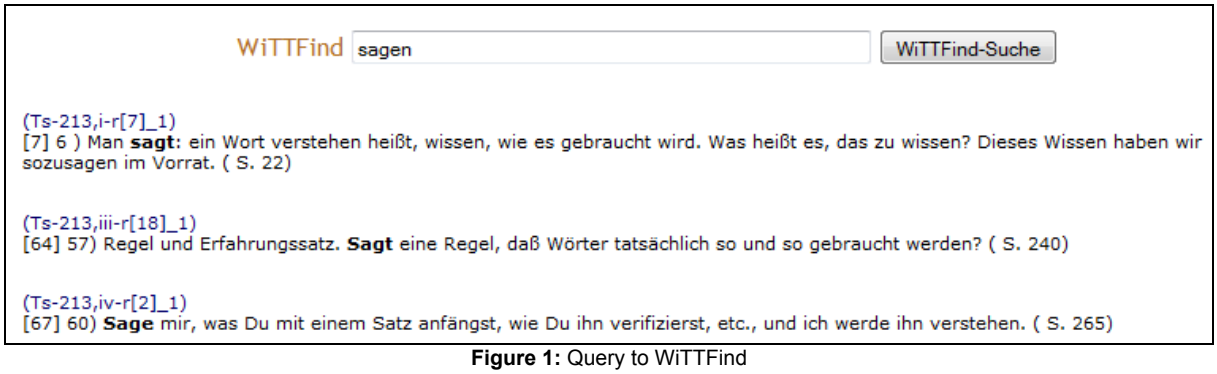

Figure 1: Query to WiTTFind

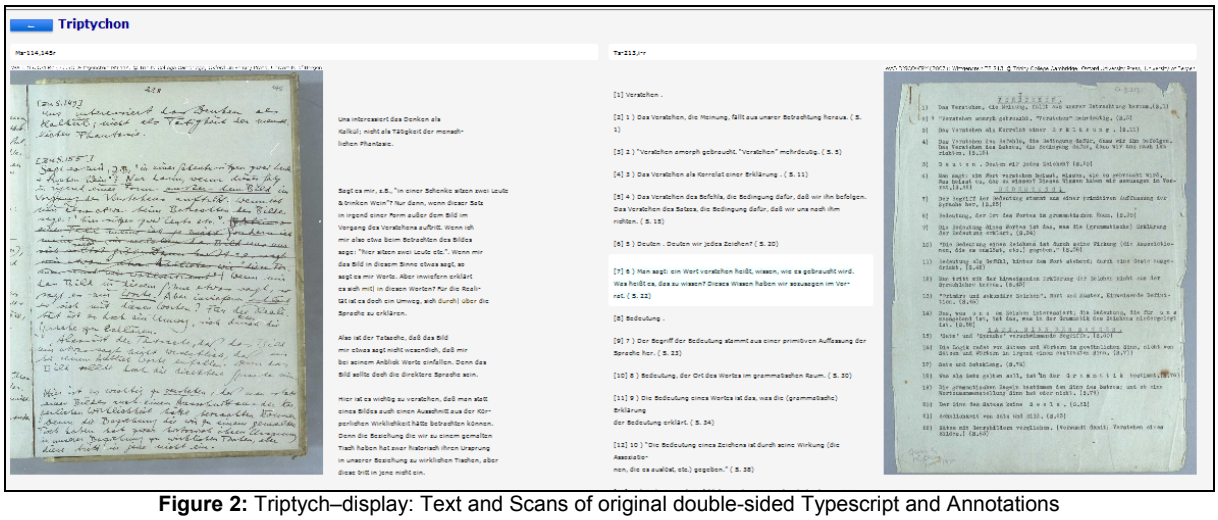

Figure 2: Triptych–display: Text and Scans of original double-sided Typescript and Annotations

* Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig-Maximilians Universität, München and Wittgenstein Archives at the University of Bergen (WAB).

2. Preparation of the Big Typescript for computational linguistic work

The XML-Transcription of the BT from the Wittgenstein Archives at the University of Bergen (WAB; the transcription can be downloaded from http://wab.uib.no/wab_hw.page/) offers an excellent basis for computational linguistic analysis, but had to be simplified for our work. From the WAB's normalized edition of the Big Typescript we extracted a html-file, which we transformed into the XML-notation CISWAB with the help of several programs written in the programming language PERL (Hadersbeck 2011). Our simplified text-version structures the BT into remarks and sentences, and still contains WAB's coding of line endings, page breaks, text alternations, mathematical notations and all the text phenomena of Ludwig Wittgenstein's Big Typescript considered relevant.

3. The electronic full-form lexicon WiTTLex

Successful computational linguistic work relies crucially on the use of an electronic full-form lexicon. For the work with the BT we constructed a special lexicon called WiTTLex. For the development of WiTTLex we were able to use CISLEX (Guenthner 1994), one of the biggest German electronic full-form lexica, which has been developed at the Centrum für Informations- und Sprachverarbeitung (CIS) over the last 18 years. WiTTLex includes all words from the normalized BT edition. Each word-entry in WiTTLex is formatted according to the DELA Format, defined at the Laboratoire d'Automatique Documentaire et Linguistique (LADL, Paris) (Gross, G. 1991). The lexicon entries contain the word's full form, lemma, and lexicographical word form, together with flexion and semantic notations (Langer 2005) for frequent words. With the help of WiTTLex search queries to WiTTFind can be processed lemmatized and grammatically. The following lines show a short extract from the lexicon:

sagst,sagen.V+refl(a):2eGi
sagten,sagen.V+tr:1mVc:1mVi:3mVc:3mVi
sagte,sagen.V:1eVc:1eVi:3eVc:3eVi
Sprachspiele,Sprachspiel.N:amN:deN:gmN:nmN
Zeichen,.EN+Hum+Nachname

4. The computational linguistic tool WiTTFind

In order to find and display the searched parts of the text, we developed the computational linguistic tool WiTTFind. The long-lasting successful work at CIS with the Corpus Word Processing Tool UNITEX, developed at the Laboratoire d'Automatique Documentaire et Linguistique (LADL) (Paumier 2002) formed the basis for our tool: We also work with the technique of local grammars (Gross, M. 1997). For highly-efficient pattern-search with local grammars, we implemented finite state transducers (Guenthner 2005; Reffle 2011); to access the electronic lexicon very quickly, we use the data-structure HAT-Trie (Akskitis 2007). All our programs are written in the programming language C++ as defined in the latest standard C++11 (Hadersbeck 2012). WiTTFind transfers every query to the BT into a local grammar, represents the grammar as a directed graph and translates this graph into an optimized automaton. With the lexicon WiTTLex in the background this automaton analyses every sentence of the BT and tries to match the search query. The text passages which fit the query are displayed with their sentence context and its "Satzsiglum". To enable browser oriented input and output we use modules of our WEB-tool CisWeb, which has been developed over the last few years at CIS. The tool CisWeb which is programmed in JAVA uses the Google-Web-Toolkit library gwt (Hanson 2007) and can be configured in a modularized way for different computational linguistic tasks. To refine the directed graph of the local grammar, the tool offers a graph-editor.

5. Display of hits in the text

Intensive discussions with Dr. Rothhaupt (Rothhaupt 2006) from the Philosophische Fakultät, Ludwig Maximilians Universität München, showed that transcriptions of the Nachlass of Ludwig Wittgenstein and resulting editions are still part of ongoing research. WAB offers many alternative or complementary editions (see for example the interactive edition format, accessible from http://wab.uib.no/wab_hw.page/, and Pichler 2010). WiTTFind only works with the normalized edition format of the BT, so we decided to present beside the sentence concordance of the edition-text also the pictures of the scanned original pages of Wittgenstein's typescript. Wittgenstein used front and back page of his typewriter sheets and had typed on the front pages his script and on the reverse side of this typewritten page he typically placed corrections to the opposite page. We programmed a so called triptych-display, which subdivides the browser window into three parts: On the left page, the facsimile of the reverse side of the last page is displayed, in the middle part you see the extracted page of the normalized edition containing the search hit and on the right side you see the facsimile of the front page of Wittgenstein's original typewritten page (see Figure 2). The triptych-display was developed within the Magisterarbeit of Kaumanns: „Entwicklung einer Ideensuchmaschine im Rahmen des Projekts Wittgenstein in Co-text" (Kaumanns 2012).

6. How to query WiTTFind

With the help of the electronic lexicon WiTTLex in the background, the tool WiTTFind can process classic word and phrase queries as well as very complex lemmatized and grammatical search tasks. We developed a special query language which gives users the possibility to define what they are looking for.

7. Full form search (Exact Search)

Users who are looking for a specific word or phrase can use the exact search. The input has to be enclosed in apostrophes. The punctuation characters inside the phrase can be suppressed. Figure 6 shows an example with more than 290 hits:

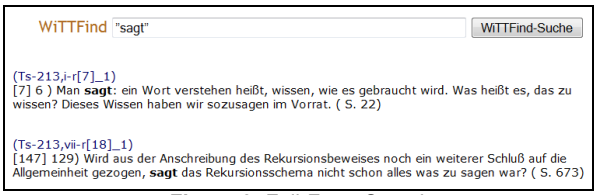

Figure 3: Full-Form-Search

8. Lemmatized search for words (Lemma-Word-Search)

With this kind of search, you can find all text passages where morphological variants of the queried word occur.

With the help of the electronic lexikon WiTTLex, our tool can access the lemma for every specified word, and from this lemma it obtains knowledge of all of its morphological variations. If the user specifies for example the word “dachte”, it finds the lemma “denken” and from here it finds all the other morphological variants like: “denkt“, „denken“, „denkst“, „dachten“, „dachtet” and so on (see Figure 4 with more than 260 hits).

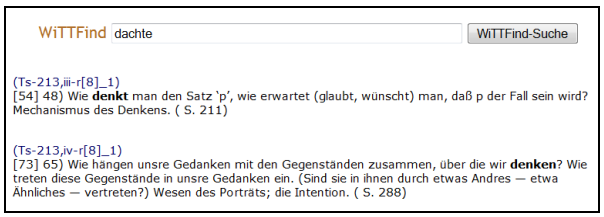

Figure 4: Lemma-Word-Search

9. Lemmatized search for phrases (Lemma-Phrase-Search)

With this kind of search, the user can enter word phrases and WiTTFind processes every word in lemmatized form. It finds all combinations of morphological variants of the single words of the entered word phrase. This search technique is particularly useful for “non-native” speakers, who are not familiar with all the morphological variants of German words (see Figure 5 with more than 39 hits).

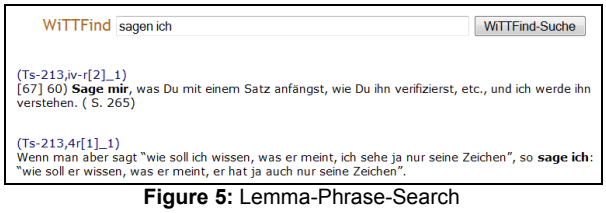

Figure 5: Lemma-Phrase-Search

10. Lemmatized search for phrases including punctuation (Lemma-Phrase-Punct-Search)

Punctuation is very important in Wittgenstein’s work, so we decided to define the special linguistic tag <PUNCT>, which has to be used in the search phrase. If the user is looking for a text passage, including punctuation, then the user must specify <PUNCT> at the appropriate position of his query.

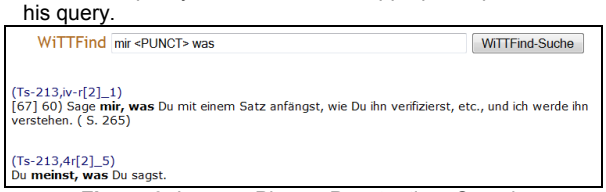

Figure 6: Lemma-Phrase-Punctuation- Search

11. Grammatical Search (Grammar-Search)

Another important piece of linguistic information about a word is its word form. In the lexicon WiTTLex we store for every word its word-form, according to the definitions in the German CISLEX: N for noun, ADJ for adjective, V for verb, DET for determiner, PRON for pronoun, ADV for adverb, PREP for preposition, KONJ for conjunction and EN for proper name. With the help of tags, labeled with the word-forms, WiTTFind can find grammatically defined words (see Figure 7a with more than 28 hits and Figure 7b with more than 60 hits).

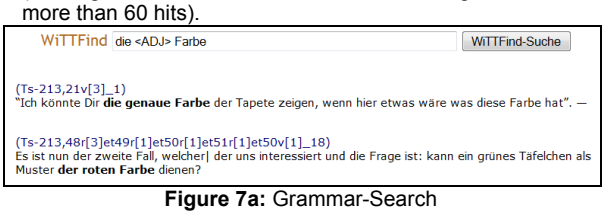

Figure 7a: Grammar-Search

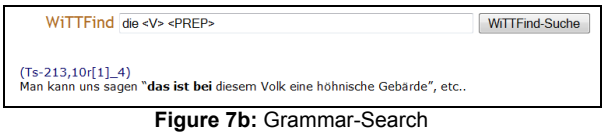

Figure 7b: Grammar-Search

12. Search on the level of sentences (Sentence - Search)

Traditional search engines hide the interpretation of search queries from the user and the user can hardly influence this interpretation. For example you can’t specify the position of the word or phrase you are looking for within a sentence. With our tool WiTTFind you can search sentences fitting a special grammatical sentence structure. The focus of our work was to offer the user a wide range of possibilities for specifying exact search patterns at different levels of information, namely word, phrase and sentence level. So we defined special linguistic tags for the search with WiTTFind:

<BOS> specifies the begin of a sentence
<EOS> specifies the end of a sentence
<WORD> specifies a word
<PUNCT> specifies a punctuation mark
<ALT> specifies an alternate character

The tag <WORD> includes pre- and postponed punctuation marks and can be followed by quantifiers: ?,+,*{n} and {n,m} to express the number of repetitions of the tag. The meaning of these quantifiers are: ‘?’ the word can occur once or none, ‘+’ : one or more occurrences, ‘*’ : allows arbitrary (including no) repetitions, ‘{n}’ : exactly n occurrences, ‘{n,m}’ : minimum n and maximum m occurrences.

Figure 8 shows an example for querying on the sentences level: We search for sentences, beginning with the word “ich” and consisting of 7 words:

Figure 8: Sentence-Search

13. Summary

In this article we presented a new Web-based access to search and research Ludwig Wittgenstein’s Big Typescript Ts-213, based on methods of computational linguistics. We developed the full-form lexicon WiTTLex, which comprises all words of WAB’s normalized edition of Ts-213 with their full-form, lemma, morphologic, syntactic and semantic information. To find and display special words, phrases and sentences in the BT, we programmed WiTTFind, together with a query language, which allows the user to specify exact, lemmatized and grammatical search-queries. WiTTFind can find all inflected forms of words and all combinations of morphological variations of

word phrases. In addition it can detect a wide rage of sentence structures, by the use of grammatical tags. All found text passages in the BT are presented with their sentence context, together with the scanned original pages of Ludwig Wittgenstein's work in a special triptych-display.

References

Askitis, Nikolas and Sinha, Ranjan 2007, *HAT-trie: A Cache-conscious Trie-based Data Structure for Strings*, School of Computer Science and Information Technology, RMIT University, Melbourne 3001, Australia.

Guenthner, Franz 1996 "Electronic Lexica and Corpora Research at CIS", *International Journal of Corpus Linguistics* 1/2, 287-301.

— and Maier, Petra 1994 *Das CISLEX Wörterbuchsystem*, CIS-Bericht-94-76.

Gross, Gaston 1991 *La form d'un dictionnaire electronique*, LADL-Report, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

Gross, Maurice 1997 "The Construction of Local Grammars", in: E. Roche and Y. Schabès (eds.), *Finite-State Language Processing*, Cambridge, Mass.: MIT Press, 329-354.

Hanson, Robert and Tacy, Adam 2007 *GWT in Action Easy Ajax with the Google Web Toolkit,* Greenwich, Conn.: Manning Publications Co.

Hadersbeck, Max 2011 *Einführung in die Programmierung für Computerlinguisten*, Centrum für Informations- und Sprachverarbeitung LMU-München, Skript zur Vorlesung.

— 2012 *Programmierung mit C++ für Computerlinguisten*, Centrum für Informations- und Sprachverarbeitung, LMU-München, Skript zur Vorlesung.

Kaumanns, David 2012 *Entwicklung einer Ideensuchmaschine im Rahmen des Projekts Wittgenstein in Co-text*, Magisterarbeit am Centrum für Informations- und Sprachverarbeitung LMU-München.

Langer Stefan, Schnorbusch Daniel (Hg.) 2005, *Semantik im Lexikon*. Tübingen: Narr.

Maurel, Denis and Guenthner, Franz 2005 *Automata and Dictionaries (Texts in Computer Science)* , London: Kings College Pubn.

Paumier, S. 2002, *Unitex manuel d'utilisation*, Université de Marne-la-Vallée, http://www-igm.univ-lv.fr/~unitex/manuelunitex.pdf (viewed 28th February 2004).

Pichler, Alois 2009 (ed., in collaboration with H.W. Krüger, D.C.P. Smith, T.M. Bruvik, A. Lindebjerg, V. Olstad 2009) *Wittgenstein Source Bergen Facsimile Edition (BTE).* In: Wittgenstein Source. Bergen: WAB, Wittgenstein Source.

— 2010 "Towards the New Bergen Electronic Edition", in: Nuno Venturinha (ed.), *Wittgenstein After His Nachlass*, Houndmills: Palgrave Macmillan, 157-172.

Reffle, Ulrich 2011 *Algorithmen und Methoden zur dokumentenspezifischen Analyse historischer und OCR-erfasster Texte,* Dissertation, Ludwig-Maximilians-Universität München (ISBN 978-3-8439-0106-2).

Rothhaupt, Josef G. F 2006 „Zur dringend notwendigen Revision des „standard view" der Genese der "Philosophischen Untersuchungen". In: Georg Gasser, Christian Kanzian, and Edmund Runggaldier (Hg.): *Cultures: Conflict – Analysis – Dialogue. Papers of the 29th International Wittgenstein Symposium 2006*, Kirchberg am Wechsel: ALWS, 278-280.