Wittgensteins Nachlass: Computerlinguistik und Philosophie

Der Finder wiTTFind und die Wittgenstein Advanced Search Tools (WAST)

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal

Maximilian.Hadersbeck@lmu.de

Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München,

Wittgenstein Archives at the University of Bergen (WAB).

1 EINLEITUNG

In meinem Vortrag möchte ich über unsere Arbeitsgruppe „Wittgenstein in Co-Text“ am Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig Maximilians Universität München (Dr. Max Hadersbeck) und dem Wittgenstein-Archiv an der Universität Bergen (WAB) / Norwegen (Dr. Alois Pichler) berichten. Vor zwei Jahren begannen wir in dieser Arbeitsgruppe für die öffentlich zugänglichen Teile des Nachlasses von Ludwig Wittgenstein (siehe Bergen Electronic Edition (BEE, 2000) und die Open Source Plattform *Wittgenstein Source* (http://www.wittgensteinsource.org, 2009-) computerlinguistische Verfahren zu entwickeln, die einen neuen WEB-basierten Zugang zu den Texten ermöglichen, um Wörter und Phrasen im „Zusammenhang des Satzes“ zu finden. Denn so schrieb schon Wittgenstein im *Tractatus logico-philosophicus* (3.3): „Nur der Satz hat Sinn; nur im Zusammenhang des Satzes hat ein Name Bedeutung“.

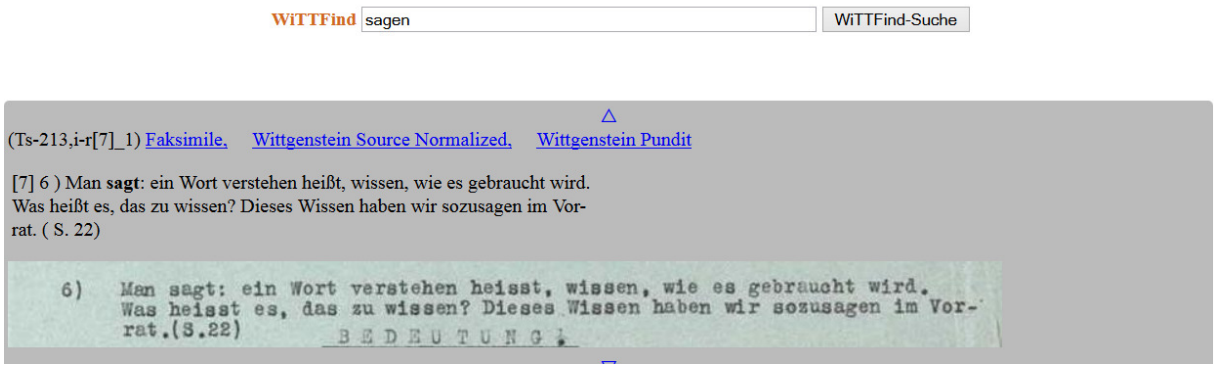
WAB und CIS entwickelten in enger Zusammenarbeit ein einfaches TEI-P5 konformes XML-Format (CISWAB), das einen optimalen Ausgangspunkt für die Zusammenarbeit von Wittgensteinforschern und Computerlinguisten darstellt.

Dazu extrahierten wir aus CISLEX, dem am CIS erstellten Vollformenlexikon des Deutschen, das Speziallexikon wiTTLex. In intensiven Gesprächen mit Computerlinguisten, Editionsspezialisten, Informatikern, Philosophen, einem interdisziplinären Seminar und einer zweitägigen Sommerschule („Digital Wittgenstein Scholarship 2013“) in München wurden der wissenschaftliche Austausch gepflegt und nach und nach Fragestellungen der Wittgensteinforscher in computerlinguistische Verfahren umgesetzt. Diese Verfahren fassten wir unter der Bezeichnung „Wittgenstein Advanced Search Tools“ (WAST) zusammen und implementierten sie in unserem WEB-basierten Finder WiTTFind.

In meinem Vortrag möchte ich WiTTFind vorstellen, die zugrunde liegenden Wittgenstein Advanced Search Tools beschreiben und über unsere Erfahrungen mit der Kooperation Computerlinguistik und Philosophie berichten.

In der folgenden Abbildung zeigen wir das Eingabefeld unseres Finders WiTTFind:

Siehe http://wittfind.cis.uni-muenchen.de:



2 ÖFFENTLICH ZUGÄNGLICHE TEXTE DES NACHLASSES

2.1 TEXT: DAS TEI-P5 KONFORME XML-FORMAT (CISWAB)

Die am WAB in Bergen entstandene XML-Transkription des Nachlasses von Ludwig Wittgenstein annotiert die Texte sehr detailliert: Alle Streichungen, Ergänzungen usw. sind im XML festgehalten. Diese genaue Auszeichnung ist aber für den Einsatz unseres Finders viel zu ausführlich, und so definierten wir ein reduziertes TEI-P5 konformes XML-Format (CISWAB), das eine geeignete Basis für die Zusammenarbeit von Wittgensteinforschern und Computerlinguisten darstellt. CISWAB wird über XSLT-Transformation aus dem umfassenderen WAB XML extrahiert (Øyvind Liland Gjesdal, WAB).

2.2 LEXIKON: DAS ELEKTRONISCHE VOLLFORMENLEXIKON MIT SEMANTISCHEN WORTKLASSEN (WITTLEX)

Das Vollformenlexikon WiTTLex umfasst alle Wörter der auf Wittgenstein Source öffentlich und frei zugänglichen Texte des Nachlasses von Ludwig Wittgenstein und ist im DELA Format, das am Laboratoire d'Automatique Documentaire et Linguistique (LADL, Paris) definiert wurde, gespeichert. Für jedes Wort werden im Lexikon die Vollform, das Lemma, die lexikographische Wortform, semantische Notationen und morphologische Varianten gespeichert (Angela Krey, CIS).

3 DIE WEB-BASIERTE APPLIKATION ZUM FINDEN VON TEXTSTELLEN: WITTFIND

Im Zentrum unserer computerlinguistischen Arbeit entwickelten wir hocheffiziente, parallelisierte C++ Client/Server-Programme (Florian Fink, CIS), die die XML-notierten Texte einlesen, das Vollformenlexikon im Hintergrund halten und alle WAST-Verfahren implementiert haben. Über eine WEB-Schnittstelle können Anfragen gestellt werden, und das Finder-Programm sucht regelbasiert nach Textstellen, die zu dieser Anfrage passen. WEB-Programme bereiten die Ergebnisse für die HTML-Ausgabe auf, extrahieren die zugehörigen Faksimileausschnitte und stellen die Treffer auf der WEB-Seite dar.

4 SUCHANFRAGEN BEI WITTFIND

4.1 LEMMATISIERTE UND INVERSE LEMMATISIERTE WORT-SUCHE

Mit Hilfe der Einträge im Vollformenlexikon WiTTLex können Anfragen an WiTTFind lemmatisiert behandelt werden. Die Suche nach dem Wort “sagen” liefert z.B. alle Textstellen, an denen lexikalische Varianten wie “sagte”, “sagten” usw. vorkommen. Das Lexikon erlaubt auch eine inverse lemmatisierte Suche: Die Anfrage “sagte” führt zum Lemma “sagen”, und daraus werden alle lexikalischen Varianten von “sagen” generiert und danach gesucht.

4.2 SUCHE ÜBER WORTFORMEN

Im Vollformenlexikon WiTTLex sind alle Wörter mit ihrer Wortform gespeichert. Nomen und Adjektive sind semantisch annotiert. Unser Finder erlaubt es, nach Wörtern zu suchen, die diesen Wortformen zugehören. Die Anfrage: „Die <ADJ> Farbe“ findet zum Beispiel alle Sätze, welche die Wortfolge „Die“ gefolgt von einem Adjektiv und dem Wort „Farbe“ enthalten. Die Anfrage: „die <COL>“ findet alle Sätze mit der Wortfolge „die“-gefolgt-von-einer-Farbe, und als weiteres Beispiel findet die Anfrage mit semantischer Wortform: „die <EN>“ alle Sätze, welche die Wortfolge „die“-gefolgt-von-einem-Eigennamen (Olga Strutynska, CIS) enthalten.

4.3 SATZSTRUKTUR UND WILDCARDS

Diese Suche erlaubt dem Nutzer festzulegen, dass ein bestimmtes Wort am Satzanfang <BOS>, bzw. Satzende <EOS> vorkommt. Stellvertretend für ein Wort oder eine Zeichenkette kann auch der Wildcard-Buchstabe „*“ verwendet werden. Die Anfrage: „<BOS> Ich den*“ findet alle Sätze, die mit „Ich“ beginnen und von Wörtern gefolgt werden, die mit „den“ beginnen. Die Anfrage: „<BOS> Ich * * * <EOS>“ findet zum Bespiel alle Sätze, die mit „Ich“ beginnen und vier Wörter lang sind.

4.4 REGELBASIERTE LINGUISTISCHE SUCHE UND PART OF SPEECH TAGGING (POS-TAGGING)

Hier wurde die Möglichkeit der deutschen Sprache implementiert, dass bei Partikelverben die Partikel vom Wortstamm getrennt vorkommen können. Partikelverben werden über unser Lexikon WiTTLex erkannt und zerlegt (Luidmilla Volos, CIS). Um Partikel von Präpositionen zu disambiguieren, verwendet unser Finder ein automatisches Part of Speech Tagging (treetagger von Dr. Helmut Schmid, CIS) und lokale Grammatiken. Die lokalen Grammatiken wurden mit dem graphischen WEB-Tool CisGraph (Shuangjaio Cao, Medieninformatik), der ebenfalls von uns

programmiert wurde, erstellt. Zum Beispiel findet die Anfrage „nachdenken“ jetzt auch Sätze wie: „Wir **denken** nie darüber **nach**, … “

4.5 Suche ohne Alternativen

Ein charakteristisches Merkmal von Wittgensteins Nachlass besteht darin, dass er in den Texten sehr viel änderte und oftmals mehrere alternative Formulierungen anbietet. In den vorliegenden Texten existieren also viele alternative Lesarten, die am WAB in XML kodiert sind. Damit die Suche auch in allen unterschiedlichen Lesarten durchgeführt werden kann, werden bei unserer Suche im Hintergrund alle Lesarten der Texte generiert, durchsucht, und die gefundenen Textstellen mit ihren Alternativen dargestellt (Pattrick Seebauer, CIS).

5 Darstellung der gefundenen Textstellen auch im Faksimile

Gerade bei sehr heterogenen Schriftensammlungen, wie die des Nachlasses von Ludwig Wittgenstein, bei der viele, auch handschriftliche, Texte, vorliegen und vom Autor häufig geändert wurden, ist es für die Wissenschaftler sehr wichtig, die gefundenen Textstellen nicht nur als edierten Text zu sehen, sondern die zugehörigen Stellen auch im Faksimile der Originale studieren zu können. Nur im Bild des Originals bekommen die Forscher die „Aura“ des gefundenen Textes zu spüren, und mit Hilfe des Faksimiles können sie sogar Editionsfehler entdecken. Am CIS (Matthias Lindiger) wurde ein Programm entwickelt, welches die Textedition und das Faksimile auf Bemerkungen-Niveau miteinander verlinkt, und welches es daher erlaubt, von einem Suchergebnis direkt zum entsprechenden Ausschnitt im Faksimile zu springen.

6 Verbindung zu bestehenden Softwaretools aus der Wittgensteinforschung

Es war uns von Anfang an wichtig, dass unser Tool keine neue digitale Insellösung darstellen sollte, sondern dass eine einfache Verbindung zu bestehenden Plattformen und Programmen herstellbar sein soll. Jede gefundene Textstelle kann durch Anklicken sofort in Wittgenstein Source dargestellt werden und mit dem Semantic Web tool Pundit (http://feed.thepund.it) weiter annotiert werden.

7 Zusammenarbeit Computerlinguistik und Philosophie

Die Zusammenarbeit zwischen dem CIS, dem WAB und Wittgensteinforschern (u.a. an der Fakultät Philosophie LMU, Department II) ist sehr intensiv und für beide Seiten sehr anregend.

Die Computerlinguisten realisierten, dass Texte, nicht wie bei herkömmlichen Suchmaschinen, statistische Ereignisse sind, sondern dass bei dieser Art von Texten jedes Wort wichtig ist und die Trefferquote bei 100% liegen muss: „Nichts darf übersehen werden!“. Außerdem erkannten die Computerlinguisten, dass nur solche Tools und Verfahren von den Kooperationspartnern akzeptiert werden, die einen einfachen WEB-basierten Zugang ermöglichen, sich an der Wissenschaftssprache der Wittgensteinforscher orientiert und für deren dahinterliegenden Formalisierungen offen oder sogar frei konfigurierbar sind. Das war ein Grund, weshalb wir von Anfang an einen regelbasierten Ansatz für unseren Finder wählten.

Die Kooperationspartner der Philosophie schätzen die technische Möglichkeit, dass die gefundenen Textstellen in den Faksimiles der Originale, die meist unzugänglich in Archiven aufbewahrt sind, sichtbar gemacht werden können. Nichtdeutschsprachige Forscher sind von der Lemmatisierung ihrer Anfrage sehr erfreut, da sie oftmals der Reichhaltigkeit der deutschen Morphologie nicht in dem Maße mächtig sind. Die Möglichkeit der gleichzeitigen Recherche über mehrere Dokumente hinweg erlaubt es hervorragend, Ähnlichkeiten und Textgenesen zu entdecken, was schon immer ein zentraler Forschungsgegenstand der Wittgensteinforschung war.

So wird aus der Zusammenarbeit der Computerlinguisten und der Wittgensteinforscher ein ständiger Kreislauf, bei dem die Philosophen zur Formalisierung ihrer Fragestellungen aufgefordert werden, die Computerlinguisten Verfahren entwickeln müssen, welche die anspruchsvollen Formalisierungen effizient implementieren, und die neuen Anfragemöglichkeiten wiederum zu einer erneuten Korrektur und Erweiterung der Formalisierung des Findens führen.

Wie schreibt schon Ludwig Wittgenstein im Ms111,178: "Wenn ich etwas suche, so ist es wesentlich, daß ich das Finden ebenso ausführlich muß beschreiben können (ob es (je so) eintritt oder nicht) ehe der Gegenstand gefunden ist.“

So nennen wir unser Tool WiTTFind nicht eine Suchmaschine, sondern einen Finder.