Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind

Max Hadersbeck, Alois Pichler, Florian Fink, Daniel Bruder, Ina Arends
Maximilian.Hadersbeck@lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München,
Wittgenstein Archives at the University of Bergen (WAB).

1 Einleitung

In dem Vortrag berichten wir über Erfahrungen, Erkenntnisse und Erweiterungen unserer schon seit 2 Jahren im Einsatz befindlichen FinderApp WiTTFind, die mit Hilfe von computerlinguistischen Verfahren den Open Access zugänglichen Teil des Nachlasses von Ludwig Wittgenstein (Wittgenstein Source, 2009) nach Wörtern, Phrasen, Sätzen und semantischen Begriffen im „Zusammenhang des Satzes“[1] durchsucht.

Im Sommer 2014 gewannen wir mit WiTTFind den EU-AWARD, der vom EU-Projekt Digitised Manuscripts to Europeana (DM2E) ausgeschrieben wurde, verbunden mit der expliziten Aufforderung zur Öffnung unseres Finders für andere Projekte der Digital Humanities. Darauf hin entwarfen wir in der disziplinübergreifenden Wittgenstein Sommerschule am CIS im Juni 2014 und in Diskussionen mit Fachleuten der Philosophie und Digital Humanities Verbesserungsmöglichkeiten, die mittlerweile in der neuen Version implementiert sind. Die Weboberfläche unseres Finders wurde optimiert, („rich-client“), jetzt können mehrere Dokumente parallel durchsucht werden, eine lemmatisierte symmetrische Vorschlagssuche und ein Faksimile E-Reader sind integriert. Der Faksimile E-Reader erlaubt es nun, dass die Faksimiles der Edition durchblättert und gefundene Textstellen automatisch visuell hervorgehoben werden. Neben den Weiterentwicklungen der FinderApp setzten die Wittgensteinforscher unseren Finder für semantische Untersuchungen ein und gewannen aus dieser Arbeit wichtige Erkenntnisse z.B. zum Thema des Verstehens in Wittgensteins Big Typescript.[2]

Der wichtigste Mehrwert unseres Finders besteht allerdings darin, dass wir die vom EU-AWARD geforderte Öffnung unseres Finders für andere Projekt konsequent umsetzten. Für die Texte der Edition, die unser Finder durchsucht, gibt es eine XML-TEI P5 kompatible Document Type Definition (DTD). Die Programme, Faksimile E-Reader und Tools sind unter der Bezeichnung „Wittgenstein Advanced Search Tools“ (WAST) in einem „docker“-Softwarecontainer zusammengefasst und werden „open source“ verfügbar sein. Somit ist unsere FinderApp mit ihren WAST-Tools in anderen Projekten der Digital Humanities einsetzbar.

Die folgende Abbildung zeigt eine Suchanfrage an unseren Finder WiTTFind:

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

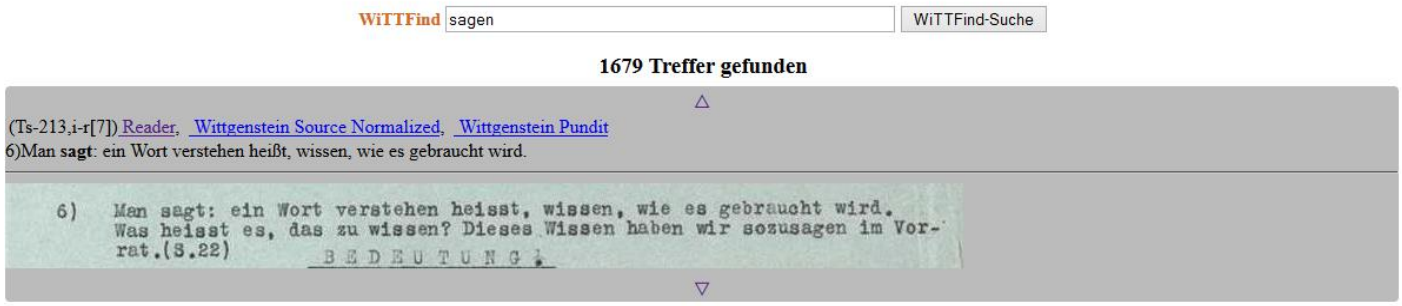

Bild 1: Suchanfrage bei WiTTFind

[1] http://www.wittgensteinsource.org/Ts-213,1r[4]_n

[2] http://www.wittgensteinsource.org/Ts-213_n

2 Erkenntnisse aus der Zusammenarbeit Computerlinguistik und Philosophie

2.1 Verbesserte Benutzeroberfläche unseres Finders

Eine der ersten Erkenntnisse unserer Zusammenarbeit war, dass die Benutzeroberfläche unserer FinderApp auf die Bedürfnisse der jeweiligen Forschergruppe abgestimmt sein muss: die Forscher sollen sich auf der Webseite „wiederfinden“. Nur dann ist die Einstiegshürde nicht zu hoch, und die Bereitschaft mit dem Finder zu arbeiten steigt. Erst für fortgeschrittene Benutzer werden in einer tieferen Schicht globale Einstellungsmenus sichtbar und spezielle Parameter einstellbar. Als Kompromiss zwischen Komplexität und gewohnter Suchmaschinenarbeit können die Nutzer verschiedene Suchumgebungen auswählen (siehe Bild 1): „Regelbasiertes Finden“, „Semantisches Finden“, „Graphisches Finden“, „Statistische Suche“ und „Geheimschriftübersetzer“.

Bild 2: Suchumgebungen bei WiTTFind

Damit die zahlreichen Suchmöglichkeiten bei WiTTFind auf einen Blick sichtbar sind, programmierten wir fachspezifische Hilfeseiten mit Beispielen:

Bild 3: Hilfeseiten bei WiTTFind

2.2 Video-Tutorials zur Nutzung von WiTTFind

Zum erleichterten Einstieg bei WiTTFind gibt es jetzt zwei Video-Tutorials in deutscher und englischer Sprache unter folgendem Link:

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

2.3 E-Reader für die Faksimile

Gerade bei komplexen Editionen mit vielen handschriftlichen Einfügungen und Streichungen, wie der des Nachlasses von Ludwig Wittgenstein, ist es für die Editionswissenschaftler eine Herausforderung, den Editionstext in der niedergeschriebenen Form als HTML-Text in einem Browser darzustellen. In der neuen Version unserer FinderApp programmierten wir einen eigenen Faksimile E-Reader, der es erlaubt, komplementär durch die Faksimile der Edition zu blättern und gleichzeitig die gefundenen Textstellen im Bild hervorhebt.

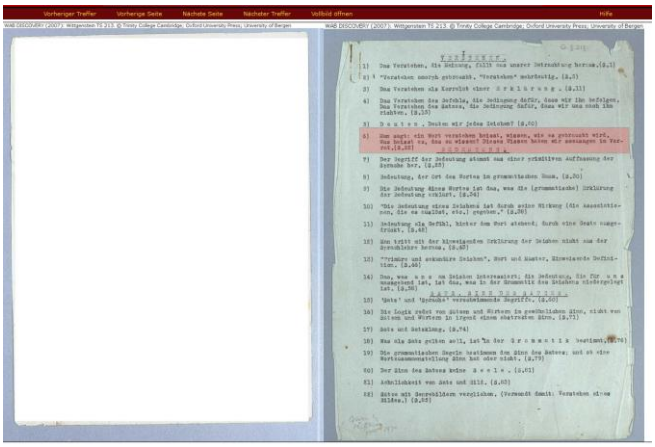

Bild 4: Faksimile Reader bei WiTTFind

2.4 Lemmatisierte Vorschlagssuche mit statistischen Angaben

Die Arbeit mit WiTTFind zeigte, dass eine komfortable Vorschlagssuche, die den gesamten Wortindex der Edition mit Frequenzlisten im Hintergrund hält, einen sehr guten Einstieg in die eigentliche Suche darstellt. Hierhin zielt unsere neueste Erweiterung von WiTTFind, eine komfortable Index-Suchfunktion, die auf einen symmetrischen Suchindex basiert. Dieser Index greift auf Einträge des zugrunde liegenden Lexikons und Wort-Frequenzlisten der Texte der Edition zurück. Dem Anwender werden nach Eingabe von wenigen Buchstaben alle Wörter mit der Häufigkeit des Auftretens im Text automatisch aufgezeigt, in denen die eingegebenen Buchstaben vorkommen; dazu werden auch noch die morphologischen Varianten dieser Wörter angezeigt. Diese Art der Autovervollständigung ist eine völlig

neue Technologie, da bisherige Autovervollständigungen die eingegebenen Buchstaben nur um die Wörter ergänzen, die mit diesen Buchstaben beginnen.

3 Von Daten zu Erkenntnissen

3.1 Semantisches Suchen: Wortfelder

Ein großes Problem semantischer Untersuchungen mit Wortfeldern stellt die Disambiguierung der Wortfeldbegriffe dar. Mit Hilfe unseres elektronischen Lexikons, der syntaktischen und semantischen Disambiguierung über Part of Speech Tagging und lokale Grammatiken können neben Einzelwörter auch Wortphrasen einem Wortfeld zugeordnet und disambiguiert werden.
Ein einfaches Beispiel wurde um das semantische Feld von "Verstehen" ausgearbeitet. Welches Interesse an Verstehen hat Wittgenstein im Big Typescript? Eine Suche nach *<N> verstehen* [Substantiv + „verstehen“] im Big Typescript ergibt, dass dort ganz klar das Verstehen von Wörtern, Sätzen, Sprachen, Befehlen ... allgemein: das Verstehen von sprachlichen Zeichen, im Vordergrund steht. Daneben gibt es aber auch bereits eine gewisse Aufmerksamkeit auf das Verstehen von Menschen und Menschlichem: von Handlungen, Gebärden, Gesten. Diese Aufmerksamkeit nimmt in Wittgensteins Spätwerk beständig zu, was eine Suche nach *<HUM> verstehen* [Substantiv für Menschliches + „verstehen“] bestätigt.
Ein zweites, komplexeres Beispiel wurde um das semantische Feld von "Grammatik" ausgearbeitet. Zuerst baten wir Wittgensteinexperten, uns eine Liste von 10-15 Wörtern zu geben, welche ihrer Ansicht nach im Wortfeld von "Grammatik" zentral sind. Dazu gehören z.B. "Anwendung", "Regel", "Kalkül" und "System". Daraufhin wurden diese Wörter im Lexikon über den Begriff "Grammatik" vernetzt. Eine WiTTFind-Suche nach *Grammatik* wird dann nicht nur Stellen mit "Grammatik" ergeben können, sondern auch Bemerkungen, welche eine Bündelung von Begriffen aus dem Wortfeld aufweisen. Erste Anwendungen ergaben, dass Wittgenstein im Big Typescript tatsächlich einen regelfixierten Begriff von Grammatik verfolgt, während dieser Aspekt später abgeschwächt werden wird (vgl. Szeltner 2013).

4 Synergien: Unsere FinderApp für andere Digital Humanities Projekte

4.1 Vorbemerkung

Wie vom DM2E Projekt bei der Preisverleihung gefordert, öffneten wir unsere FinderApp für andere Projekte der Digital Humanities. Editionsprojekte müssen ihre Dokumente in unser reduziertes XML-TEI P5 Format (CISWAB) konvertieren und die Open-Source Software *docker*[3] auf ihrem Rechner installieren. Dann können sie unseren Finder bei ihren Editionstexten anwenden. Zur Darstellung und Highlighting der Treffer im Faksimile sind allerdings umfangreiche OCR-Arbeiten notwendig. In den nächsten Unterkapiteln beschreiben wir im Detail, wie unser Finder einsetzbar wird.

4.2 Die Texte der Edition

Unsere FinderApp findet Wörter, semantische Begriffe und Satzphrasen über mehrere Dokumente hinweg, sofern die Dokumente in unserem XML-TEI-P5 Format vorliegen. Wir nennen dieses XML-Format CISWAB und beschreiben es in einer eigenen Document Type Definition (DTD). Die einzelnen Dokumente sind bis auf Satzebene über Siglen eindeutig zu spezifizieren:
(z.B. `<s n="Ts-213,i-r[7]_1" ana="facs:Ts-213,i-r abnr:7 satznr:15">6)Man sagt: ein Wort verstehen heißt, wissen, wie es gebraucht wird.</s>`)

4.3 Elektronisches Vollformenlexikon

Zu den Texten einer Edition benötigt unsere FinderApp ein elektronisches Lexikon im DELA Format (Laboratoire d'Automatique Documentaire et Linguistique, Paris). Bei der Erstellung des Lexikons können wir behilflich sein, da wir am CIS das größte deutsche Vollformenlexikon erstellt haben.

4.4 Syntaktische Disambiguierung: Part of Speech Tagging

Grundvoraussetzung für die syntaktische Disambiguierung ist es, dass die Texte mit einem Part of Speech Tagger bearbeitet werden. Zu unseren WAST-Tools gehört das automatische Taggen der Texte. Dazu verwenden wir den *treetagger* von Dr. Helmut Schmid, der am CIS entwickelt wird. Der *treetagger* konvertiert

[3] siehe: https://www.docker.com/

die Textdatei in eine getaggte XML Datei, die die Eingabedatei für unsere FinderApp darstellt.

4.5 Darstellung der Treffer im Faksimile Reader

Um die Treffer in unserem Faksimile-Reader darzustellen, müssen die Faksimile mit der open source Software *tesseract* bearbeitet werden, und je nach Qualität der Faksimiles manuell nachbearbeitet werden. Wir entwickelten Tools, die diese manuelle Arbeit erleichtern.

4.6 Praktische Voraussetzung zur Verwendung unserer FinderApp

Wir haben unser Ziel, dass die FinderApp WiTTFind und die WAST-Tools möglichst auf jedem Rechner lauffähig sind, erreicht. Mit Hilfe der neuesten Open Source Software Technologie *docker* werden die unterschiedlichen Programmiersprachen und Libraries, die wir einsetzen, in einem Softwarecontainer, genannt WAST-dockerimage, zusammengefasst. Jeder Anwender, der auf seinem Rechner die *docker*-Serversoftware installiert hat, kann das WAST-dockerimage herunterladen und virtualisiert läuft die FinderApp WiTTFind unter dem Dockerserver auf dem Rechner. Die Dockerserversoftware funktioniert nahezu unter jedem Betriebssystem (Linux, Windows, MACOS).

4.7 Vorstellung und Vorführung unseres Finders auf der Tagung

Neben diesem Vortrag wollen wir auf der Tagung in einem Poster den Aufbau und den Einsatz der FinderApp WiTTFind als Open Source Tool vorstellen: Die optimierte Browseroberfläche, zugrunde liegende Texte der FinderApp, Faksimile mit OCR, Faksimile Reader und den Einsatz des Finders als Open Source Programm. Für Interessierte wird die FinderApp unter verschiedenen Betriebssystemen an Laptops vorgeführt.

5 EU-AWARD und Publikationen

EU AWARD 2014: http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal: Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST). Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014) Madrid: 91-96

Szeltner, Sarah: 'Grammar' in the Brown Book. Papers of the 36th International Ludwig Wittgenstein-Symposium, vol 21. Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society; 2013.

Wittgenstein Source: Bergen Text and Facsimile Edition. In: Pichler A., collaboration with, Krüger H.W., Lindebjerg A., Smith D.C.P., BruvikT.M., Olstad V., editors. Bergen: Wittgenstein Archives at the University of Bergen; 2009. http://www.wittgensteinsource.org/