FemSMA Corpus Workbench.
Ein Werkzeug zur Unterstützung der qualitativen und quantitativen Analyse von textuellen Daten*

Brigitte Krenn (Wien)

Abstract

In various areas of (linguistic) research, there is a need to analyse larger amounts of textual data. Digitisation and the availability of computational linguistics tools offer substantial support in qualitatively and quantitatively analysing those data sets. Keeping, maintaining and presenting data and their metadata within one system facilitate data inspection and browsing. Quick assessment of data sets for the presence or absence of specific textual characteristics is supported by the possibility to manually annotate segments of text with theory-driven meta-information in combination with automatic analysis employing computational linguistics tools and computerized search.

In the present contribution, the FemSMA Corpus Workbench CWB is introduced. CWB is a computational linguistics tool for manual and automatic annotation and analysis of text documents. CWB supports storage and maintenance of, and annotation and search in textual data and related metadata. CWB is a client-server application with a web interface as frontend for data inspection and manual annotation. Data storage and automatic processing is done at server side. Automatically annotated are word-level features such as parts of speech; general word features such as capitalisation, character reduplication, abbreviation; swear words and emotion words. Due to its modular system architecture, CWB can be flexibly extended, which, however, requires the involvement of computational linguists to adapt and extend CWB's automatic analysis and search functionalities, and represent the new functionality in the web interface.

*__Danksagung:__ Die Entwicklung und Implementierung der CWB wurde ermöglicht durch das Forschungsprogramm FEMTech „Frauen in Forschung und Technologie“ des Österreichischen Bundesministeriums für Verkehr, Innovation und Technologie, konkret im Rahmen des Forschungsprojektes „FemSMA – Automatisierte, gendersensible Verfahren zum Ausbau von *Social Media* Analysen als EDV-gestützte Forschungsmethodik“. Besonderer Dank gilt meinem OFAI-Kollegen Johannes Matiasek für die technische Konzeption und Implementierung und Karin Wetschanow für ihre Beiträge zur Konzeption der CWB und ihre Anregungen zur konkreten Umsetzung der Systemfunktionalität aus Anwendersicht. Weiterer Dank gilt dem/der anonymen Reviewer_in für Kommentare hinsichtlich der besseren Verständlichkeit der Darstellung der CWB.

Anfragen bezüglich einer Nutzung der CWB: brigitte.krenn@ofai.at

1 Einleitung

In verschiedenen Bereichen der (linguistischen) Forschung besteht die Notwendigkeit Sammlungen von Texten anhand theoretischer Fragestellungen qualitativ zu untersuchen. Das Vorliegen digitalisierter Texte und der Einsatz computerlinguistischer Werkzeuge ('Tools') sind eine hilfreiche Unterstützung bei der qualitativen Analyse von Text-Korpora. Vor allem wenn größere Textmengen bearbeitet werden sollen, stellt bereits die Möglichkeit der Datenhaltung und -bearbeitung in einem Gesamtsystem eine große Arbeitserleichterung dar. Des Weiteren lassen sich Textsammlungen durch die Möglichkeit einer flexiblen manuellen Annotierung in Kombination mit der automatischen Verarbeitung von Texten und mit computergestützter Suche viel schneller auf das Vorhandensein, bzw. die Ausprägung bestimmter Merkmale untersuchen als es mittels rein manueller Vorgehensweisen möglich ist.

Im vorliegenden Beitrag wird die FemSMA[1] Corpus Workbench (CWB) vorgestellt, als aktuelles Beispiel für ein computerlinguistisches Instrument, das folgende Funktionalitäten verbindet: automatische Suche in Textdokumenten, manuelle und automatische Annotierung von Texten mittels computerlinguistischer Analysetools. Die CWB wurde ursprünglich mit dem Ziel entwickelt Social Media Postings zu analysieren und zu annotieren, (i) um zu studieren inwieweit Autor_innengender anhand von Textmerkmalen in den Postings vorhergesagt werden kann; (ii) um ein Referenzkorpus zum Training von statistischen Modellen aufzubauen, die zur Klassifizierung von Texten nach Autor_innengeschlecht herangezogen werden können. Obwohl die CWB ursprünglich für die Analyse von genderspezifischen Merkmalen in Social Media Texten entwickelt wurde, kann sie für die Untersuchung anderer textlinguistischer Fragestellungen eingesetzt und adaptiert werden, wie z. B. zur Analyse von Zitationspraxen in wissenschaftlichen Texten.

Die CWB wurde insbesondere für die Zusammenarbeit von Linguist_innen und Computerlinguist_innen entwickelt. Ein gemeinsam von Linguist_innen und Computerlinguist_innen getriebener Analyse- und Annotationsprozess ist besonders von Bedeutung, wenn größere Mengen an textuellen Daten sowohl von einer theoriegetriebenen (*top-down*) als auch einer datengetriebenen (*bottom-up*) Perspektive untersucht werden sollen. Auf diese Weise können qualitative und quantitative Untersuchungen textueller Merkmale miteinander verschränkt werden, wobei die automatischen, unter Einsatz von computerlinguistischen Werkzeugen durchgeführten Analysen als Vorverarbeitung und Unterstützung für die manuellen qualitativen Analysen dienen und somit die Arbeit der Linguist_innen unterstützen.

Der Grundgedanke hinter der CWB ist, dass in einem ersten, manuellen Arbeitsschritt bestimmte für die Analyse relevante Textabschnitte (Belegstellen) gekennzeichnet und beispielhaft manuell annotiert werden. Für das Auffinden entsprechender Belegstellen in einem vorliegenden Corpus steht eine Suchfunktionalität auf Basis von regulären Ausdrücken zur Verfügung.[2] Diese ermöglicht es den menschlichen Expert_innen einen raschen Überblick

[1] Forschungsprojekt „FemSMA – Automatisierte, gendersensible Verfahren zum Ausbau von *Social Media* Analysen als EDV-gestützte Forschungsmethodik". Siehe http://femsma.ofai.at/.

[2] Reguläre Ausdrücke sind eine Art formale Sprache mittels derer generalisierte Muster definiert werden können, die dazu eingesetzt werden, um in Texten bestimmte Zeichenketten zu finden. Zum Beispiel passt der reguläre

über das Vorhandensein bestimmter Textmerkmale in einem vorgegebenen Korpus zu bekommen. Auf diese Art und Weise kann schnell festgestellt werden, ob der vorhandene Datensatz für die Analyse bestimmter Textmerkmale geeignet ist, oder ob dieser um weitere Texte mit entsprechenden Textmerkmalen ergänzt werden muss.

Eine weitere Suchfunktion der CWB ermöglicht es, sich bereits manuell annotierte Textpassagen gruppiert nach ihrer Klassifizierung („Labels“) in Listen ausgeben zu lassen. Diese Listen wiederum dienen den Computerlinguist_innen im Team als Grundlage um zu bewerten, welche der für eine (text)linguistische Analyse relevanten Merkmale mittels welcher computerlinguistischer Verfahren und mit welcher Treffsicherheit automatisch identifiziert und annotiert werden können. Somit lässt sich abschätzen, welche der in der CWB aktuell vorhandenen Analysefunktionalitäten zur Merkmalsanalyse eingesetzt werden können und welche weiteren Funktionalitäten speziell für die zu analysierenden Merkmale zusätzlich implementiert werden sollten.

Ziel der automatischen Annotierung ist es, möglichst viele Belegstellen im gesamten Textkorpus zu identifizieren und diese mit entsprechenden Labels zu versehen. Die jeweilige Qualität der automatischen Annotierungen wiederum muss manuell überprüft und gegebenenfalls müssen die automatischen Annotierungen manuell korrigiert werden. Die so gewonnenen Belegstellen können für maschinelles Lernen von Klassifiern (Manning, Schütze 1999: 575–608, Kapitel 16 „Text Categorization“) verwendet werden. Dieses Ineinandergreifen von automatischer Annotierung und manueller Korrektur entspricht dem Bootstrapping eines annotierten Referenzkorpus (*Gold Standard Corpus*) in der Computerlinguistik, siehe z. B. Baroni und Bernardini (2004).

In den nachfolgenden Abschnitten wird die aktuell vorhandene Funktionalität der CWB anhand von Fallbeispielen aus der Social Media Analyse von Autor_innengender genauer beschrieben: siehe Abschnitt 0 für die manuelle Annotierung, Abschnitt 0 für die automatische Verarbeitung von Texten mittel computerlinguistischer Werkzeuge, sowie Abschnitt 0 für die in der CWB implementierte automatische Suche in den Textdokumenten. In Abschnitt 0 wird diskutiert, wie die Funktionalität der CWB auf die Analyse von Zitationspraxen in wissenschaftlichen Texten, wie in Wetschanow (in diesem Heft) dargelegt, umgelegt werden kann. Abschnitt 0 liefert eine Zusammenfassung und einen Ausblick zur geplanten weiteren Entwicklung der CWB.

2 Corpus Workbench CWB

Die CWB ist eine ajax-basierte[3] (Client-Server) Webapplikation, ein Werkzeug für (i) die Verwaltung und das Durchblättern von Textdokumenten, (ii) die manuelle Annotierung von Textstellen, (iii) die automatische Tokenisierung, d. h. die Zerlegung von Texten in einzelne Wörter, und deren Annotierung mit Merkmalen aus verschiedenen Merkmalsklassen, dazu gehören morphosyntaktische Klassen wie Nomen, Verb, Adjektiv etc. (siehe Abschnitt 0 für

Ausdruck [0-9]+jährig unter anderem auf folgende Zeichenketten *100jährig, 39jähriger, 2jähriges* usw. Die in der CWB vorhandene Funktionalität zum Schreiben regulärer Ausdrücke entspricht jener der Programmiersprache Perl, siehe www.cs.tut.fi/~jkorpela/perl/regexp.html.

[3] Garrett (2005)

eine genauere Erklärung zur Tokenisierung), (iv) die Suche von Textstellen anhand deren manueller Annotierung, mittels Volltextsuche und Suche basierend auf regulären Ausdrücken.

Der serverseitige Teil der CWB besteht aus einer Datenbank[4], dem zentralen Datenspeicher, in dem die Textdokumente mit ihren Metadaten abgespeichert sind, siehe Abbildung 1. Das Kernstück der Metadaten sind die manuell erstellten Annotierungen der Textsegmente anhand von Labelgruppen und Labels, welche theoriegetrieben von den Linguist_innen entwickelt werden. Labelgruppen, Labels und Annotierungen gelangen mittels eines Webinterfaces und entsprechender Systemfunktionalität für die manuelle Annotierung – der Annotierkomponente – in die Datenbank. Für die Tokenisierung und automatische Annotierung der Dokumente mit Merkmalen aus diversen Merkmalsklassen sorgt die Tokenisierungskomponente. Die Suchkomponente operiert ebenso wie die Tokenisierungskomponente auf den Textdokumenten. Die Suche erfolgt auf den Korpusdokumenten und gibt je nach Anfrage die folgenden Ressourcen aus: (i) die gefundenen Dokumente, (ii) eine alphabetisch und nach Labelgruppen und Labels geordnete Liste von Belegstellen, (iii) eine Liste von Textstellen, die aufgrund einer Volltextsuche oder einer Suche mittels eines regulären Ausdrucks gefunden wurden.

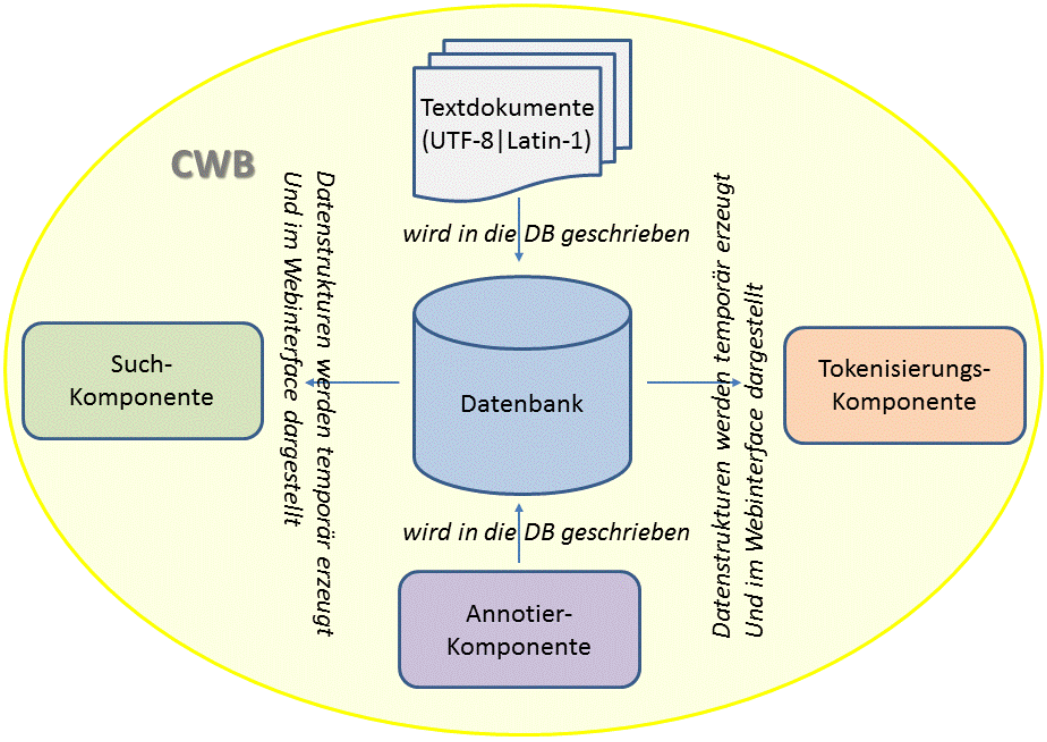

Abbildung 1: CWB – Architektur und Systemkomponenten

[4] Für die technische Umsetzung der Datenbank wurde SQLite verwendet (www.sqlite.org), da SQLite Open Source Technologie sowie kostenfrei ist, und im Vergleich zu anderen Datenbanktechnologien einfach zu betreiben ist, so ist z.B. kein eigener Datenbankserver notwendig und die Datenbank wird in einem einzigen File gehalten. Eine SQLite Datenbank erfordert keine Administration und ist daher besonders geeignet in Kontexten, in denen es keinen IT-Support gibt.

3 Funktionalitäten für die manuelle Annotierung

Der Hauptzweck der CWB ist, Personen dabei zu unterstützen, Textdokumente zu annotieren. Annotierungen in der CWB erfolgen über die Zuordnung von Labels zu einzelnen Wörtern oder Textstellen in einem Dokument. Diese Labels werden nach theoretischen Kriterien definiert, je nach dem, welchem Untersuchungszweck die Annotierungen dienen sollen. In FemSMA wurden die Labels aufgrund genderlinguistischer Fragestellungen und der Einbeziehung von genderlinguistischen Theorien erstellt, in Abschnitt 0 werden die Labels für die Annotierung anhand der in Wetschanow (in diesem Heft) vorgeschlagenen Analyse von Zitationspraxen in wissenschaftlichen Texten definiert.

Wesentliche Eigenschaften der Annotierfunktionalität in der CWB sind:

1. Ein Dokument kann mit verschiedenen Annotierungen versehen werden.
2. Bei jeder Annotierung wird mitgespeichert, wer die annotierende Person war.
3. Innerhalb eines Dokuments dürfen Annotierungen von einer Annotierer_in und ein und derselben Labelgruppe nicht überlappen.
4. Annotierungen von unterschiedlichen Annotierer_innen können überlappen. Von zwei sich überlappenden Labels wird nur der zweite angezeigt.
5. Die Annotierung eines Dokuments ist nur möglich, wenn eine Person explizit als Annotierer_in ausgewählt ist (siehe Abbildung 3, rechte Seite) und eine Labelgruppe für die Annotierung selegiert wurde (Abbildung 4, Mitte).

Die Punkte 1 und 2 gewährleisten, dass ein und dasselbe Dokument von mehreren Personen annotiert werden kann, ohne dass die Annotierer_innen die Annotationen der jeweils anderen Person(en) sehen. Dies ist die Voraussetzung für die Erstellung von Referenzkorpora (in der Computerlinguistik auch Gold Standard Korpora genannt). Die in den Punkten 3 und 4 angesprochenen Möglichkeiten bzw. Unmöglichkeiten der Überlappung von Labels hängen mit der html-basierten Darstellung der annotierten Textstellen im Webinterface zusammen, und sind einer technisch-formalen Einschränkung von HTML geschuldet, nämlich dass in HTML sich überlappende Strukturen nicht zulässig sind.

Abbildung 2 zeigt die Eingangsseite der CWB: In einer Scrolldown-Liste („choose resource“) werden alle Einzelressourcen, die in der Datenbank vorhanden sind, aufgelistet. Die über diese Liste zugänglichen Texte bilden das Gesamtkorpus. Im FemSMA sind die einzelnen Textdokumente nach ihrer Zugehörigkeit zu bestimmten Foren oder Subforen zusammengefasst, wie zum Beispiel zum Unterthema „Bitte um Bewertung meiner Hormonwerte“ im Forum Gynäkologie, oder Forumpostings zum Artikel „Berufsarmee ist nicht gleich Berufsarmee“ der Onlinezeitung derstandard.at. Aus dieser Liste wird die aktuell zu annotierende Ressource ausgewählt.

Im Bereich „Select Annotations“ können einerseits Namen für Annotierer_innen vergeben werden, bzw. aus einer bereits vorhandenen Namensliste der Name der Annotierer_in, die den Text aktuell annotieren wird, ausgewählt werden (siehe Abbildung 3). Andererseits werden in diesem Bereich die Labelgruppen (Label Group) und Labels für die Annotierung vergeben, bzw. aus bereits vorhandenen Labelgruppen diejenige ausgewählt, deren Labels gerade für die Annotierung verwendet werden sollen. Das zeigt Abbildung 4: Die linke Seite zeigt das Interface zur Definition von neuen Labelgruppen und zugehörigen Labels. Neben der

Bezeichnung für die Labelgruppe werden jeweils ein kurzer Labelname, eine Beschreibung und die Farbkodierung für den Label angegeben. Es können beliebig neue Labels und Labelgruppen definiert werden. Sie spiegeln den theoretischen Zugang zur Analyse der zu annotierenden Phänomene wider. Im annotierten Text erscheint die Annotierung in der jeweiligen Labelfarbe (siehe Abbildung 5).

Ist ein Dokument aus einer bestimmten Ressource ausgewählt, wie z. B. in Abbildung 5 ein Posting von User „Zuckerwatte“ aus „Haarausfall Forum Frauen/Superunglücklich“, werden für diese Ressource eine Reihe von Informationen angezeigt, wie: um welche Art von Ressource (Type = Forum) und welches Thema (Topic = Haarausfall) es sich handelt, die Quelle (bei Social Media Dokumenten ist das der URL der Ressource), wie viele Texte unter der Ressource subsumiert sind („Number of messages“), wie lange (Anzahl der Zeichen) ein Text der Ressource durchschnittlich ist („Average message length“), von wie vielen Personen die Texte stammen und von wie vielen davon das Geschlecht bekannt ist („User statistics“), welche Labelgruppen annotiert wurden und wie viele Annotierungen es pro Labelgruppe gibt („Annotation counts per label group“). Des Weiteren werden die Terme ausgegeben, die für die Ressource typisch sind. Dies erfolgt mittels Gewichtung der Terme basierend auf einer Kombination von Termfrequenz (Term Frequency TF) und inverser Dokumentfrequenz (Inverse Document Frequency IDF),[5] siehe Manning, Schütze 1999: 541–544 „15.2.2 Term weighting“. Je nach Wichtigkeit werden die Terme mit unterschiedlicher Schriftgröße dargestellt (Wichtigkeit entspricht Größe), siehe z. B. „Haare, Pille, Haarausfall, Arzt“ im vorliegenden Beispiel. Die einer Ressource zugehörigen Dokumente (Texte) können der Reihe nach für ihre Bearbeitung zur Ansicht gebracht werden (siehe die Prev- und Next-Buttons), bzw. über ihren Index aufgerufen werden. (Die Dokumente sind aufsteigend von 1 bis *n* nummeriert.) In der vorliegenden Beispielansicht ist das Dokument Nummer 1 präsentiert und die Labelgruppe „Gender indexicality“ ist für die Annotierung ausgewählt. Die im Beispiel markierten Stellen sind Belegstellen für Gender indexicality female, d. h. explizite Äußerungen im Text, die darauf hinweisen, dass es sich bei der Autor_in um eine Frau handelt (zu Gender Indexicality siehe Ochs: 1992; Kotthoff: 2012). Rechts vom Text wird angezeigt, welche Labelgruppe mit den entsprechenden Labels gerade für die Annotierung aktiv ist. Das zu annotierende Label wird angeklickt und mit der Maus werden die Belegstellen markiert. Annotierungen müssen explizit gespeichert werden („Save Annotations“) und können jederzeit wieder gelöscht werden („Delete Selected“, „Delete All“).

[5] TFIDF ist ein Maß um zu bewerten, wie wichtig ein Wort für ein Dokument aus einer Sammlung von Dokumenten (Korpus) ist. In je weniger Dokumenten aus der Sammlung ein Wort vorkommt, desto wichtiger ist es für die Dokumente, in denen es vorkommt. *Term Frequency* TF ist die Anzahl, wie oft ein Wort in einem Dokument vorkommt. *Inverse Document Frequency* IDF ist ein Maß dafür, wieviel Information ein Wort trägt, d. h. ob es in vielen oder wenigen Dokumenten des Korpus auftritt. IDF wird berechnet, indem die Gesamtanzahl der Dokumente im Korpus durch die Anzahl der Dokumente, in denen das Wort vorkommt, dividiert wird und dann der Logarithmus des Quotienten berechnet wird. TFIDF ist schließlich die Multiplikation der TF- und IDF-Werte. TFIDF ist null für Wörter, die in allen Dokumenten vorkommen. Je höher der TFIDF-Wert für ein Wort ist, desto spezifischer ist es für das jeweilige Dokument.

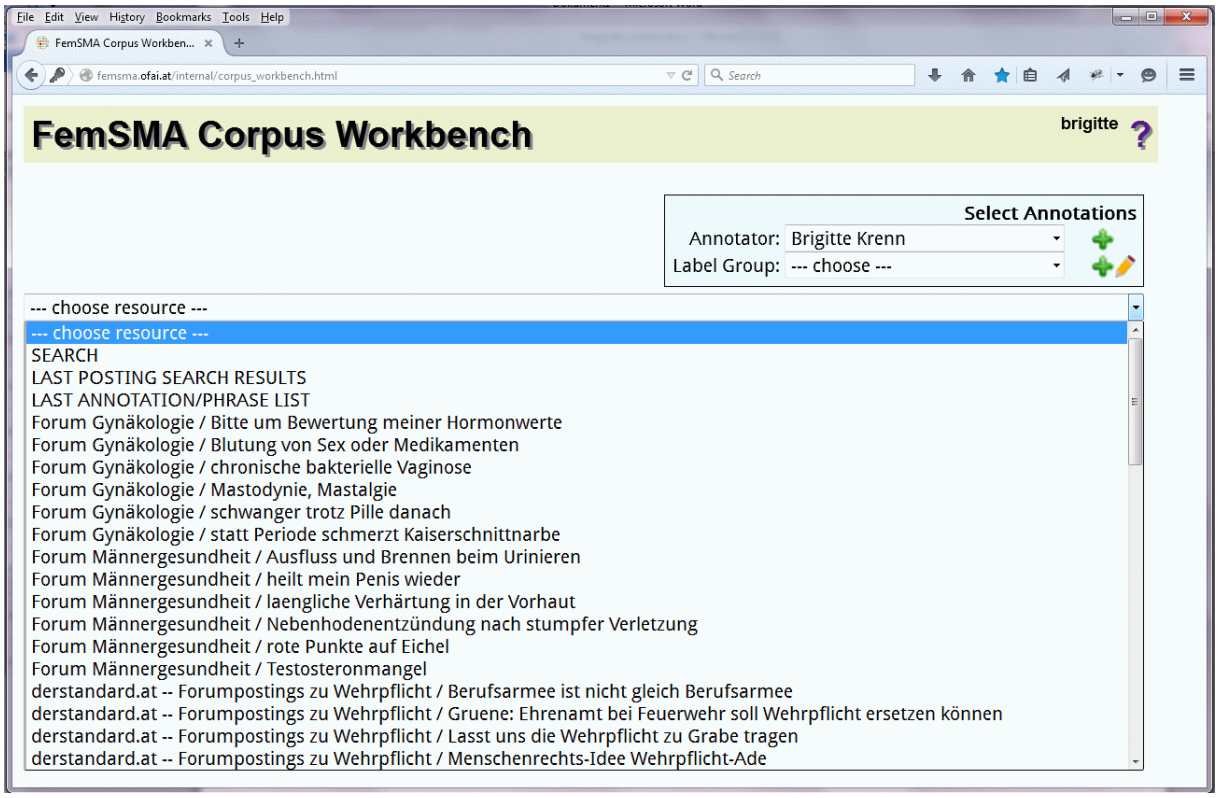

Abbildung 2: Eingangsseite CWB

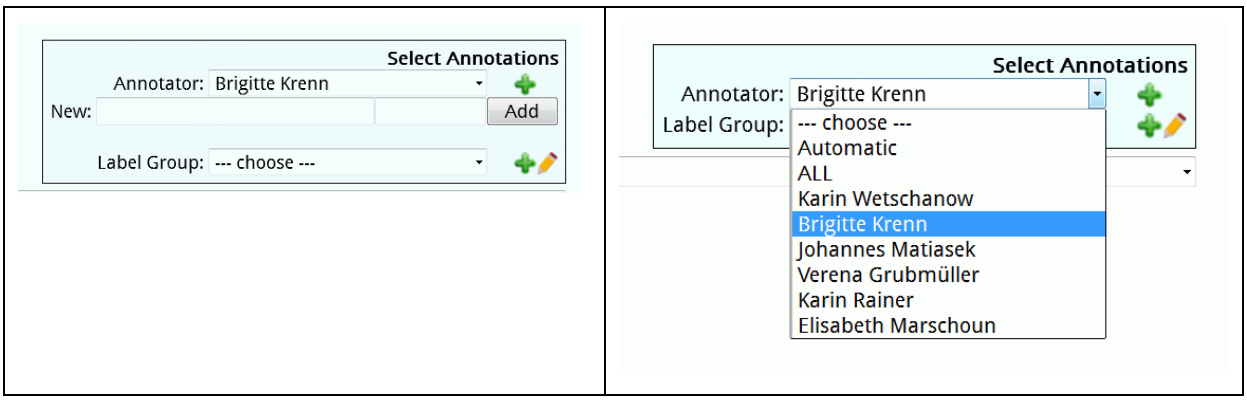

Abbildung 3: Funktionalität für die Bestimmung und Auswahl von Annotierer_innen: linke Seite – Hinzufügen neuer Namen („New“); rechte Seite – Auswahl aus bereits vorhandener Annotierer_innenliste („Annotator“)

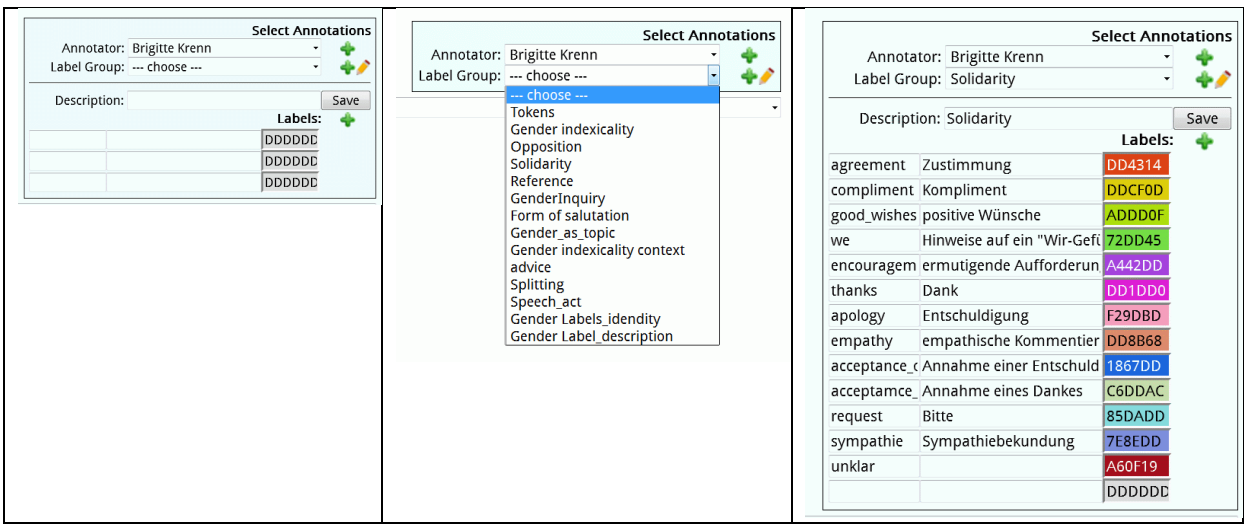

Abbildung 4: Select Annotations: linke Seite -- neue Labelgruppe und zugehörige Labels werden definiert; Mitte -- Auswahl aus bereits bestehender Liste von Labelgruppen; rechte Seite – Liste der Unterlabels zur Labelgruppe Solidarity

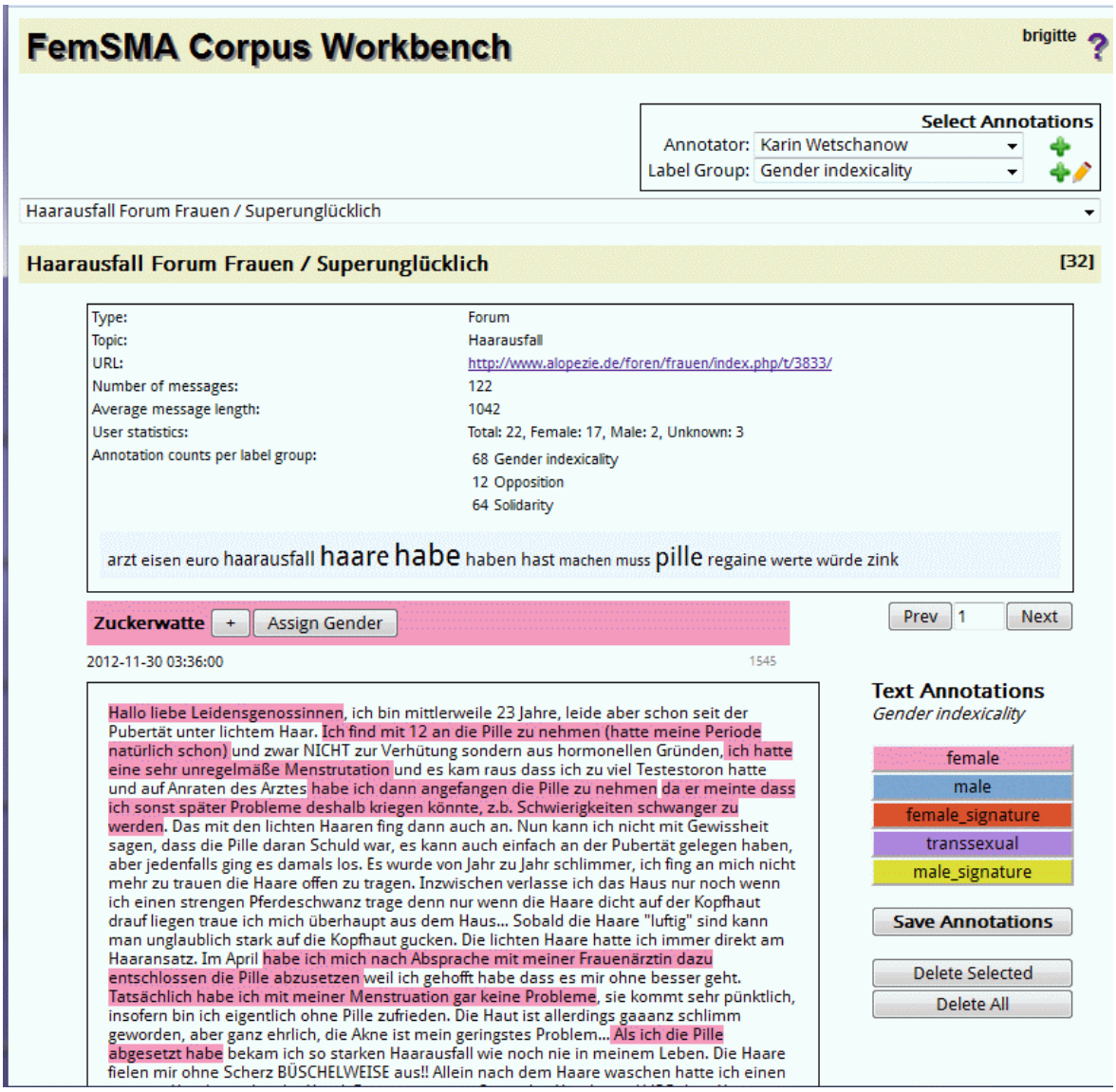

Abbildung 5: CWB: Annotierinterface

4 Automatische Verarbeitung von Texten mittels computerlinguistischer Analysetools

Neben der manuellen Annotierung von Texten beinhaltet die CWB auch eine Reihe von computerlinguistischen Tools zur automatischen Analyse bzw. Annotierung. Die Annotierung erfolgt auf Wortebene im Zuge der Tokenisierung der Texte. In der Computerlinguistik wird jener Prozess Tokenisierung genannt, der den Text in bedeutungstragende Einheiten zerlegt. Das können Wörter, Phrasen, Symbole oder andere bedeutungsvolle Einheiten sein wie z. B. Emoticons oder Twitter-Hashtags (siehe auch Manning/Schütze 1999: 124–130, 4.2.2 „Tokenization: What is a word?“). In der vorliegenden Version der CWB wird jedes Token einer Tokenkategorie zugeordnet. Beispiele für Tokenkategorien sind Ellipse, Emotikon,

Satzzeichen, Wort, URL, Twitter Hashtag udgl. Tokens der Kategorie Wort werden des Weiteren mit einer Reihe von Wortmerkmalen versehen. Dazu gehören:

1. Allgemeine Wortmerkmale, wie Abkürzung, Kapitalisierung, Reduplikation von Buchstaben z. B. *sooo*, Interjektion, Schimpfwort.
2. Part-of-Speech Tags, d. h. morphosyntaktische Kategorien wie Nomen, Verb, Adjektiv udgl. Konkret wird das Stuttgart-Thübingen Tagset (STTS, www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html) verwendet.
3. Wörter, die positive oder negative Emotion bzw. Bewertung ausdrücken, wie z. B. glücklich, Glück, Trauer, traurig, gut, schön, böse, schlecht, miserabel, usw. Zur Identifizierung solcher Wörter werden verschiedene Wortlisten bzw. Lexika verwendet, wie zum Beispiel SentiStrength_DE (www.ofai.at/research/interact/resources/SentiStrength_DE/download_form.html).

Abbildung 6 zeigt ein Beispiel für automatische Annotierung. Rechts sind die verschiedenen Tokenlabels zu sehen, die im angezeigten Dokument als Ergebnis der Tokenisierung vorhanden sind. Wird mit der Maus auf ein Label geklickt, so werden alle Wörter, die mit diesem Label annotiert sind mit der entsprechenden Labelfarbe unterlegt. Im vorliegenden Beispiel sind alle Wörter, die automatisch als attributives Adjektiv (ADJA) erkannt wurden, gelb unterlegt.

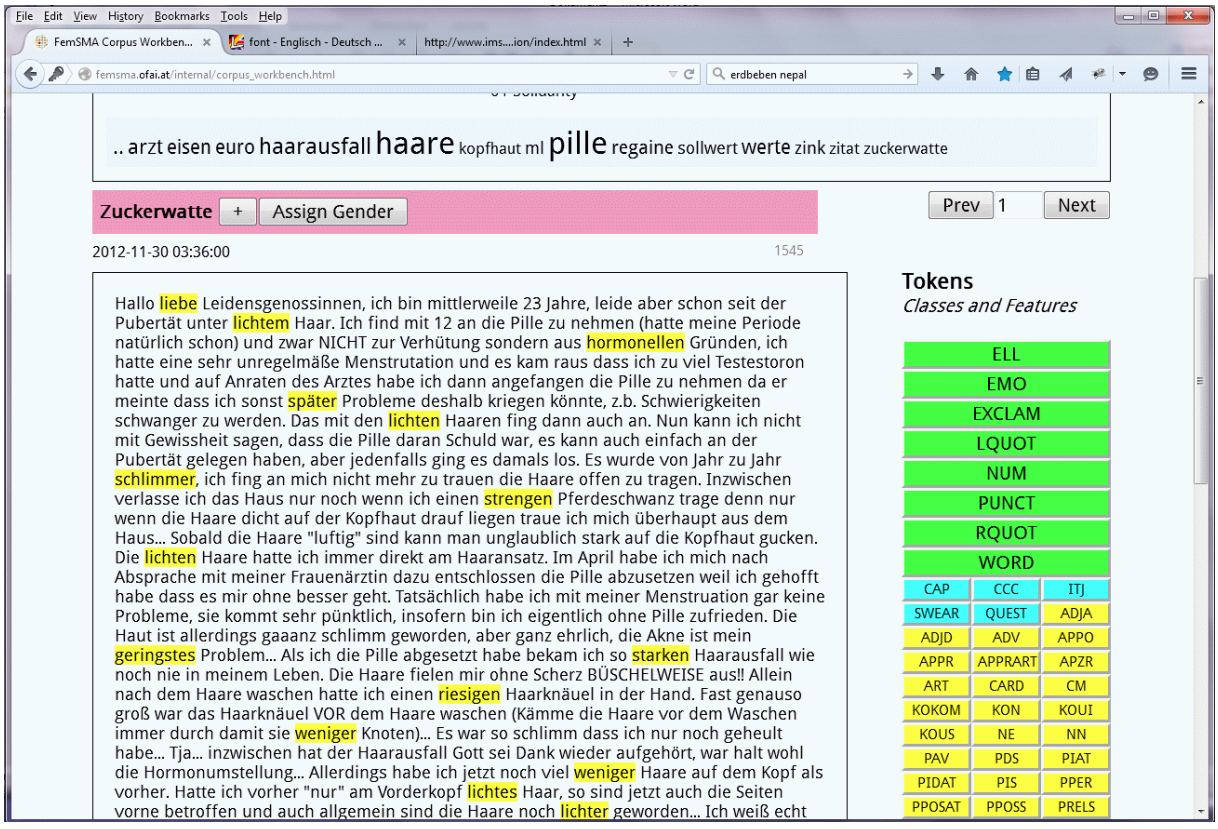

Abbildung 6: Automatische Annotierung

5 Automatische Suche in Textdokumenten

Die Suchkomponente operiert ebenso wie die Tokenisierungskomponente auf den Textdokumenten. Es gibt eine Reihe von Möglichkeiten die Suche einzuschränken, diese

können miteinander verknüpft werden. Die in FemSMA implementierten Einschränkungen für die Suche sind (siehe Abbildung 7):

„Restrict User" – es wird in den Dokumenten einer bestimmten User_in/Autor_in gesucht.

„Restrict Resource" – es wird in ausgewählten Dokumenten und nicht im gesamten Korpus gesucht.

„Restrict Markup" – es wird nach bestimmten Labels gesucht.

„Restrict Content" – es wird nach Vollformen oder regulären Ausdrücken gesucht.

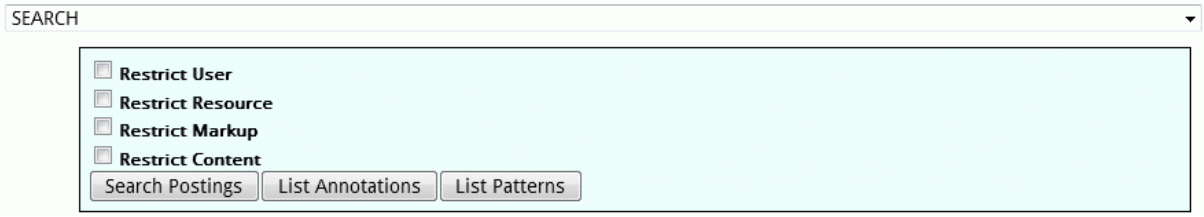

Abbildung 7: Suche: Einschränkungskriterien für die Suche und die Ausgabe der Suchergebnisse

Je nach Anfrage („Search Postings", „List Annotations", „List Patterns") werden die folgenden Ressourcen ausgegeben: Erstens, die auf die Suchanfrage passenden Dokumente (Search Postings). Abbildung 8 illustriert das Suchergebnis zur generellen Anfrage „Gib mir alle im aktuellen Korpus vorhandenen Dokumente aus". D. h. es werden keine Einschränkungen der Suche vorgenommen und der Search-Postings-Button wird gedrückt. Im Ausgabeinterface ist zu sehen, dass das gesamte Korpus 46.726 Dokumente enthält („Postings found"). Diese sind von 1 bis 46.726 durchnummeriert und können einzeln mittels Eingabe der Dokumentnummer oder der Prev- und Next-Buttons (über dem Text-Annotations-Feld in der Abbildung) eingesehen werden. Zusätzlich kann im Select-Annotations-Feld eine Labelgruppe eingegeben werden, die für die manuelle Inspektion der Korpusdaten von Interesse ist. Das vorliegende Beispiel zeigt die Sicht auf das Dokument Nummer 5 und dessen bereits vorhandene Annotierungen innerhalb der Labelgruppe Solidarität ('Solidarity'). Es ist zu sehen, dass die Textstelle „vielen dank" mit dem Label *thanks* annotiert wurde.

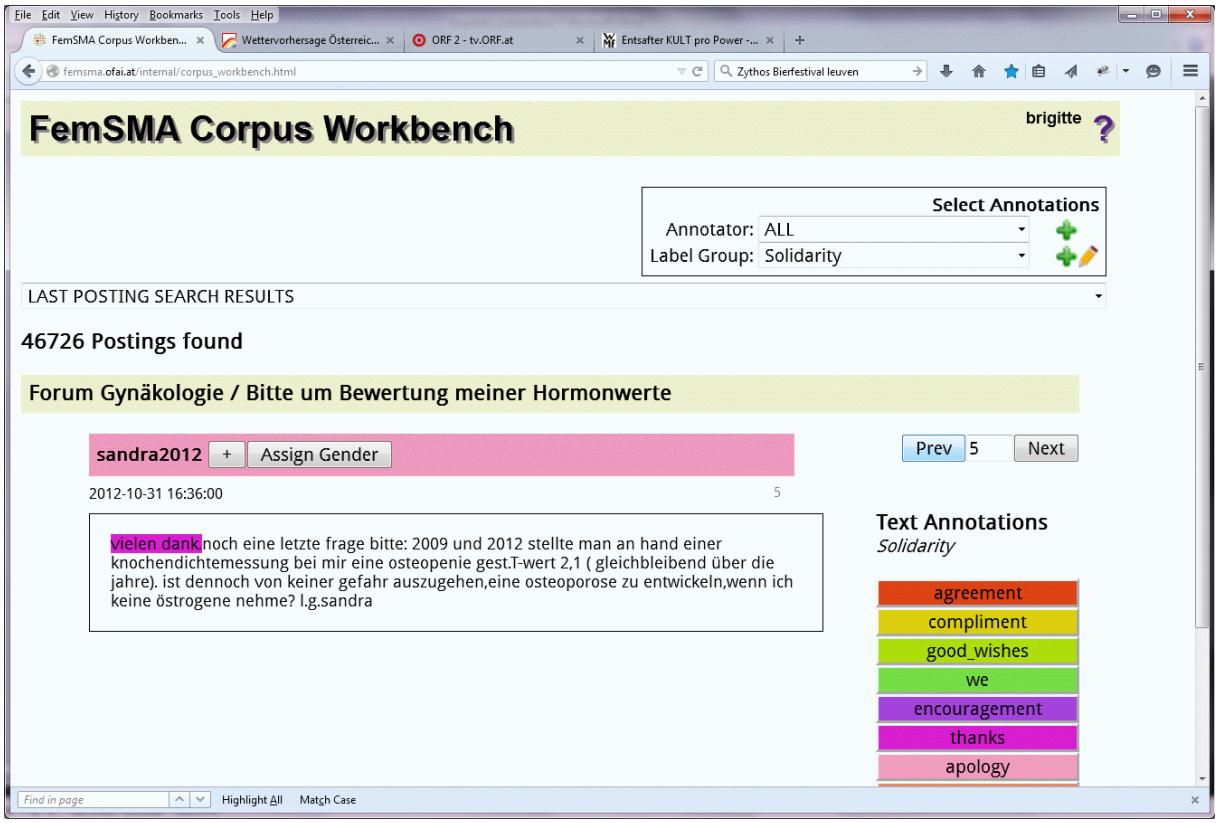

Abbildung 8: Search Postings

Außerdem kann zweitens eine alphabetisch und nach Labelgruppen und Labels geordnete Liste von in den Dokumenten gefundenen manuell annotierten Phrasen (“List Annotations”) ausgegeben werden. Ein Ausschnitt der Ergebnisliste von manuell annotierten Belegstellen für Sarkasmus (Label *sarcasm* aus der Labelgruppe Opposition) ist in Abbildung 9 zu sehen. Jede Belegstelle wird als Hyperlink dargestellt und ist mit dem/den jeweiligen Dokument(en)/Text(en), in dem/denen sie auftritt, verknüpft. Betrachten wir zum Beispiel die Belegstelle „Bravo Grüne … dann wird bald jeder bei der Feuerwehr sein“ (Abbildung 10), sehen wir, dass sie in einem Forumposting der Online-Zeitung *derstandard.at* zum Thema Wehrpflicht vorkommt, dass der Text von einem Autor mit Nicknamen Cotton verfasst wurde, und dass der User Cotton als männlich identifiziert wurde (blau unterlegter Username). Im besten Fall können all diese Informationen bereits beim *Scraping*, d. h. der automatischen Verarbeitung der das Posting enthaltenden Webseite, gefunden werden. Ist das nicht möglich, wird die Information beim Upload eines Dokuments in die CWB manuell eingegeben.

In Abbildung 10 sehen wir des Weiteren, dass im vorliegenden Text nicht nur eine sarkastische Äußerung vorkommt, sondern auch eine Beschimpfung, siehe die manuell als „insult“ markierte Textstelle „Solche Anrechnungen sind Schwachsinn“. Je nachdem welche „Label Group“ im Select-Annotations-Bereich eingestellt wird, werden, wenn vorhanden, die mit den Labels der ausgewählten Labelgruppe annotierten Textstellen farblich unterlegt. Diese Funktionalität unterstützt die manuelle Inspektion der Texte hinsichtlich gemeinsam in einem Text auftretender Merkmale.

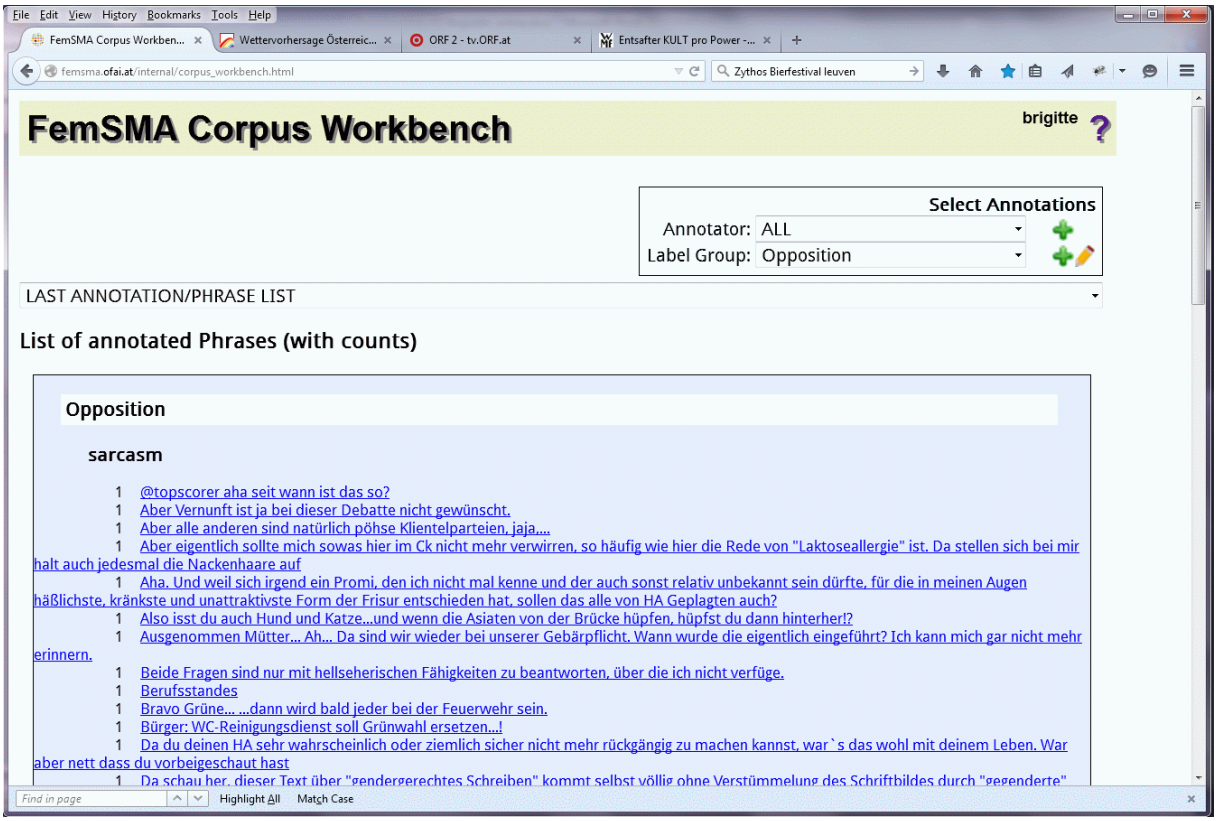

Abbildung 9: List Annotations: Belegstellen, die manuell mit dem Label *sarcasm* annotiert wurden

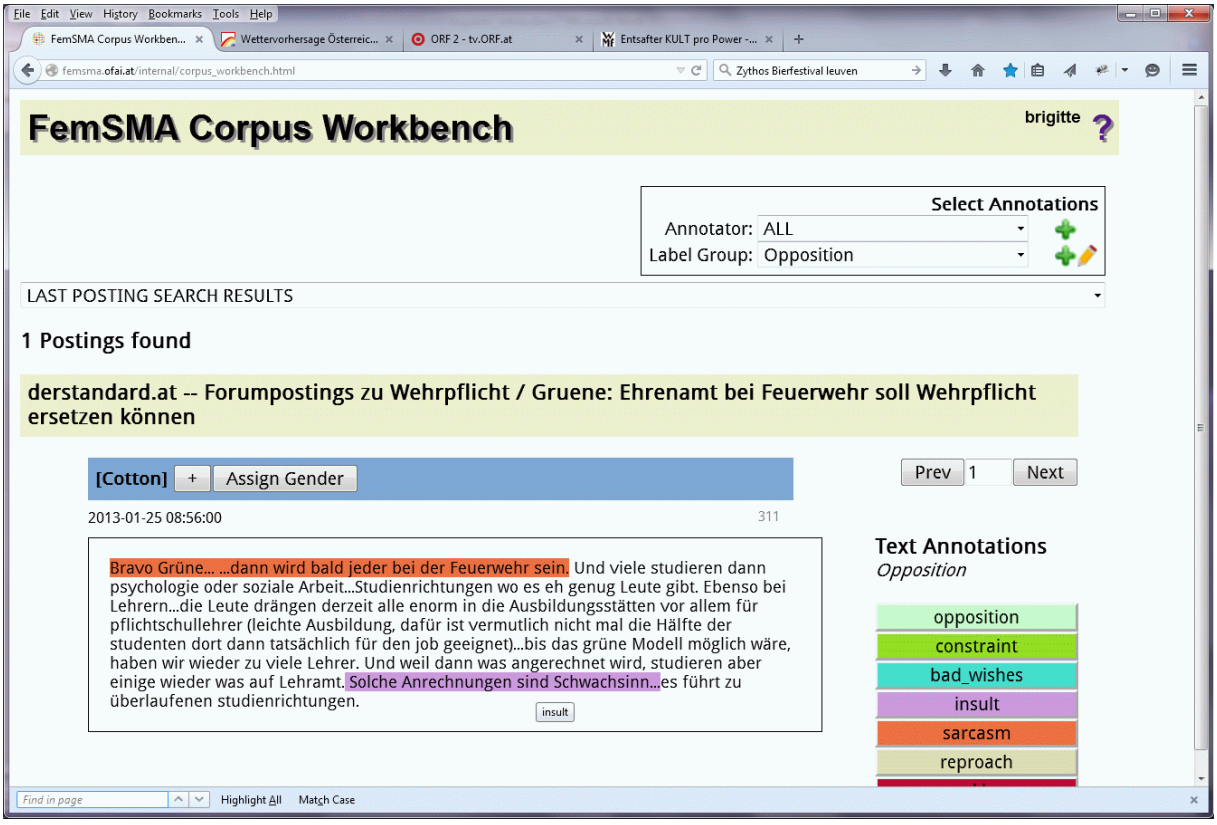

Abbildung 10: Belegstelle „Bravo Grüne ...“

Schließlich kann drittens eine Liste von Textstellen, die aufgrund einer Volltextsuche oder eines regulären Ausdrucks gefunden wurden, ausgegeben werden (List Patterns). Abbildung 11, zum Beispiel, zeigt einen Ausschnitt aus der Ergebnisliste zur Suche nach dem Wort

(Suchstring) „Haarausfall“. Aufgelistet wird der Suchstring selbst und die 12 Zeichen rechts und links vom Suchstring, in welcher Ressource (Chefkoch.de/Laktoseintoleranz, Forum Haarausfall allgemein Männer Optik QS usw.) und bei welcher Autor_in (Dakota, Gauloises, Morrissey, …) der Suchstring vorkommt. Zusätzlich kann für jede Belegstelle der vollständige Text, in dem der Suchstring vorkommt, angezeigt werden, siehe Abbildung 12.

Resultierend aus der List-Patterns-Funktionalität sind in der Phrasenliste die Suchstrings je nach Autor_innengeschlecht unterschiedlich farblich unterlegt. So kann ein erster Eindruck darüber gewonnen werden, ob, bezogen auf das vorliegende Textkorpus, eine Belegstelle eher in Texten von Autorinnen, Autoren oder geschlechterausgewogen vorkommt. Als Nebeneffekt wird die List-Patterns-Funktionalität in FemSMA auch für das Nachannotieren bzw. Korrigieren von Geschlechterzuordnungen verwendet. Siehe den Assign-Gender-Button rechts oben in Abbildung 11 und Abbildung 12.

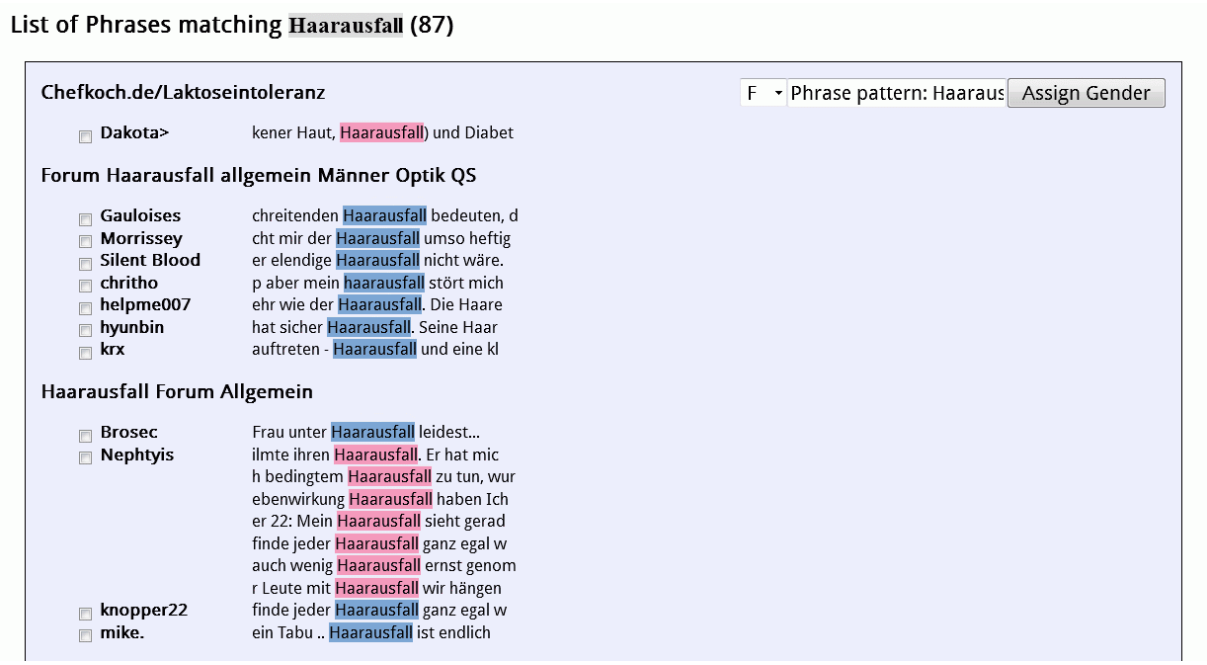

Abbildung 11: List Patterns: Ausgabeliste zum Suchstring „Haarausfall“

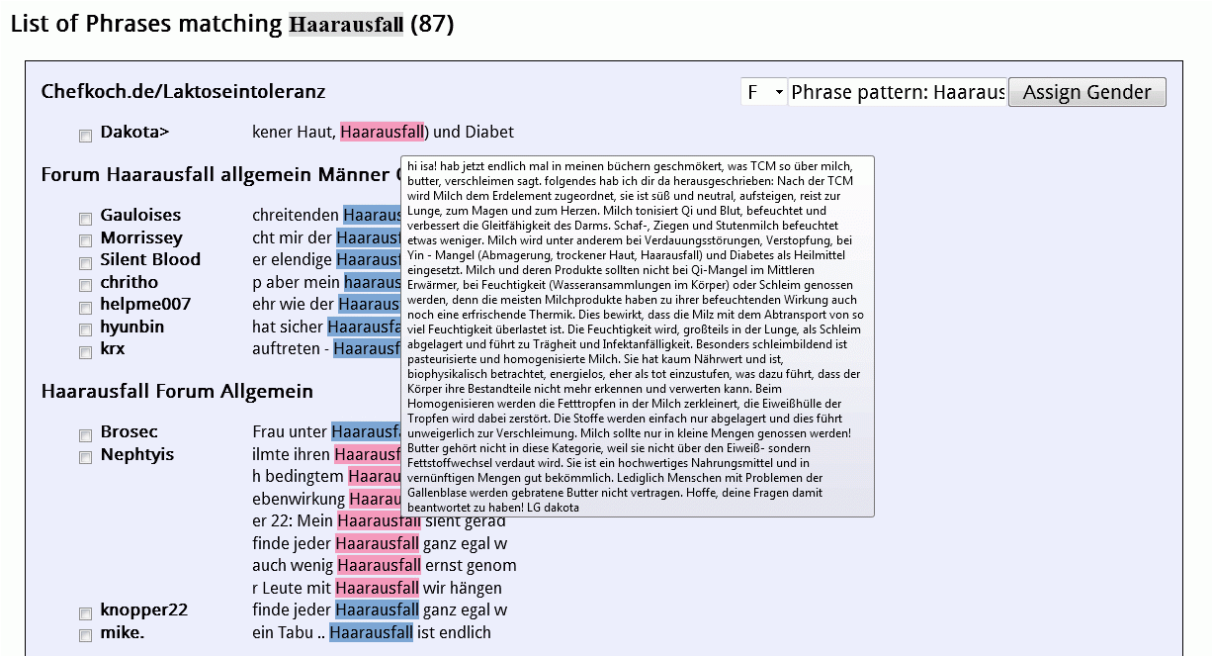

Abbildung 12: List Patterns: gesamter Text, in dem der Matchstring "Haarausfall" vorkommt

6 Verwendung der CWB zur Analyse und Annotierung von Zitationspraxen in funktionalen Abschnitten deutschsprachiger Fachaufsatzeinleitungen

Im Folgenden wird diskutiert, wie die CWB für die Annotierung und Untersuchung von Zitationspraxen in funktionalen Abschnitten deutschsprachiger Fachaufsatzeinleitungen herangezogen werden kann und welche Adaptierungen erforderlich sind. Zu diesem Zweck betrachten wir die Eingangsseite der CWB (Abbildung 13), sowie das Annotierinterface (Abbildung 14), und diskutieren die Abbildung des in Wetschanow (in diesem Heft) aufgestellten Analyserasters auf Labelgruppen und Labels, sowie welche Tools zur automatischen Verarbeitung speziell zur Unterstützung der Annotierung von Zitationspraxen eingesetzt werden können.

In einem ersten Schritt wurde eine Kopie der FemSMA CWB hergestellt und die Datenbank mit den in Wetschanow verwendeten Ressourcen befüllt. Während bei der Analyse von Social Media Postings in FemSMA Webseiten von Foren etc. als übergeordnete Ressourcen definiert sind, denen die jeweiligen Textdokumente zugeordnet sind, kann bei der Analyse und Annotierung von wissenschaftlichen Texten der Artikel selbst als Ressource betrachtet werden, unter der die einzelnen Textabschnitte (Kapitel) repräsentiert sind. Sind in FemSMA die einzelnen Postings die zu annotierenden Texteinheiten, so sind es bei der textlinguistischen Untersuchung von wissenschaftlichen Artikeln die einzelnen Kapitel, wobei in Wetschanow konkret nur die Einleitungen untersucht wurden. Entsprechend wurden nur die Einleitungen als zu annotierende Textdokumente in die Datenbank der CWB-Kopie aufgenommen. Analog zu Abbildung 2 zeigt Abbildung 13 die Eingangsseite der CWB mit der Liste der in der CWB geladenen Artikel, darunter auch jene, die in Wetschanow untersucht wurden. Über diese Auswahlliste wird analog zu Abbildung 5 auf die mit der Ressource verbundenen Texte zugegriffen. Im konkreten Fall gibt es pro Ressource nur einen Text, das Einleitungskapitel. Während die Modellierung der User-Statistik (= Autor_innen-Statistik) an die Gegebenheiten bei wissenschaftlichen Artikeln, d. h. typischerweise mehrere

Autor_innen pro Artikel angepasst werden muss, konnte der Rest der CWB-Funktionalität für manuelle Annotierung gleich belassen werden. Das gilt sowohl für die Metainformation, als auch für die Annotierung und Präsentation der Texte. Type = article, Topic = science, "Number of messages" pro Ressource ist 1, weil nur die jeweiligen Einleitungskapitel Untersuchungsgegenstand in Wetschanow sind. Ebenso bleibt die Funktionalität für „Annotations count per label group" und für die Erzeugung der Termliste unverändert, siehe Abbildung 14.

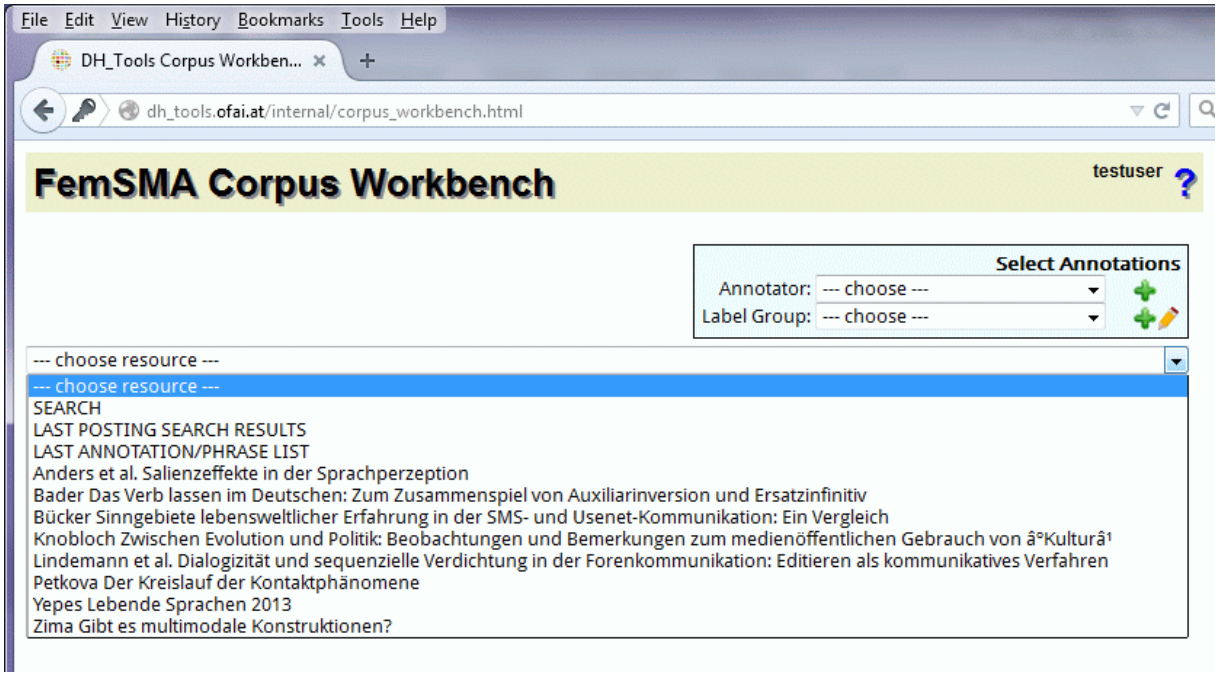

Abbildung 13: Eingangsseite CWB: Liste wissenschaftlicher Artikel

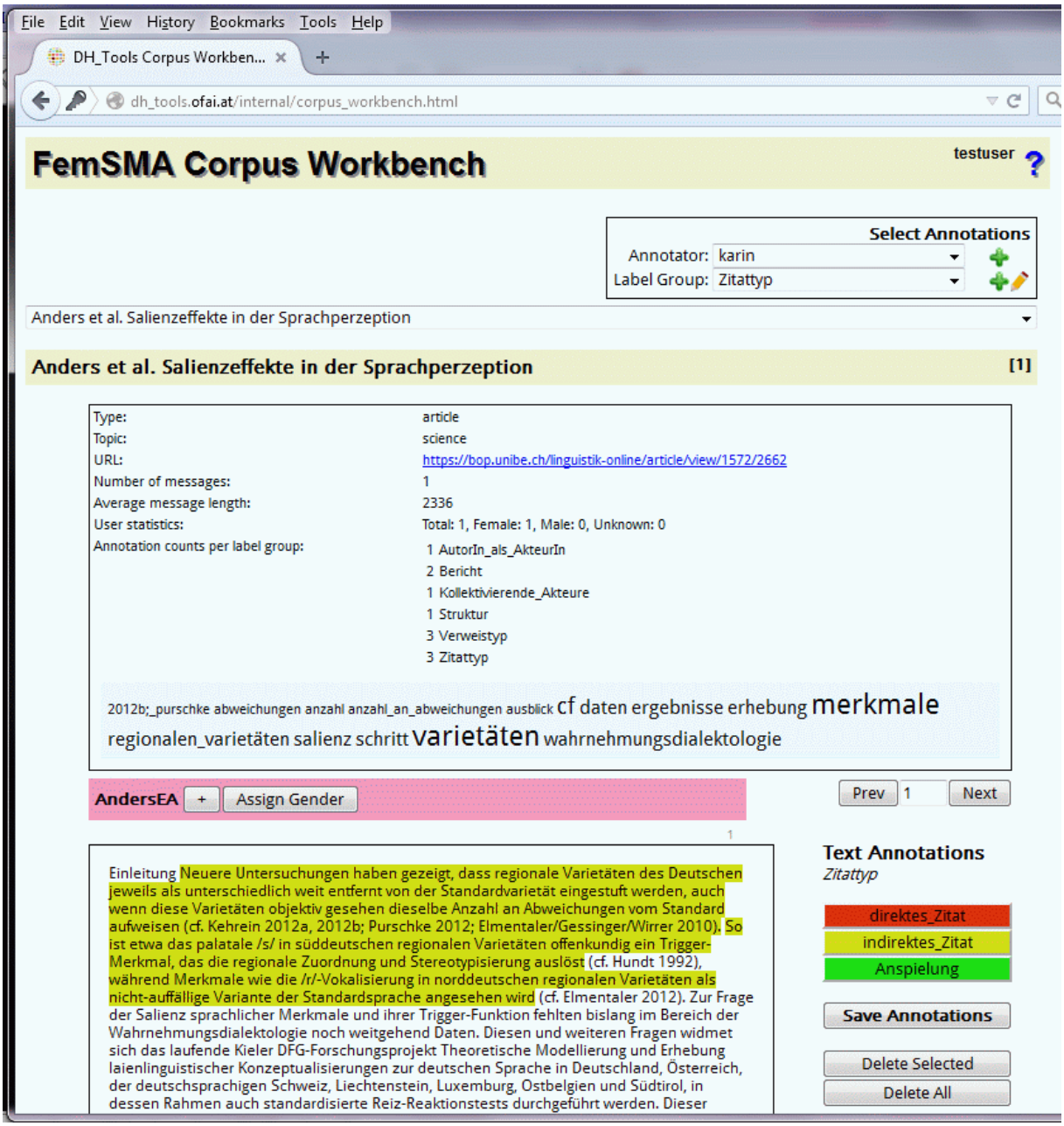

Abbildung 14: CWB: Annotierinterface, wissenschaftlicher Artikel, Einleitung

6.1 Definition von Labelgruppen und Labels

Wie in Abschnitt 0 beschrieben, erlaubt die in der CWB implementierte Annotierfunktionalität die Definition von Labelgruppen und zugehörigen Labels, wobei letztere annotiert werden. Dem gegenüber steht das in Wetschanow (in diesem Heft) vorgeschlagene Analyseraster, das teilweise drei- und vier-stufig ist. Im Folgenden wird ein Vorschlag zur Übertragung des Wetschanow'schen Analyserasters in das in der CWB verfügbare Annotierformat gemacht. Ziel ist, die Information aus der ursprünglichen, theoretischen Analyse zu erhalten, ohne die in der CWB vorhandene Annotierfunktionalität umprogrammieren zu müssen. Eine gängige Vorgehensweise in der Computerlinguistik ist, die Hierarchie der Labelbeziehungen in den Labelnamen abzubilden. Siehe dazu ein Beispiel aus der Kodierung der Verben im STTS: Die folgenden drei Labels VVFIN, VAFIN, VMFIN beschreiben finite Verben, wobei der erste Buchstabe V für die Oberkategorie Verb steht, der

zweite für die spezifische Verbkategorie Vollverb V, Auxiliar A, bzw. Modalverb M und FIN für finit. Auf ähnliche Weise wurde mit dem in Wetschanow aufgestellten Labelraster verfahren, welches aus Gründen der Übersichtlichkeit im Folgenden (

Funktionale Abschnitte	• Territorium etablieren: o Zentralität behaupten ▪ Aktuaität der Forschung ▪ Aktualität in der Gesellschaft o Thema: Verallgemeinerungen o bisherige Forschung besprechen o bisherige Forschung besprechen und bewerten • Eine Nische etablieren: o Widersprechen o Lücke aufzeigen ▪ Datenmangel behaupten ▪ Fehlende Perspektive behaupten o Frage aufbringen o Tradition weiterführen • Die Nische besetzen: o Zwecke aufzeigen o aktuelle Forschung ankündigen o wichtigste Resultate ankündigen o RA-Struktur zeigen
o intertextuelle Verweistypen	• Direkte Zitate/ Anspielungen • Indirekte Zitate
o Strukturtypen	• syntaktisch-semantisch nicht-integriert • syntaktisch-semantisch integriert o Subjekt o Nicht-Subjekt/passiv o Konstituent auf Phrasenebene
o Berichtstyp	• berichtend • nicht-berichtend
o Akteur_innentypen	• Individuelle Akteur_innen • Kollektivierte Akteur_innen o Texte o Personen o Theorien oder Ideologien o Disziplinen o soziale Felder o Kulturen bzw. Kulturkreise
o Funktion des Quellverweises	• Integrierte Quellverweise o Quellverweis als Teil eines Werknamens: o Nichtverweisende „Quellangabe“ • Nicht-integrierte o Beleg o Beispielhafte Nennung o Quellangabe

Tabelle 1) noch einmal zusammengefasst ist. Tabelle 2, hingegen, gibt Beispiele für die Übersetzung des von Wetschanow vorgeschlagenen Analyserasters in Labelgruppen und Labels in der CWB. Kriterien für die Definition der Labelgruppen und Labels waren: (1) Weitgehende Nähe zu den Benennungen im ursprünglichen Analyseraster bei gleichzeitiger Kompaktifizierung der Labels, damit sie (a) sprechend und (b) kurz genug sind, um in das Layout des Webinterfaces zu passen. Längere Beschreibungen zu den Labelnamen können im Description-Feld (siehe den Select-Annotations-Bereich im CWB Webinterface) angegeben werden. (2) Annotierpraktische Gründe, es soll eine Balance gefunden werden zwischen der Notwendigkeit des Umschaltens von Labelgruppe zu Labelgruppe und der Länge der unter einer Labelgruppe verfügbaren Labelnamen.

○ Funktionale Abschnitte	• Territorium etablieren: ○ Zentralität behaupten ▪ Aktuaität der Forschung ▪ Aktualität in der Gesellschaft ○ Thema: Verallgemeinerungen ○ bisherige Forschung besprechen ○ bisherige Forschung besprechen und bewerten • Eine Nische etablieren: ○ Widersprechen ○ Lücke aufzeigen ▪ Datenmangel behaupten ▪ Fehlende Perspektive behaupten ○ Frage aufbringen ○ Tradition weiterführen • Die Nische besetzen: ○ Zwecke aufzeigen ○ aktuelle Forschung ankündigen ○ wichtigste Resultate ankündigen ○ RA-Struktur zeigen
○ intertextuelle Verweistypen	• Direkte Zitate/ Anspielungen • Indirekte Zitate
○ Strukturtypen	• syntaktisch-semantisch nicht-integriert • syntaktisch-semantisch integriert ○ Subjekt ○ Nicht-Subjekt/passiv ○ Konstituent auf Phrasenebene
○ Berichtstyp	• berichtend • nicht-berichtend
○ Akteur_innentypen	• Individuelle Akteur_innen • Kollektivierte Akteur_innen ○ Texte ○ Personen ○ Theorien oder Ideologien ○ Disziplinen ○ soziale Felder ○ Kulturen bzw. Kulturkreise
○ Funktion des Quellverweises	• Integrierte Quellverweise ○ Quellverweis als Teil eines Werknamens: ○ Nichtverweisende „Quellangabe“ • Nicht-integrierte ○ Beleg ○ Beispielhafte Nennung ○ Quellangabe

Tabelle 1: Analyseraster Gliederung Wetschanow (in diesem Heft)

Labelgruppe Label	**Labelgruppe** Label
FA_Territorium_etablieren Zentralität_Aktualität Zentralität_Aktualität_Forschung Zentralität_Aktualität_Gesellschaft Thema_Verallgemeinerung bisherige_Forschung_besprechen bisherige_Forschung_besprechen_bewerten	**Akteur_innentyp** individuell Texte Personen Theorien/Ideologien Disziplinen soziale_Felder Kulturen/Kulturkreise
FA_Nische_etablieren Widersprechen Lücke_Datenmangel	**Quellverweis_integriert** Teil_von_Werknamen nichtverweisende_Quellangabe

Lücke_Perspektive Frage Tradition	
FA_Nische_besetzen Zweck aktuelle_Forschung Resultate RA_Struktur	**Quellverweis_nicht_integriert** Beleg Beispielhafte_Nennung Quellangabe
Syn-Sem-Struktutypen nicht_integriert Subjekt Nicht_Subjekt(passiv) Konstituent(syn)	

Tabelle 2: Übertragung des Analyserasters in Wetschanow (in diesem Heft) in entsprechende für die Verarbeitung in der CWB taugliche Labelgruppen und Labels; nur Kategorien mit ursprünglich mehr als zwei Beschreibungsebenen sind in die Beispielliste aufgenommen

6.2 Automatische Verarbeitung zur Unterstützung der manuellen Analyse und Annotierung

Automatische Unterstützung für die manuelle Analyse und Annotierung von Zitationspraxen kann auf drei Arten erfolgen: (1) Quellverweise identifizieren, (2) Kontexte zu den Quellverweisen bestimmen, (3) Kontexte analysieren. Beispiele für Quellverweise im laufenden Text sind z. B.:

(cf. Hundt 1992)
(cf. Kehrein 2012a, 2012b; Purschke 2012; Elmentaler/Gessinger/Wirrer 2010)
(vgl. Schegloff/Sacks 1973; Schegloff 2007)
(vgl. Schegloff/Sacks 1973: 295-296)
(Münker 2009: 57)
(Bader/Fritz 2011: 59)

Diese bilden verschiedene Muster und können mittels regulärer Ausdrücke identifiziert werden. Einerseits können sie unter Verwendung der in der CWB vorhandenen Suchfunktionalität „Restrict Content“ systematisch im Korpus gesucht und in weiterer Folge manuell annotiert werden. Andererseits können die für die Suche verwendeten regulären Ausdrücke in die Tokenisierungskomponente der CWB eingebaut werden und so die Quellverweise in der Tokenansicht im CWB Webinterface sichtbar gemacht werden.

Ist der Quellverweis identifiziert, können in einem weiteren automatischen Verarbeitungsschritt die zu den einzelnen Quellverweisen gehörigen Kontexte identifiziert werden. Typischerweise sind es Sätze oder Satzteile, in denen der Quellverweis vorkommt. Siehe z. B. einen Quellverweis im Satzkontext bei Zima (2014),

> Dabei gibt er allerdings innerhalb der Kognitiven Grammatik folgerichtig zu bedenken, dass nicht jede ko-verbal gebrauchte Geste automatisch als Teil der sprachlichen Einheit angenommen werden kann.
>
> (Langacker 2008: 250–251)

sowie ein Beispiel für Quellverweise, die sich auf Satzteile beziehen aus Petkova (2012).

> So fasst Myers-Scotton das ganze Spektrum der Sprachkontaktphänomene unter dem Begriff code-switching zusammen (Myers-Scotton 1997, 2006), während Muysken dies unter dem Begriff codemixing tut.
>
> (Muysken 2000)

Diese Art von Kontexten kann unter Ausnutzung von Satzzeichen automatisch identifiziert werden. Sind die Quellverweise und ihre Kontexte identifiziert, können die Kontexte weiter analysiert werden, z. B. hinsichtlich des Vorhandenseins bestimmter thematischer Verben oder Verbgruppen, wie Verben des Argumentierens (*argumentieren*, *vorschlagen*, *darauf hinweisen, Bezug nehmen*, usw.), des Denkens (*meinen*, *annehmen*, *empfinden*, usw.), des Zeigens (*zeigen*, *aufzeigen*, *belegen*, *demonstrieren*, usw.), des Findens (*entdecken*, *etablieren*, *beobachten*, usw.), sowie damit verbundener *dass*-Sätze. Des Weiteren können über die syntaktische Analyse Subjekte identifiziert werden. Auch kann, bis zu einem gewissen Grad, die semantische Klasse der Subjekte ermittelt werden, um automatisch Vorschläge für die Annotierung von Akteur_innentypen zu machen. Wie bei allen automatischen Sprachverarbeitungsverfahren ist mit einem gewissen Prozentsatz an Fehlerhaftigkeit der Ergebnisse zu rechnen. Die automatischen Analysen helfen jedoch potentielle Belegstellen schneller aufzufinden und dienen der Unterstützung der Linguist_innen in der qualitativen Arbeit.

7 Zusammenfassung

Im vorliegenden Beitrag wurde die Corpus Workbench CWB vorgestellt. Die CWB ist ein computerlinguistisches Werkzeug zur manuellen und automatischen Annotierung und Analyse von Textdokumenten. Zusammenfassend zeichnet sich die CWB durch folgende Merkmale aus: Die CWB unterstützt bei der Verwaltung, der Annotierung und dem Durchsuchen von Textdokumenten. Es werden sowohl die Dokumente als auch verschiedene mit den Dokumenten verbundene Metadaten wie Art, Thema und Name der Ressource, sowie deren manuelle Annotierungen in einer zentralen Datenbank gehalten. Der Zugang zu den Daten erfolgt über ein Webinterface. So kann über ein Scroll-Down-Menü auf alle in der CWB vorhandenen Dokumente zugegriffen werden. Beim einzelnen Dokument können die manuellen sowie die automatischen Annotierungen angezeigt werden. Es können neue Annotierungen hinzugefügt werden, wobei die Dokumente gesondert von mehreren Personen annotiert werden können. Die manuellen Annotierungen erfolgen über Markierung der betreffenden Textstelle mit dem Mauscursor. Die Labels für die Annotierung können frei definiert werden. Zusätzlich zu den manuellen Annotierungen, werden die Texte automatisch analysiert und mit verschiedenen Merkmalen auf Wortebene versehen, wie z. B.: morphosyntaktische Kategorien (*Part-of-Speech*), allgemeine Wortmerkmale, wie z. B. ob es sich um ein Wort mit Reduplikation (*sooo*), eine Abkürzung (*z. B.*), um Kapitalisierung (*DIE*) udgl. handelt, ob ein Wort ein Schimpfwort ist, oder einen positiven oder negativen emotionalen bzw. wertenden Gehalt hat (*schön*, *hässlich*, *böse*, *Liebe*, usw.). Zusätzlich werden für jedes Dokument die charakteristischen Terme berechnet und in Form einer Termcloud in der Dokumentansicht ausgegeben. Das erlaubt einen ersten, groben Blick auf den Inhalt des Dokuments. Darüber hinaus können die Textdokumente nach unterschiedlichen Kriterien durchsucht werden: nach bestimmten handannotierten Labels, mittels Volltextsuche

und über reguläre Ausdrücke. Gesucht werden kann im gesamten Textkorpus, oder auf ausgewählten Dokumenten.

Die CWB ist insbesondere für die enge Zusammenarbeit von Linguist_innen und Computerlinguist_innen ausgelegt. Erstere unterstützt sie bei der qualitativen Arbeit mittels flexibler Annotier- und Suchfunktionalitäten. Durch die Einbindung von Computerlinguist_innen können die manuell annotierten, exemplarischen Belegstellen im Textkorpus hinsichtlich ihrer Verarbeitbarkeit mittels computerlinguistischer Tools analysiert und die CWB um entsprechende Funktionalitäten erweitert werden. Die gezielte automatische Verarbeitung unterstützt einerseits die Linguist_innen bei der qualitativen Arbeit. Andererseits liefert die automatische Identifikation und Annotierung potentieller Belegstellen Input für: (i) quantitative Analysen des Auftretens bestimmter Merkmale und Merkmalskombinationen im gesamten Textkorpus oder in ausgewählten Subkorpora; (ii) maschinelles Lernen, wie z. B. das Erlernen von Modellen für die Klassifikation von Textabschnitten in Belegstelle/Nicht-Belegstelle, bzw. das Clustering ähnlicher Belegstellen.

Die CWB ist *Work-in-Progress*. Aufgrund ihres modularen Aufbaus kann sie gezielt erweitert werden, je nach Anforderung der konkret zu analysierenden Merkmale. Ein Beispiel für die Vorgehensweise bei der Anpassung der CWB an einen neuen Untersuchungsgegenstand wurde im vorliegenden Beitrag mit der Analyse von Zitationspraxen in wissenschaftlichen Texten gegeben. Sofern die in der CWB vorhandene Funktionalität für ein bestimmtes Analysevorhaben nicht ausreicht, müssen Zusatzfunktionalitäten in die CWB eingebaut werden. Dazu bedarf es der Kenntnis verschiedener Programmier- und Repräsentationssprachen wie Perl, JavaScript, jQuery, HTML. Als Webserver für die CWB wird Apache 2 verwendet. Gegenwärtig werden 3 Instanzen der CWB am OFAI betrieben: die im Beitrag beschriebene FemSMA-Instanz und eine erste Adaption für die Untersuchung von Zitationspraxen, sowie eine Version der FemSMA-Instanz, die in der Lehre eingesetzt wird und Student_innen als Übungsplattform dient. Bevor die CWB frei zugänglich gemacht wird, sollen noch 2 bis 3 weitere Testprojekte unter Verwendung und Adaption der CWB am OFAI durchgeführt werden. Als längerfristiges Ziel soll eine Kernversion der CWB als virtuelle Maschine mittels VirtualBox (www.virtualbox.org) zur Verfügung gestellt werden.

Literaturverzeichnis

Baroni, Marco/Bernardini, Silvia (2004): "BootCaT: Bootstrapping corpora and terms from the web". In: Lino, Maria Teresa et al. (eds.): *Proceedings of LREC* 2004. Lisbon, ELDA: 1313–1316.

Garrett, Jesse James (2005): "Ajax: A New Approach to Web Applications". http://adaptivepath.org/ideas/ajax-new-approach-web-applications/ [03.11.2015].

Manning, Christopher/Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Kotthoff, Helga (2012): „'Indexing gender' unter weiblichen Jugendlichen in privaten Telefongesprächen". In: Günthner, Susanne et al. (eds.): *Genderlinguistik. Sprachliche Konstruktionen von Geschlechtsidentitäten*. Berlin, de Gruyter: 251–285.

Ochs, Elinor (1992): "Indexing Gender". In: Duranti, Alessandro/Goodwin, Charles (eds.): *Rethinking Context.* Cambridge, Cambridge University Press: 335–358.

Petkova, Marina (2012): „Der Kreislauf der Kontaktphänomene". *Sociolinguistica* 26/1: 1–17.

Wetschanow, Karin (in diesem Heft): „Zitationspraxen in deutschsprachigen Fachaufsatzeinleitungen“.

Zima, Elisabeth (2014): „Gibt es multimodale Konstruktionen? Eine Studie zu [V(motion) in circles] und [all the way from X PREP Y]“. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 15: 1–48.