Stefanie Dipper

Digitale Korpora in der Lehre — Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik

In diesem Artikel werden verschiedene Szenarien aus der Lehre vorgestellt, in denen Korpora (und andere Sprachressourcen) Einsatz finden. Jeweils zwei Beispiele illustrieren die Nutzung von Korpora in der Theoretischen Linguistik und in der Computerlinguistik. In der Theoretischen Linguistik dienen Korpora als Belegquellen oder Testdaten für die Hypothesen aus der theoretischen Forschung. In der Computerlinguistik werden Korpora für die Anwendungsentwicklung oder für den Ressourcenaufbau eingesetzt.

1 Einführung[1]

In diesem Artikel geht es um den Einsatz digitaler Sprachressourcen in der universitären Lehre. Die hier angesprochenen Sprachressourcen sind unterschiedlicher Art und reichen von "rohen Texten", wie man sie z.B. im Internet findet, über sorgfältig aufbereitete, annotierte Korpora und spezialisierte Korpus-Suchtools bis hin zu automatischen Analysetools, z.B. für die automatische Wortart-Erkennung. Die Lehrveranstaltungen, in denen diese Sprachressourcen zum Einsatz kommen, richten sich zum einen an Studenten der Theoretischen Linguistik, zum andern an Studenten der Computerlinguistik.

Es werden insgesamt vier Erfahrungsberichte aus der Lehre vorgestellt. Sie sind so ausgewählt, dass sie möglichst verschiedenartige Aspekte von Sprachressourcen, ihrer Nutzung und der an sie gestellten Anforderungen illustrieren. Konkret werden zwei Einsatzszenarien aus der Theoretischen Linguistik (Abschnitt 2) und zwei aus der Computerlinguistik (Abschnitt 3) beschrieben. In Abschnitt 4 folgen abschließende Anmerkungen.

2 Korpora in der Theoretischen Linguistik

In den ersten beiden Erfahrungsberichten geht es um (kleinere) Lehreinheiten, wie sie im Rahmen eines Bachelor-Studiengangs für Studenten der Linguistik angeboten werden.

2.1 Korpora als Belegquellen

Im ersten vorgestellten Einsatzszenario dienen Korpora vorwiegend als Belegquellen. In den Korpora wird nach sprachlichen Belegen gesucht, die Instanzen eines im Kurs behandelten linguistischen Phänomens sind.

[1] Alle angegebenen URLs wurden am 8.3.2011 abgerufen.

Als ein Beispiel für ein solches Phänomen soll die *Vorfeldbesetzung* im Deutschen dienen. Dabei geht es um die Frage, was für Faktoren bestimmen, welche Konstituente in einem deutschen Hauptsatz die Position vor dem finiten Verb, das sogenannte *Vorfeld*, einnimmt. Im Gegensatz zum Englischen kann hier eine beliebige maximale Phrase stehen, z.B. eine Subjekt- oder Objekt-NP, eine PP oder auch eine Adverbialphrase, wie in Beispiel (1)[2] (die Vorfeld-Konstituente ist unterstrichen).

(1) Manchmal mußten erst Mahnschreiben in Sahlins Briefkasten landen, bevor die mit rund 10.000 Mark monatlich nicht unbedingt schlecht versorgte Ministerin sich zur Rückzahlung bequemte.

Die Faktoren, die hier eine Rolle spielen, sind recht gut untersucht: Neben der grammatischen Funktion der Vorfeld-Konstituente (es sind mehrheitlich Subjekte) spielt auch (Nicht-)Vorerwähntheit eine wichtige Rolle (z.B. Filippova and Strube (2007); Speyer (2007)). So wird die Vorfeldposition u.a. auch bevorzugt von Ausdrücken eingenommen, die für den Hörer neue Information darstellen, und von sogenannten *scene-setting elements* (zu denen vermutlich auch Bsp. (1) gerechnet werden kann).

Es liegt nun nahe, die Studenten — nach einer Einführung in das Thema und die involvierten Faktoren — gezielt nach Belegen suchen zu lassen, die gegen die genannten "Regeln" verstoßen, also z.B. nach Sätzen, in denen ein vorerwähntes Objekt im Vorfeld steht. Solche Beispiele, die potenziell die linguistischen Hypothesen widerlegen, können dann gezielt untersucht und die Hypothesen gegebenenfalls verfeinert werden.

Für eine effiziente Suche nach Belegen sind zwei Dinge notwendig: (i) entsprechend annotierte Korpora und (ii) Korpus-Suchtools, mit denen diese Annotationen abgefragt werden können.

Korpora Für unsere Beispielsuche von oben — "vorerwähntes Objekt im Vorfeld" — benötigen wir eine *Baumbank*, d.h. ein Korpus, das mit syntaktischer Information (Konstituenten mit ihren Kategorien und Funktionen) annotiert ist. Zusätzlich wollen wir die Position im Vorfeld sowie Vorerwähntheit abfragen können. Tabelle 1 listet die aktuell verfügbaren Baumbanken für das Deutsche[3] und gibt an, ob die Zusatzmerkmale im Korpus mit annotiert sind (die eingeklammerten Werte werden unten näher erläutert).

Korpus-Suchtools Die Korpora liegen typischerweise in mehreren Formaten vor, z.B. im NEGRA-Exportformat (einem Datenbank-ähnlichen Spaltenformat, Brants (1997))

[2] Entnommen aus der TüBa-D/Z: `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`

[3] TüBa-D/Z: "Tübinger Baumbank des Deutschen/Zeitungskorpus", `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`
TüBa-D/S: "Tübinger Baumbank des Deutschen/Spontansprache" (Projekt Verbmobil), `http://www.sfs.uni-tuebingen.de/tuebads.shtml`
NEGRA-Korpus: Projekt "Nebenläufige grammatische Verarbeitung" (SFB 378), `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/`
TIGER-Korpus: "Linguistic Interpretation of a German Corpus", `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/`
PCC: "Potsdam Commentary Corpus", `http://www.ling.uni-potsdam.de/pcc/pcc.html`

Baumbank	Vorfeld	Vorerwähntheit	#Sätze	#Tokens
TüBa-D/Z	✓	✓	55.000	980.000
TüBa-D/S	✓	–	38.000	360.000
NEGRA-Korpus	(✓)	–	20.000	350.000
TIGER-Korpus	(✓)	–	50.000	900.000
PCC	(✓)	✓	2.900	44.000

Tabelle 1: Deutsche Baumbanken und die Merkmale 'Vorfeld' und 'Vorerwähntheit'. Der Eintrag '(✓)' bedeutet, dass das Merkmal nicht explizit annotiert ist, aber "simuliert" werden kann. In den letzten Spalten finden sich Angaben zur Größe der Korpora.

oder in TIGER-XML (Mengel and Lezius, 2000). Um Linguisten einen einfachen Zugang zu den Daten zu ermöglichen, wurden spezialisierte Korpus-Suchtools entwickelt. Diese erlauben es dem Nutzer, Anfragen an das Korpus in einer Anfrage-Sprache zu formulieren, die auf Grundkonzepten der Linguistik beruht. Beispielsweise stellen solche Anfrage-Sprachen Operatoren für die Relationen *Dominanz* und *Präzedenz* bereit. Neben der Funktionalität der *Suche* auf den Daten bieten Korpustools außerdem geeignete *Visualisierungen* der Daten an, z.B. in Form von Phrasenstruktur-Bäumen.

Ein oft verwendetes Suchtool für Baumbanken ist TIGERSearch[4], das unter vielen anderen Formaten auch Daten in TIGER-XML verarbeiten kann. Dieses Tool hat allerdings den (in unserem Fall relevanten) Nachteil, dass es satzweise operiert. D.h. mit TIGERSearch kann nur nach Relationen innerhalb eines Satzes gesucht werden, und das Tool kann nur Bäume für einzelne Sätze darstellen.

Ein Suchtool, das satzübergreifende Abfragen erlaubt, ist ANNIS2[5], das als Multifunktionstool neben baumartigen Strukturen auch Annotationen von Spannen und Zeigerstrukturen erfasst. Die Baumbank TüBa-D/Z, in der das Merkmal der Vorerwähntheit (in Form von Anaphern- und Koreferenz-Relationen) annotiert ist, kann mit ANNIS2 abgefragt werden.

Die in (2) gezeigte Anfrage sucht nach vorerwähnten Objekten im Vorfeld: Zuerst wird allgemein nach Knoten der Kategorie 'VF' (= Vorfeld, Zeile 1) und nach Knoten der Kategorie 'NX' (= Nominalphrasen/NPs, Zeile 2) gesucht. Als zusätzliche Einschränkung wird angegeben, dass die NP als 'OA' (= Objekt im Akkusativ) fungiert und im Vorfeld steht (Zeile 4); die Angaben '#1' und '#2' beziehen sich dabei auf die Knoten, die in Zeile 1 bzw. 2 spezifiziert wurden.[6] Schließlich, als weitere Einschränkung, wird eine

[4] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`

[5] ANNIS: "Annotation of Information Structure, `www.sfb632.uni-potsdam.de/annis/`

[6] Der Ausdruck sucht nach einem Knoten der Kategorie 'VF', der einen Knoten der Kategorie 'NX' dominiert, wobei die Dominanz-Kante ebenfalls ein Label erhält: der 'NX'-Knoten fungiert als 'OA'. Dass die Position 'Vorfeld' als Kategorie-Label annotiert wird, kommt für linguistische Nutzer

nicht näher spezifizierte 'relation' (= Anaphern- oder Koreferenz-Relation) eingeführt (Zeile 3), die auf die NP zutreffen soll (Zeile 5); mit anderen Worten: die NP soll also in irgendeiner Form bereits vorerwähnt sein.

(2)
```
cat="VF" &
cat="NX" &
relation=/.*/ &
#1 >[func="OA"] #2 &
#2 _=_ #3
```

Diese Anfrage findet 139 Treffer im Korpus, darunter den Treffer, der in Abb. 1 gezeigt wird.[7] Die so gefundenen Belege können nun einzeln durchgesehen und genauer untersucht werden.[8]

Für das PCC-Korpus, das ebenfalls eine Koreferenz-Annotation enthält, sieht die entsprechende Anfrage in ANNIS2 wie in (3) aus. Die Form des Suchausdruckes unterscheidet sich deutlich von dem in (2). Dafür gibt es mehrere Gründe: Das PCC verwendet das Syntax-Annotationsschema des TIGER-Korpus[9], in dem keine explizite Markierung des Vorfeldes vorgesehen ist. Stattdessen machen wir die (schwächere) Einschränkung, dass das Objekt die *linke Tochter* des Satzes oder der VP ist (Zeile 5: '@l'). Außerdem gibt es keine einheitliche Auszeichnung für NPs in TIGER; daher lassen wir die Kategorie unterspezifiziert ('node', Zeilen 2–4). Schließlich unterscheidet sich die Form der Koreferenz-Annotation in beiden Korpora (Zeile 6).

(3)
```
cat=/(S|VP)/ &
node &
node &
node &
#1 >@l[func="OA"] #2 &
#3 ->anaphor_antecedent #4 &
#2 _=_ #3
```

sicher zunächst unerwartet. Solche Annotationsentscheidungen müssen den Annotationsguidelines entnommen werden, die als Dokumentation ein unverzichtbarer Bestandteil des Korpus sind (für die TüBa-D/Z: `http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-0911.pdf`).

[7] Die Abfrage in (2) wurde nicht auf der aktuellen Version 6, sondern auf der Vorgängerversion 5 der TüBa-D/Z durchgeführt, die 800.000 Tokens enthält.

[8] Allerdings muss bei solchen Untersuchungen berücksichtigt werden, dass das Konzept von Vorerwähntheit, wie es in der TüBa-D/Z annotiert ist, nicht notwendigerweise mit dem Konzept, wie es in den linguistischen Theorien verstanden wird, identisch ist. Z.B. ist die Vorfeld-NP *das Geld* in Bsp. (i) laut den Annotationen der TüBa-D/Z nicht "vorerwähnt" (da das Geld *als solches* noch nicht thematisiert wurde). Im Vorkontext ist jedoch schon die Rede von "bezahlen", daher stellt der Ausdruck *das Geld* natürlich keine völlig neue (unerwartete) Information im Sinne von Filippova and Strube (2007) oder Speyer (2007) dar.

(i) Seit 1991 bezahlte sie mal neue Schuhe, eine Lederjacke oder gleich die Urlaubsreise für die ganze Familie mit der Staatskarte. Das Geld zahlte sie zurück, aber dummerweise nicht gleich und nicht unaufgefordert.

[9] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation/tiger_scheme-syntax.pdf`

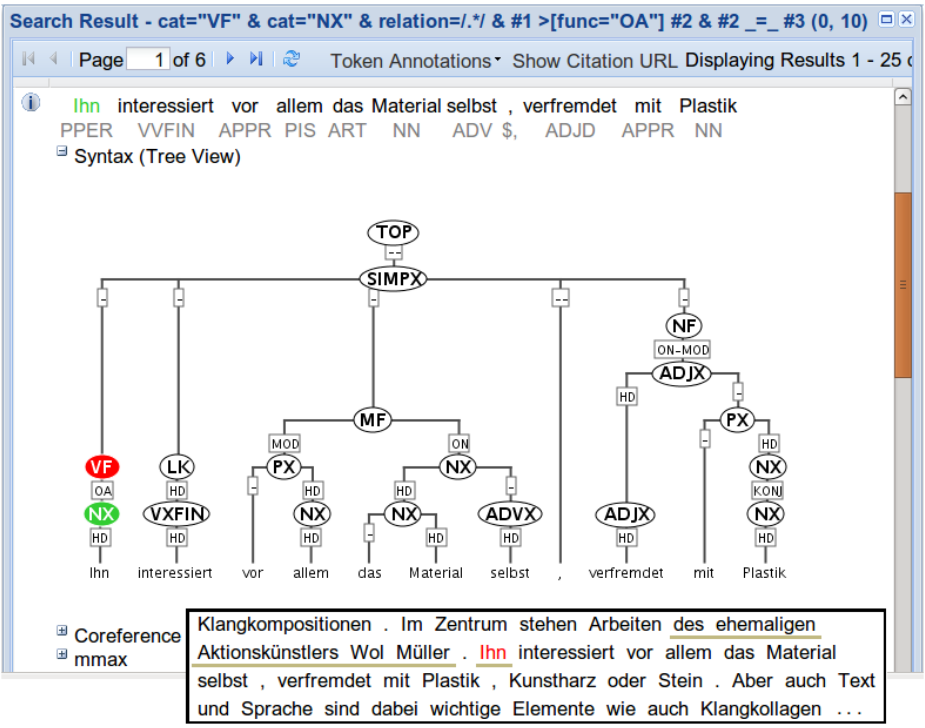


Abbildung 1: Screenshot von ANNIS2: Treffer der Suche nach einem vorerwähnten Objekt im Vorfeld (aus TüBa-D/Z). ANNIS2 markiert die Knoten 'VF' und 'NX', die in der Anfrage spezifiziert wurden, im Baum farbig. In der eingefügten Box unten ist ein Ausschnitt aus der Koreferenzansicht von ANNIS2 gezeigt, koreferente Ausdrücke werden hier mit der gleichen Farbe unterstrichen (im Beispiel: *des ehemaligen Aktionskünstlers Wol Müller* und *ihn*).

Zugang zu den Ressourcen In dem oben geschilderten Szenario sollen Studenten selbständig in Korpora nach relevanten Belegen suchen. Voraussetzung dafür ist zum einen, dass die Korpora und Suchtools frei verfügbar sind (für die Lehre). Alle oben aufgeführten Korpora und Tools erfüllen diese Bedingung: sie sind für nicht-kommerzielle Anwendungen kostenlos lizenzierbar.

Zum andern sollten die technischen Hürden möglichst niedrig sein, sowohl bei der Installation wie auch bei der Bedienung der Tools. Zum aktuellen Zeitpunkt bieten sich dafür Tools an, die auf einem Server installiert und von den Studenten über einen Webbrowser bedient werden können. Auf diese Weise müssen Prozesse wie das Installieren des Suchtools oder das Aufbereiten und Einlesen der Korpora nicht unzählige Male durchgeführt werden (nämlich von jedem einzelnen Studenten auf dem eigenen Rechner), sondern passieren einmalig auf dem zentralen Server.

Von den oben erwähnten Korpustools ist ANNIS2 ein solches webbasiertes Tool. Ein weiteres Beispiel ist CQPweb[10], das allerdings etwas andere Funktionalitäten als ANNIS2 bietet: Es unterstützt keine Baum- oder Zeigerstrukturen, sondern nur Wort- oder Spannen-basierte Annotationen; andererseits können in CQPweb auch Parallelkorpora abgefragt werden.

Eine andere Möglichkeit ist, Korpora zu nutzen, die zwar nicht lizenzierbar sind, aber über ein Web-Interface abgefragt werden können. Dazu zählen z.B. die DWDS-Korpora, die mit POS und Lemmata annotiert sind, oder COSMAS II, dessen Korpora mit POS, Lemmata und Morphologie annotiert sind.[11]

Approximationen Bei Korpora, die z.B. nur mit Wortarten ('parts of speech', POS) annotiert in CQPweb vorliegen, lassen sich syntaktische Anfragen mit Hilfe *regulärer Ausdrücke* über den POS-Annotationen approximieren. Der Suchausdruck in (4), formuliert in der Anfrage-Sprache CQP, erfasst beispielsweise eine Teilmenge der Akkusativ-NPs am Satzanfang, denen ein finites Verb folgt.[12]

(4) `[pos="ART" & word="(Den|Einen)"][pos="ADJA"]*[pos="N."][pos="V.FIN"]`

Schon in der Anfrage in (3) haben wir diese "Technik" genutzt, nicht explizit annotierte Information mit Hilfe der vorhandenen Annotation zu approximieren. Die Annotationen im NEGRA-, TIGER- und dem PCC-Korpus erlauben eine gute Approximation der Vorfeld-Position, daher ist dieses Merkmal in Tabelle 1 mit '(✓)' markiert.

[10] Web-Interface für den "Corpus Query Processor", `http://cwb.sourceforge.net/`. Über CQPweb wird auch eine Reihe von Demo-Korpora angeboten, `http://cwb.sourceforge.net/demos.php`.

[11] DWDS: "Digitales Wörterbuch der Deutschen Sprache", `http://www.dwds.de/`
COSMAS: "Corpus Search, Management and Analysis System", `https://cosmas2.ids-mannheim.de/cosmas2-web/`

[12] Die POS-Tags in (4) folgen dem STTS (Schiller et al., 1999): 'ART' markiert Artikel, 'ADJA' attributive Adjektive, 'N.' steht für 'NN' oder 'NE', d.h. für allgemeine Nomen, 'V.FIN' für 'VAFIN', 'VMFIN' oder 'VVFIN', d.h. für finite Verben. Die abgefragte NP beginnt mit einem definiten oder indefiniten maskulinen Artikel, da dieser eindeutig für Akkusativ markiert ist; die Großschreibung bewirkt, dass nur Artikel am Satzanfang gefunden werden. Danach stehen beliebig viele Adjektive, gefolgt vom Kopfnomen. Der Ausdruck erfasst also beispielsweise weder feminine noch artikellose NPs oder NPs mit postnominalen Modifikatoren.

Die Beispiele in diesem Abschnitt illustrieren, dass das linguistische Wissen, das bei der Suche nach relevanten Belegen nötig ist, auf unterschiedliche Weise einfließen kann: Entweder das linguistische Wissen ist bereits im Korpus vorhanden, in Form von komplexen Annotationen, die nur noch abgefragt werden müssen. Oder das Korpus enthält nur einfache Annotationen, die in einer komplexen Anfrage miteinander kombiniert werden; hier verteilt sich das linguistische Wissen auf Annotation und Abfrage. Schließlich sichtet man gegebenenfalls die Treffer von Hand, was ebenfalls den Einsatz von linguistischem Wissen erfordert.

2.2 Korpora als "Testdaten"

Im zweiten Einsatzszenario geht es um ein linguistisches Phänomen, das nicht so deutlich abgrenzbar ist wie die Vorfeldbesetzung: die Realisierung des *(Aboutness-)Topiks*. In der linguistischen Forschung wird als ein Test für die Bestimmung des Topiks eines Satzes die Einleitungsphrase *Ich erzähle dir etwas über X* genannt. Die Konstituente, die an Stelle des 'X' eingesetzt werden kann, ohne dass der Sinn des nachfolgenden Satzes verändert wird, ist die Topik-Konstituente des Satzes. Im Beispiel (5) (entnommen aus Frey (2004)) ergibt der Test, dass *Maria* die Topik-Konstituente ist (im Beispiel unterstrichen). Als alternative Einleitungsphrasen werden genannt: *Was ist mit X?* oder *Was X anbelangt, ...* (Götze et al., 2007).

(5) (Ich erzähle dir etwas über Maria.)
Nächstes Jahr wird Maria wahrscheinlich nach London gehen.

Topiks sind oft Subjekte, außerdem stehen sie oft im Vorfeld oder am Anfang des Mittelfeldes (Frey, 2004). Geht man allerdings weg von den konstruierten Beispielsätzen aus der Forschungsliteratur (wie Bsp. (5)) hin zu Alltagstexten, stellt man schnell fest, dass die Bestimmung des Satztopiks oftmals nicht trivial ist. Angewendet auf fortlaufenden Text führen die genannten Tests mit Einleitungsphrasen häufig zu unnatürlichen Ergebnissen und unklaren Intuitionen. (6) zeigt einen Beispielsatz im originalen Kontext, der zwei gleichwertige Topik-Kandidaten enthält (entnommen aus Cook and Bildhauer (2011)).

(6) (Dazugelernt habe ich besonders im Bereich der Öffentlichkeitsarbeit. Ich merkte, welche Handlung welche Reaktion auslöst und wie man gewisse Ereignisse richtig kommuniziert.)
Von dieser Erfahrung$_{top?}$ kann ich$_{top?}$ am neuen Ort selbstverständlich profitieren.

D.h. im Vergleich zum Begriff des Vorfeldes ist der Begriff des Topiks vage und schwierig anzuwenden. Entsprechend gibt es bislang auch nur wenige Versuche, Korpora mit Topiks zu annotieren — und folglich stehen für die Lehre erstmal keine vorannotierten Korpora zur Verfügung. Eine Möglichkeit ist daher, die Studenten nach einer theoretischen Einführung in die Thematik selbst Texte annotieren zu lassen. Für die (manuelle) Annotation benötigt man, neben dem zu annotierenden Text, (i) ein Annotationstool und (ii) Annotationsguidelines.

Annotationstools Der Einsatz von Annotationstools im Kurs hat mehrere Vorteile gegenüber der Annotation auf Papier: Die Studenten lernen den Umgang mit den Tools; die Daten können nachhaltig gespeichert werden (und z.B. in nachfolgenden Sitzungen weiter verwendet werden); Daten können kollaborativ annotiert werden; die Annotatoren-Übereinstimmung kann berechnet und kontrovers annotierte Sätze können automatisch bestimmt werden.

Aus den schon oben genannten Gründen gilt auch hier, dass das Annotationstool möglichst webbasiert operieren sollte, also der Zugang und die Bedienung über einen Webbrowser möglich sein sollte. Da der Nutzer hier eigene Texte annotieren will, muss das Tool eine Upload-Funktion oder eine Texteingabe anbieten. Neben einer Download-Funktion wäre außerdem ein direkter Weg vom Annotationstool in ein Korpus-Suchtool ideal, in dem die annotierten Daten gleich durchsucht werden können.

Für den gelegentlichen Einsatz im Linguistik-Unterricht ist es wichtig, dass das Annotationstool möglichst einfach und intuitiv bedient werden kann. Das Annotations-Interface sollte sich daher an gängigen Editoren orientieren (z.B. in der Verwendung von Tastaturbefehlen). Eine Unterstützung des Text-Uploads durch einen integrierten Tokenisierer ist denkbar.

Zum aktuellen Zeitpunkt ist der überwiegende Teil der verfügbaren Annotationstools als *stand alone*-Anwendung realisiert, die lokal installiert werden muss. Zu den wenigen Ausnahmen gehören die Tools Serengeti, das für die Annotation von Koreferenz-Beziehungen entwickelt wurde, und Typecraft, mit dem wortbasierte Annotationen erstellt werden können.[13] Daneben werden seit einiger Zeit Web-Interfaces für (meist kommerzielle) *crowd sourcing*-Annotationen entwickelt.

Annotationsguidelines Werden die bekannten Tests als Kriterien für die Annotation genommen, so werden die Studenten schnell feststellen, dass die Tests häufig nur unbefriedigende Ergebnisse liefern und die Intuitionen oft unklar bleiben. Eine (anspruchsvolle) Aufgabe für die Studenten könnte dann darin bestehen, das Konzept 'Topik' besser zu operationalisieren, d.h. bessere linguistische Tests dafür zu entwickeln. Angewendet auf fortlaufende Alltagstexte und sukzessive erweitert und verfeinert, können diese Tests schließlich in "robuste" Guidelines münden, die für alle denkbaren Fälle eindeutige Kriterien der Annotation vorsehen.

Auf diese Weise wird das Korpus letztlich zum "Testkorpus" für die linguistische Theorie: Mit einer wachsenden Menge von klar definierten Kriterien und von annotierten Daten ergeben sich Hinweise darauf, welche sprachliche Realität Konzepte wie Topik im Deutschen bzw. in Korpora des Deutschen haben. Ein annotiertes Korpus kann außerdem als Test für die Vorhersagen der Theorie dienen.

Ein kurzes Zwischenfazit: Im ersten Szenario war das Phänomen klar umrissen (Vorfeld-Besetzung); der Fokus lag auf der Suche nach markierten Einzelbelegen in annotierten Korpora. Dagegen ist im zweiten Szenario das Phänomen schwerer fassbar (Topik) und

[13] Serengeti: "Semantic Relations Annotation Tool", `http://coli.lili.uni-bielefeld.de/serengeti/`
Typecraft: `http://www.typecraft.org/`

die linguistische Analyse noch unklar; der Fokus liegt daher auf der Untersuchung und Annotation von fortlaufendem Text.

Für beide Szenarien gilt, dass die Annotations- und Suchtools satzübergreifende Relationen mit abdecken und die Bedienung webbasiert erfolgen sollte. Idealerweise sollte der Output der Annotationstools direkt in ein Suchtool eingespeist werden können.

3 Korpora in der Computerlinguistik

Die beiden nächsten Erfahrungsberichte stammen aus einem Bachelor-Studiengang für Computerlinguistik. Hier geht es um kleinere Implementationsaufgaben, die von den Studenten weitgehend eigenständig bearbeitet werden (im Rahmen sogenannter "Forschungsprojekte").

3.1 Korpora in der Anwendungsentwicklung

Als Anwendungsbeispiel soll die Aufgabe der *Autorenschaft-Zuschreibung* dienen: Gegeben eine Menge an Texten von verschiedenen Autoren wird ein Klassifikator gesucht, der einen neuen Text einem der Autoren zuweist.

Vorgabe im Kurs war es, nicht einen reinen *bag of words*-Ansatz zu implementieren, bei dem der Klassifikator ausschließlich *N-Gramme* (Sequenzen) von Wortformen nutzt. Stattdessen sollten auch linguistische Merkmale wie Wortart (POS), Morphologie o.ä. eine Rolle spielen.

Die erste Aufgabe der Studenten bestand darin, sich nach geeigneten Ressourcen umzusehen. Das betraf zum einen die Textgrundlage, d.h. eine Menge von geeigneten, vergleichbaren Texten, bei denen die Autorenschaft bekannt ist. Zum andern betraf es automatische Analysetools, die zu diesen Texten die linguistischen Merkmale für den Klassifikator liefern sollten.

Im Folgenden werden zwei Studentenprojekte mit verschiedenen Textgrundlagen vorgestellt.

Beispiel-Korpus: Enron Eine Gruppe wählte als Korpus das *Enron Email Dataset* (Klimt and Yang, 2004; Shetty and Adibi, 2004). Enron war ein amerikanischer Energiekonzern, der nach umfangreichen Bilanzfälschungen 2001 Insolvenz anmelden musste. Im Zuge der Ermittlungen gegen Enron stellte die amerikanische Bundesaufsichtsbehörde für Energie den (privaten wie beruflichen) Email-Verkehr von 150 leitenden Enron-Mitarbeitern im Internet frei zur Verfügung. Von verschiedenen Forschungsinstituten wurden diese Daten für Forschungszwecke weiter aufbereitet, indem z.B. Duplikate und leere Mails entfernt wurden. Diese Daten, das Enron Email Dataset, können frei heruntergeladen werden.[14] Das Korpus enthält die Ordner der Posteingänge der Mitarbeiter, d.h. die Emails sind nach ihren Empfängern sortiert. Abb. 2 zeigt eine Email, wie sie im Korpus enthalten ist.

[14] Z.B. unter `http://www.cs.cmu.edu/~enron/`

```
Date: Tue, 5 Dec 2000 02:51:00 -0800 (PST)
From: denise.lagesse@enron.com
To: drew.fossum@enron.com
Subject: Susan's expense report 11/16/00
Cc: susan.scott@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: susan.scott@enron.com
X-From: Denise LaGesse
X-To: Drew Fossum
X-cc: Susan Scott
X-bcc:
X-Folder: \Drew_Fossum_Dec2000_June2001_1\Notes Folders\Notes inbox
X-Origin: FOSSUM-D
X-FileName: dfossum.nsf

Drew,

Did you already approve this expense report and send it to accounting?

Denise

---------------------- Forwarded by Denise LaGesse/ET&S/Enron on 12/05/2000
10:50 AM ---------------------------


DENISE LAGESSE
11/16/2000 04:04 PM
To: Drew Fossum/ET&S/Enron@ENRON
cc: Susan Scott/ET&S/Enron@ENRON

Subject: Susan's expense report 11/16/00

Please approve the attached expense report and cc: me on your approval.
Thanks!
```

Abbildung 2: Email aus dem Enron-Korpus. Es handelt sich um die Datei 'fossum-d/notes_inbox/103' aus dem Release vom 21. August 2009. Im Email-Header finden sich neben den Standard-Einträgen gemäß dem Internet-Standard RFC 5322 auch Nutzer-definierte Einträge, erkennbar am Präfix 'X-', die z.B. vom Email-Client-Programm eingefügt sein können. Der untere Teil der Email besteht aus einer weitergeleiteten Email, die durch eine Strich-Markierung eingeleitet wird. Der eigentliche Text, der dem Absender als Autor zugeordnet werden kann, besteht also aus nur einer Zeile (*Did you already ...*), dazu kommen die Grußformel (*Drew,*) und die Abschiedsformel (*Denise*).

Vor der eigentlichen Projektaufgabe stand eine Reihe von Vorverarbeitungen an: Der komplette Header wurde gelöscht, die Information über Absender und Adressat in anderer Form abgespeichert. Abschnitte, die aus anderen Mails in Form von Zitaten übernommen wurden, mussten erkannt und gelöscht werden (da sie von anderen Autoren stammen). Ebenso wurden Abschnitte, die aus weitergeleiteten Mails bestanden, entfernt (s. die Beispiel-Email in Abb. 2). Gegebenenfalls musste Markup, wie z.B. HTML-Tags, gelöscht werden. Alle verbleibende Information wurde in einem XML-Format abgelegt. Zuletzt wurden die Mails umgeordnet: statt nach ihrem Empfänger wurden sie nach ihrem Absender sortiert (da es im Projekt ja um die automatische Erkennung des Autors ging).

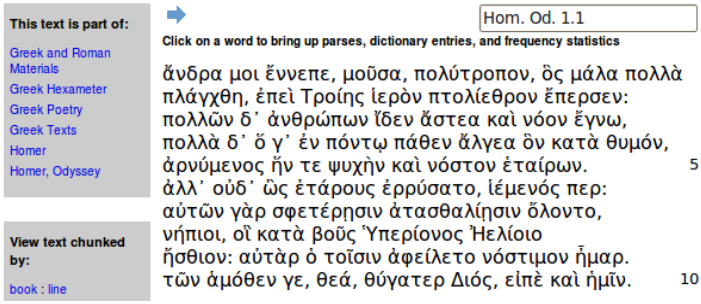
Als nächster Schritt stand die automatische Analyse mit Hilfe frei verfügbarer Tools an. Hierfür wählte die studentische Gruppe als POS-Tagger den Stanford-Tagger und als Parser RASP.[15] Der ursprüngliche Plan sah vor, ausgewählte Merkmale, die aus den Analysen des Taggers und Parsers extrahiert werden sollten, mit WEKA[16] weiter zu verarbeiten, einem Tool, das eine Reihe von Algorithmen für *data mining* für die Anwendung bereit stellt. Allerdings erwies sich die Datenmenge als zu groß für eine Verarbeitung durch den Tagger und Parser (im Rahmen des Kurses). Letztlich wurde der Klassifikator daher auf oberflächennahen Merkmalen trainiert: Anzahl der Sätze in der Email, durchschnittliche Satzlänge, Type-Token-Ratio, Grußformel (abstrahiert), prozentualer Anteil an Satzzeichen. Als besonders distinktives Merkmal stellte sich die Grußformel heraus.

Beispiel-Korpus: Homer Eine zweite Gruppe von Studenten wählte als Textgrundlage die Ilias und Odyssee von Homer. In der *Homerischen Frage* geht es darum, ob die beiden Epen von nur einem Autor geschrieben wurden — eine bis heute nicht gelöste Frage. Laut dem *Unitarismus* gab es einen einzigen Dichter beider Werke. Im Gegensatz dazu besagt die *Oral-Poetry-Theorie*, dass die Epen zuerst mündlich improvisiert und tradiert wurden, bevor sie schriftlich festgehalten wurden. Nach der *Analyse*-Theorie wiederum gab es für beide Epen jeweils einen "Hauptdichter", dessen Werk im Laufe der Zeit durch mehrere "Nebendichter" ergänzt und angereichert wurde.

Die Studenten hatten entsprechend zwei Aufgaben: zum einen zu berechnen, wie homogen die beiden Epen in sich sind (also Teilabschnitte mit dem Gesamtepos zu vergleichen); zum anderen zu berechnen, wie ähnlich die beiden Epen zueinander sind (und das Ergebnis in Relation zu setzen zum ersten Ergebnis).

Die Texte Homers werden über das Projekt Perseus angeboten und sind (automatisch) annotiert mit POS und Morphologie.[17] Allerdings sind die Texte und Annotationen nicht direkt downloadbar, sondern nur über ein Web-Interface zugänglich, s. Abb. 3. Durch einen Mausklick auf ein Wort erhält man die zugehörige POS- und morphologische Analyse.

[15] Stanford Tagger: `http://nlp.stanford.edu/software/tagger.shtml`
RASP: `http://www.informatics.sussex.ac.uk/research/groups/nlp/rasp/`
[16] `http://www.cs.waikato.ac.nz/ml/weka/`
[17] `http://www.perseus.tufts.edu/hopper/`




```
<a href="morph?l=a%29%2Fndra&amp;la=greek">ἄνδρα</a>
<a href="morph?l=moi&amp;la=greek&amp;prior=a)/ndra">μοι</a>
<a href="morph?l=e%29%2Fnnepe&amp;la=greek&amp;prior=moi">ἔννεπε</a>,
<a href="morph?l=mou%3Dsa&amp;la=greek&amp;prior=e)/nnepe">μοῦσα</a>,
<a href="morph?l=polu%2Ftropon&amp;la=greek&amp;prior=mou=sa">πολύτροπον</a>,
<a href="morph?l=o%28%5Cs&amp;la=greek&amp;prior=polu/tropon">ὃς</a>
<a href="morph?l=ma%2Fla&amp;la=greek&amp;prior=o(\s">μάλα</a>
<a href="morph?l=polla%5C&amp;la=greek&amp;prior=ma/la">πολλὰ</a><br />
```

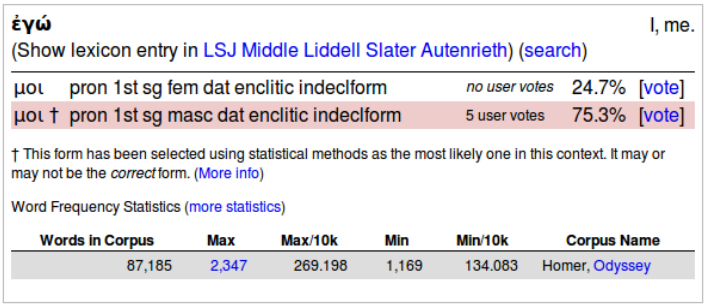


Abbildung 3: Screenshots vom Perseus-Projekt: (i) Oben ist der Beginn der Odyssee von Homer abgedruckt. (ii) In der mittleren Box steht die (verkürzte) HTML-Darstellung der ersten Verszeile. Aus dem 'href'-Attribut wird ersichtlich, dass für die morphologische Analyse, die durch Mausklick abgerufen wird, intern Betacode benutzt wird (z.B. ist '%29%2F' die Hexadezimal-Darstellung von ')/' und steht für die altgriechischen Diakritika über dem ersten Buchstaben im Text). Außerdem kann man sehen, dass für die morphologische Analyse das jeweilige Vorgängerwort ('prior') mitgenutzt wird. (iii) Unten wird der Lexikoneintrag zum zweiten Wort gezeigt, dem Personalpronomen *μοι* 'mir', mit POS- und morphologischer Analyse. Die statistische Disambiguierung präferiert hier die maskuline Lesart des Pronomens.

Mit Hilfe von Download-Programmen wie *wget* konnte die Projektgruppe den vollständigen Text der Ilias und der Odyssee (in Betacode) herunterladen. Eine Nutzung des Perseus-Tools für die POS- und Morphologie-Analyse war jedoch losgelöst vom Web-Interface nicht möglich, so dass letztlich das heruntergeladene Korpus nicht automatisch annotiert werden konnte. Statt linguistische Merkmale zu nutzen, musste die Projektgruppe daher doch auf einen *bag of words*-Ansatz zurückgreifen.

3.2 Korpora für den Ressourcenaufbau

Im zweiten Beispiel aus der Computerlinguistik werden Korpora für den Aufbau von Sprachressourcen genutzt. Konkrete Aufgabe für die Studenten war, mit Hilfe eines Korpus ein deutsches Nomenlexikon mit Angabe der Flexionsklasse zu erstellen.

Die Aufgabe wurde von zwei Gruppen bearbeitet, die mit unterschiedlichen Ressourcen arbeiteten: Gruppe 1 verwendete ein deutsches Zeitungskorpus, das mit Hilfe des RFTaggers[18] automatisch annotiert worden war, und zwar mit POS-, Lemma- und morphologischer Information. Hier bestand die Aufgabe darin, aus Tupeln bestehend aus einem Lemma, den zugehörigen Wortformen und der Flexionsinformation die damit kompatible Flexionsklasse zu bestimmen. Für die Fälle, in denen nicht genügend Belege für die verschiedenen Flexionsformen gefunden wurden, mussten Heuristiken entwickelt werden, um die wahrscheinlichste Klasse zu bestimmen. Außerdem waren Lösungen zu entwickeln für homographe Lemmata mit unterschiedlichen Flexionsklassen.[19]

Gruppe 2 verwendete dasselbe Korpus, aber nutzte neben den Wortformen nur die POS-Annotation. Hier musste also zusätzlich eine Methode entwickelt werden, zusammengehörige Wortformen als solche zu erkennen. Mit diesem Ansatz konnten dann auch (Vorschläge für) Lexikoneinträge von Nomen, die vom RFTagger nicht lemmatisiert wurden, erzeugt werden — das waren im betreffenden Korpus immerhin rund 20% aller Nomen (11% der normalen Nomen, 45% der Eigennamen).

Als Fazit aus den Szenarien in diesem Abschnitt lässt sich festhalten, dass es für die computerlinguistische Lehre essenziell ist, dass die Korpora frei zum Download und zur lokalen Weiterverarbeitung zur Verfügung stehen. Außerdem sollten Analysetools ebenfalls für den lokalen Einsatz genutzt werden können. Alternativ könnten (frei verfügbare) annotierte Korpora dazu verwendet werden, eigene Tools zu trainieren.

Während im ersten Szenario die Textauswahl eine große Rolle für die Aufgabe spielte, war die Textgrundlage im zweiten Szenario sekundär. Hier kam es vorwiegend darauf an, dass die Texte mit gängigen Analysetools mit zufriedenstellender Performanz automatisch annotiert werden konnten.

[18] `http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/`

[19] Z.B. lautet für das Lemma *Bank* die Form des Nominativ Plural in der Bedeutung als Geldinstitut *Banken*, in der Bedeutung als Sitzmöbel *Bänke*.

4 Schluss

In den vorgestellten Szenarien aus der Lehre spielen Korpora eine unterschiedliche Rolle: In der Theoretischen Linguistik dienen sie einerseits als Belegsammlung, aus der interessante Einzelbelege mit Hilfe der Annotationen und geeigneter Suchanfragen gezielt herausgesucht werden können. Andererseits werden Korpora als Gegenstand für die eigene (manuelle) Analyse und Annotation gesehen; die relevanten Merkmale sind hier noch nicht (vollständig) annotiert. In beiden Fällen geht es darum, Korpora zur Entwicklung und Validierung linguistischer Theorien zu nutzen.

Die Computerlinguistik kann in ähnlicher Weise auf bereits annotierte Korpora zurückgreifen und ausgewählte Merkmale für die (automatische) Analyse nutzen. Häufig sind geeignete Korpora allerdings nicht oder nicht frei verfügbar. Dann ist es Teil der Aufgabe, Korpora selbst aufzubereiten und automatisch zu annotieren.

Aus den Szenarien ergeben sich unterschiedliche Anforderungen an die Sprachressourcen. Auf der einen Seite sollen technische Details und Hürden (wie z.B. die interne XML-Kodierung der Daten oder die Tool-Installation) in der Lehre für theoretische Linguisten ausgeklammert werden können. Daher plädiere ich dafür, dass Korpora und Annotationstools (für die manuelle Annotation) über das Web benutzbar sein sollen. Wichtig wäre ein Anschluss der Annotationstools an Korpus-Suchtools.

Auf der anderen Seite benötigt man in der Lehre für die Computerlinguistik den vollen Zugriff auf die Ressourcen. Für viele Anwendungen werden zudem größere Datenmengen benötigt, was allerdings dank freier Quellen wie Wikipedia für viele Sprachen kein Problem mehr ist. Freie Tools für automatisches POS-Tagging gibt es (mit Sprachmodellen) für "gängige" Sprachen wie Englisch oder Deutsch in genügender Anzahl. Tools für das Parsing oder Chunking hingegen sind seltener und oft nicht effizient genug.

Bei der Arbeit mit "echten" Daten, in realistischen Szenarien, werden die Computerlinguistik-Studenten früh mit Problemen wie dem Daten-Encoding oder der Datengröße konfrontiert, die mit Sicherheit auch in ihrem späteren Berufsleben eine Rolle spielen werden.

Zuletzt ein Vergleich zum Verhältnis zwischen dem Gegenstand *Text* und dem daraus gewonnenen Wissen in den verschiedenen Szenarien. Texte stellen linguistisches Wissen in *extensionaler* Weise dar, dagegen haben linguistische Hypothesen und Theorien den Anspruch, linguistisches Wissen in *intensionaler* Weise abzubilden. Anders formuliert könnte man sagen, dass Texte die "Produkte angewandten linguistischen Wissens" sind, während linguistische Theorien die zugrunde liegenden Regeln und Generalisierungen, d.h. das linguistische Wissen selbst erfassen wollen.

Aufgabe der Theoretischen Linguistik ist es nun, aus dem Text, d.h. aus den "beobachtbaren" sprachlichen Daten, auf die dahinter liegenden Regularitäten und Gesetzmäßigkeiten rückzuschließen. Auf ganz ähnliche Art versuchen Computerlinguisten, mit statistischen Methoden geeignete Sprachmodelle zu erstellen, die die beobachteten Daten möglichst optimal erklären.

In beiden "Lagern", der Theoretischen Linguistik wie der Computerlinguistik, wird der Weg vom extensionalen Objekt (dem Text) zum intensionalen Objekt (der Theorie/dem Modell) dadurch vereinfacht, dass geeignete *Abstraktionen* vorgenommen werden: Durch manuelle oder automatische Annotationen (POS-Tagging und andere Analyseschritte) werden sprachliche Daten *reduziert* auf die für die Theorie- und Modellbildung relevanten Eigenschaften und vereinfachen dadurch das Aufspüren von Regularitäten. In diesem Abstraktionsschritt werden die Daten allerdings nicht nur reduziert, sondern gleichzeitig auch *angereichert*: Durch die Abstraktion wird nämlich sprachliches Wissen, das im Text nur implizit (nämlich extensional) ausgedrückt ist, *explizit* gemacht.

Literatur

Brants, T. (1997). The NeGra export format. CLAUS Report Nr. 98. Universität des Saarlandes, Computerlinguistik, Saarbrücken.

Cook, P. and Bildhauer, F. (2011). Annotating information structure: The case of topic. In Dipper, S. and Zinsmeister, H., editors, *Proceedings of the Workshop 'Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, volume 3 of *Bochumer Linguistische Arbeitsberichte (BLA)*, pages 45–56.

Filippova, K. and Strube, M. (2007). German Vorfeld and local coherence. *Journal of Logic, Language, and Information (JoLLI)*, 16(4):465–485. Special Issue on Coherence in Dialogue and Generation.

Frey, W. (2004). A medial topic position for German. *Linguistische Berichte*, 198:153–190.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., and Stoel, R. (2007). Information structure. In Dipper, S., Götze, M., and Skopeteas, S., editors, *Information Structure in Cross-linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS)*, pages 147–187. Universitätsverlag Potsdam.

Klimt, B. and Yang, Y. (2004). Introducing the Enron corpus. In *Proceedings of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS-2004)*.

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC)*, pages 121–126.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Universitäten Stuttgart und Tübingen, `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf`.

Shetty, J. and Adibi, J. (2004). The Enron email dataset: Database schema and brief statistical report. Technical report, Information Sciences Institute. `http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf`.

Speyer, A. (2007). Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 26:83–115.