Noah Bubenhofer

Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge

Für die sprachwissenschaftliche Ausbildung an den Universitäten ist es zwar unabdingbar, die Studierenden in die Theorie und Methoden der Korpuslinguistik einzuführen, doch als Lehrperson kämpft man dabei mit einer Reihe von Problemen, denn das technische und methodische Know-how der Studierenden ist oft sehr heterogen. Zudem zeigt sich die Wichtigkeit, die Studierenden für korpuslinguistisches Arbeiten begeistern zu können, indem sie an attraktives Anschauungsmaterial herangeführt werden. Im Folgenden zeige ich an einigen Beispielen, welche Themen in den Bereichen Semantik, Textlinguistik, Diskurs- und der Kulturanalyse sinnvollerweise korpuslinguistisch bearbeitet werden können. Zudem versuche ich anhand des Nutzungsverhaltens meiner Online-Einführung in die Korpuslinguistik die Bedürfnisse von Anwendern an Methoden und Werkzeuge der Korpuslinguistik abzuleiten.

1 Einleitung

Korpuslinguistische Methoden gehören inzwischen zum Standardrepertoire vieler Forscherinnen und Forscher in der Sprachwissenschaft. So verwundert es nicht, dass die Korpuslinguistik auch Eingang in die linguistische Lehre findet – auch wenn dies erst zaghaft und nicht in allen Teildisziplinen gleichermaßen geschieht.

Dabei müssen allerdings eine Reihe von Hürden überwunden werden:

- Korpuslinguistisches Arbeiten setzt Wissen über empirische Forschung generell voraus. Diese Grundlage muss teilweise erst aufgebaut werden.
- Teilweise fehlt bei den geisteswissenschaftlich geprägten Studierenden das technische Know-how, um korpuslinguistisch arbeiten zu können. Zudem ist das manchmal gepaart mit dem fehlenden Selbstvertrauen, sich dieses Know-how anzueignen.
- Oft bedingen korpuslinguistische Arbeiten einen großen Aufwand, sowohl für Lernende als auch die Betreuenden, der im Rahmen eines Studiums nicht geleistet werden kann.

Trotzdem ist es gewinnbringend, mit Studierenden korpuslinguistisch zu arbeiten. Im Folgenden möchte ich deshalb an einigen Beispielen aus meiner eigenen Lehrpraxis

zeigen, welche Themen in den Bereichen Semantik, Textlinguistik, Diskurslinguistik und Kulturanalyse bearbeitet werden können.

Bei korpuslinguistischem Arbeiten geht es immer auch um technische Aspekte: Welche Ressourcen und Werkzeuge gibt es, die von Anwenderinnen und Anwendern ohne Informatik-Hintergrund leicht eingesetzt werden können? Die Präsentation der Beispiele studentischer Arbeiten wird deshalb ergänzt durch Hinweise auf Ressourcen und Werkzeuge, die sich in der Lehre bewährt haben. Trotzdem bleiben Wünsche an die Entwicklung von Analysewerkzeug offen, die ich skizzieren möchte.

Um diese Wünsche und Bedürfnisse auf eine breitere Basis zu stellen, habe ich die Zugriffe auf die online verfügbare „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011) analysiert und ein Nutzerprofil erstellt, das ich ebenfalls präsentiere.

2 Korpuslinguistik in der Lehre

2.1 Konzepte und Ziele

Meine Erfahrung zu Korpuslinguistik in der Lehre beruht auf folgenden Lehrveranstaltungen und Betreuungsangeboten:

- Kurse „Einführung in die Korpuslinguistik“ im Grundstudium/Bachelor-Studiengang am Deutschen Seminar der Universität Zürich.
- Seminare an den Universitäten Zürich und Mannheim (Hauptstudium/Master-Studiengang) zu den Themen Diskurs-/Kulturanalyse und Semantik (Lexikographie), in denen korpuslinguistisches Arbeiten die methodische Grundlage darstellte.
- Betreuung von Seminar- und Abschlussarbeiten zu verschiedenen Themen mit korpuslinguistischer Methodik.
- Verschiedene Workshops und Beratung für fortgeschrittene Studierende und Doktorierende zu korpuslinguistischen Themen in den Bereichen Semantik, Grammatik, Textlinguistik, Diskurs- und Kulturanalyse.

Das Ziel sowohl der einführenden Kurse als auch der Seminare war, die Studierenden an empirisches Arbeiten mit Korpora heranzuführen. Es geht also darum, den Weg von einer Hypothese zur Operationalisierung und Analyse aufzuzeigen und sich grundlegende Gedanken über den Stellenwert von Korpora als Datengrundlage für Analysen des Sprachgebrauchs zu machen. Darauf aufbauend lernen die Studierenden verfügbare Ressourcen und Werkzeuge kennen und anhand eigener Fragestellungen anzuwenden. Ein typischer Seminarplan einer solchen Einführung ist in Tabelle 1 dargestellt.[1] Dabei bewegt man sich zwischen den folgenden Polen:

[1] Als begleitende Literatur nutzte ich neben themenspezifischer Literatur die Einführungen Lemnitzer/Zinsmeister (2006) und Scherer (2006).

1. **Grundlagen:** Begriffsklärung, korpuslinguistische Denkweise (Arm-Chair Linguist vs. Corpus Linguist), Anwendungen
2. **Empirisches Arbeiten:** Thesenbildung, Operationalisierung
3. **Korpora Grundlagen:** Repräsentativität, Korpusgröße, Korpustypen, Annotation
4. **Bestehende Korpora nutzen:** DeReKo IDS (o. J.), DWDS (o. J.), Baumbanken
5. **Methoden:** Recherchen, Ergebnisdarstellung, Kollokationen, n-Gramme, statistische Auswertungen
6. **Eigene Korpora aufbauen**

Tabelle 1: Typischer Seminarplan „Einführung in die Korpuslinguistik“

Pol 1: Analyseebenen Belege finden ↔ Muster entdecken ↔ systematische statistische Auswertungen durchführen

Die Studierenden lernen verschieden elaborierte Methoden der Korpusanalyse kennen, wobei der einfachste Zugang über klassische, korpusbasierte Analysen erfolgt: Konkordanzen und Belege analysieren und kategorisieren, Kollokationsanalysen, Vergleiche mit anderen Korpora. Ziel ist es aber, fortgeschrittenere Analysen zu machen, mit denen die Beobachtung von Einzelbelegen systematisiert und abstrahiert werden.

Pol 2: Datengrundlage bestehende Korpora nutzen ↔ eigene Korpora aufbauen

Es sind bereits viele sofort nutzbare Korpora verfügbar, die den Einstieg in Korpusanalysen erleichtern. Für viele Forschungszwecke ist es aber sinnvoll, eigene Korpora erstellen zu können.

Pol 3: Know-how viel technisches Know-how ↔ wenig technisches Know-how

Das vorhandene computertechnische Know-how der Studierenden ist meistens sehr heterogen. Deshalb ist es wichtig, auf Studierende mit eher wenig Know-how Rücksicht zu nehmen und sie aber gleichzeitig zu ermutigen, trotzdem anspruchsvolle Analysemethoden auszuprobieren. Oft können Projekte auch in Gruppen durchgeführt werden, in denen sich die Studierenden mit ihren unterschiedlichen Fähigkeiten ergänzen.

Aus diesen Polen leiten sich bereits die Bedürfnisse nach Ressourcen und Werkzeugen her, die idealerweise für die Lehre verfügbar sind, um alle Aspekte abdecken zu können: Man benötigt Software, um allen Analyseebenen in bestehenden aber auch selbst aufgebauten Korpora nachgehen zu können, wobei diese Software sowohl für „Poweruser“ als auch technisch wenig bewanderte Studierende nutzbar sein sollte.[2]

[2] Dies gilt natürlich nicht nur im Bereich der Lehre, sondern genau so in der Forschung.

Wichtig, um die Studierenden zu korpuslinguistischen Arbeiten zu animieren, sind anschauliche Beispiele solcher Arbeiten aus der Forschung.[3] Einerseits machen intuitiv nutzbare Korpusrecherche-Tools wie Wortschatz Leipzig (o. J.), der Google Books Ngram Viewer (o. J.)[4] oder eine anschauliche Anwendung wie die „Wortwarte“ (Lemnitzer, 2011) bereits Lust darauf, Sprachdaten zu analysieren. Andererseits dienen dazu auch Forschungsarbeiten, durchaus populärwissenschaftlich aufbereitet, wie die Analysen zu den vergangenen US- und Bundestagswahlen (Bubenhofer u. a., 2008, 2009), von denen ich als Forschungsgruppenmitglied aus erster Hand berichten konnte.

2.2 Beispiele studentischer Arbeiten

Im Folgenden berichte ich von einigen korpuslinguistischen Arbeiten von Studierenden, die ich betreute.[5] Es geht weniger darum, die genauen Inhalte zu referieren, sondern die eingesetzten Methoden und Werkzeuge zu erwähnen. Die Arbeiten sind naturgemäß unveröffentlicht.

David Papst nimmt sich in **„*klein* und *winzig*. Eine Korpusuntersuchung zur Synonymie“** (Papst, 2005) eine klassische Fragestellung der Semantik vor. Als Basis für die Untersuchung dieser Quasisynonyme dient ihm das DWDS (o. J.), aus dem er eine Zufallsauswahl von Belegen für die Lemmata extrahiert und manuell kategorisiert. Er verzichtet völlig auf Standardverfahren wie eine Kollokationsanalyse, kann die semantisch-funktionalen Differenzen der beiden Lemmata aber trotzdem gut beschreiben.

Technisch und theoretisch etwas avancierter ist eine Arbeit von Igor Matic: **„Konzeptuelle Metaphern der Wirtschaftskrise in der NZZ am Sonntag“** (Matic, 2009). Er stellt sich ein eigenes Korpus aus 53 Zeitungsartikeln zusammen und kategorisiert darin manuell Metaphern. Um die Zeitungsartikel zu durchsuchen verwendet er „AntConc“[6] (vgl. Abbildung 1), eine Konkordanzsoftware, die aber auch Kollokationen, n-Gramme und bei Korpusvergleichen Schlüsselwortanalysen vornehmen kann.

Paul Rauber benutzt in seiner diskurslinguistisch ausgerichteten Arbeit **„Intellektuelle im Diskurs. Zwischen Hybris und Machtkritik“** (Rauber, 2009) mehrere Korpora. Einerseits verwendet er ein eigenes Korpus von Zeitungsartikeln aus dem Schweizer „Tages-Anzeiger“, die er über eine Schlüsselwortsuche mit den Lemmata *Intellektueller/intellektuell* zusammenstellt und mit der Konkordanzsoftware „AntConc“ verwaltet. Als Referenzkorpora benutzt er die „Frankfurter Rundschau“ über die COSMAS II-Schnittstelle des DeReKo IDS (o. J.), das gesamte DWDS (o. J.) und die Kollokationsprofile von Wortschatz Leipzig (o. J.). In allen Korpora nutzt er die Funktionen zum

[3] Dabei spielt die eigene Forschung naturgemäß eine besonders wichtige Rolle, wie bei mir die Arbeit zu korpuslinguistischen Methoden in der Diskurs- und Kulturanalyse (Bubenhofer, 2009).

[4] Der Google Ngram Viewer erzielte einige Aufmerksamkeit durch die populärwissenschaftliche Lancierung des Forschungszweigs „Culturomics“ (Michel u. a., 2011; Lieberman u. a., 2007).

[5] Einen Teil der Arbeiten betreute ich zusammen mit Angelika Linke.

[6] Die Software wurde von Laurence Anthony entwickelt: `http://www.antlab.sci.waseda.ac.jp/antconc_index.html` (23. März 2011), Anthony (2010).

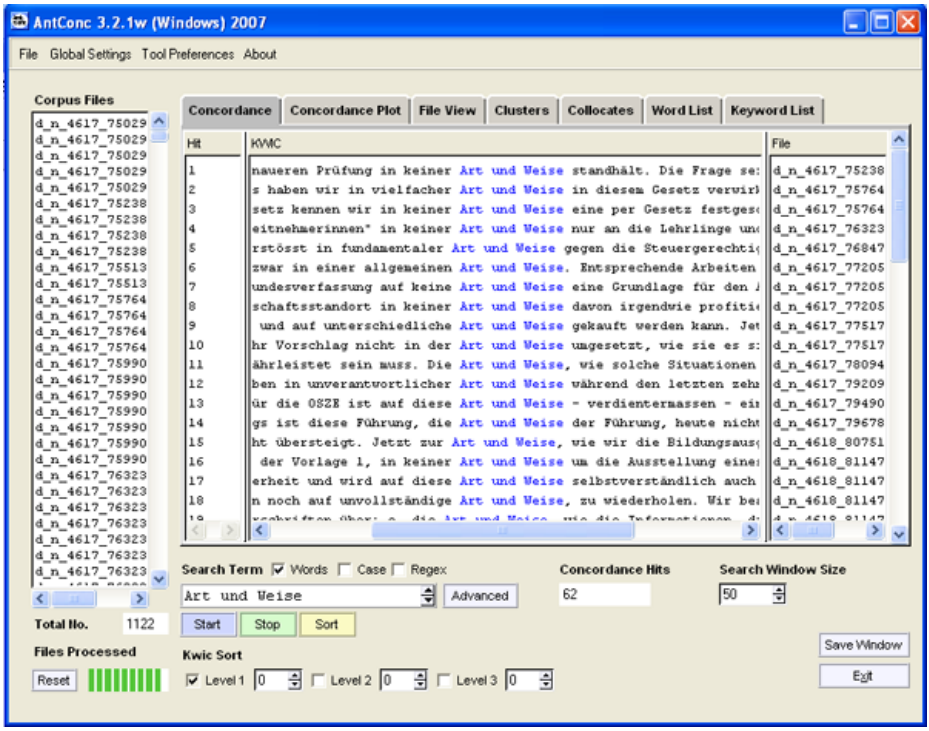

Abbildung 1: „AntConc“ ist eine einfach bedienbare Konkordanzsoftware, mit der auch Kollokationen, n-Gramme und Schlüsselwörter berechnet werden können (Anthony, 2010).

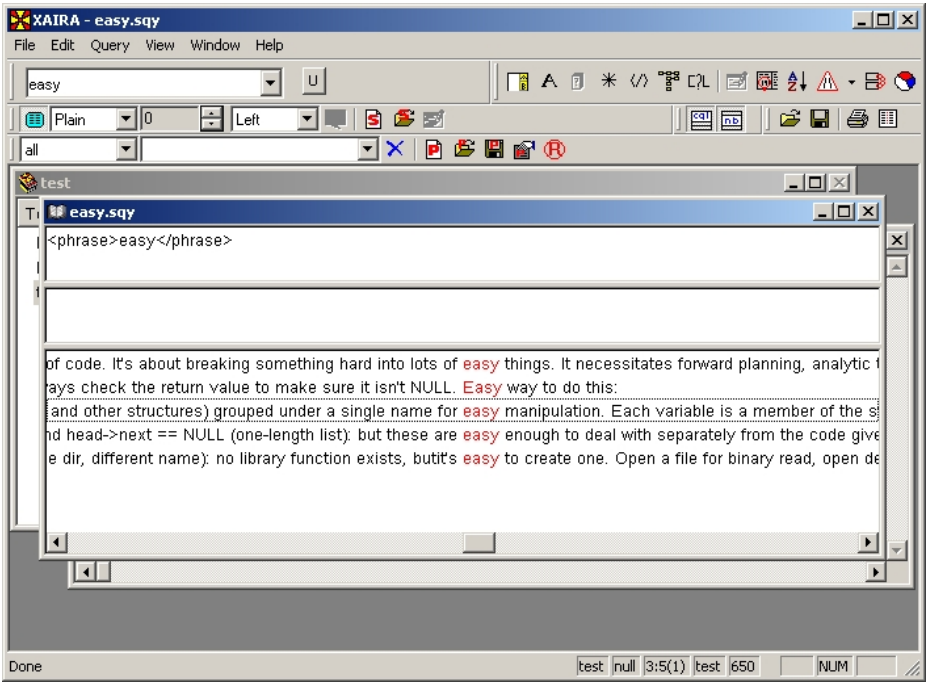

Abbildung 2: „Xaira“ ist eine Konkordanzsoftware, die XML-Dokumente verarbeiten kann (Oxford University Computing Services, 2011).

Berechnen von Kollokationen, um die Verwendungsweisen des Konzeptes *intellektuell* semantisch und diskurslinguistisch zu beschreiben.

In eine ähnliche Richtung zielt Verena Casanas Arbeit **„Homosexualität. Analyse der Paradigmengruppe *homosexuell* – *gleichgeschlechtlich* anhand der taz“** (Casana, 2009). Im selber anhand der Schlüsselwörter zusammengestellten Korpus der „tageszeitung“ von 1994–2008 untersucht sie mit „AntConc“ die Belege und Kollokationen zu den untersuchten Lemmata und deutet den Wandel der Verwendungsweise diskurslinguistisch.

Während die bisher beschriebenen Arbeiten mit relativ wenig Aufwand auskamen, um die Daten für die Analyse vorzubereiten, ist das in Tamara Weibels Arbeit anders: **„*Mieterinnen* oder *Mieter* – *Schweizervolk* oder *Ausländer*? Parteispezifische Personen- und Personengruppenbezeichnungen der SP und SVP im Schweizer Parlament“** (Weibel, 2009). Grundlage sind die Protokolle des Schweizer Parlaments, die allerdings von einem computerlinguistisch bewanderten Kommilitonen aufbereitet werden mussten, um die Redebeiträge zu extrahieren und den jeweiligen

Sprechern und Parteien zuordnen zu können. So aufbereitet nutzt die Autorin ebenfalls „AntConc“, um die darin vorhandene Funktion „Keywords“ zu verwenden, die beim Vergleich zweier Teilkorpora die jeweils statistisch signifikanten Wörter berechnet.[7] Die so als parteitypisch eruierten Lemmata werden dann durch Kollokationsanalysen und eine manuelle Kategorisierung semantisch-funktional gedeutet.

Auch Sara Baertschi bewegt sich mit ihrer Abschlussarbeit **„Der Berg ruft. Sprachgebrauchsmuster von 1920-1945 in der Literatur des Schweizer Alpen-Clubs“** (Baertschi, 2010) im Feld der Diskursanalyse. Basis ist ein Korpus alpinistischer Texte (Text+Berg-Korpus, 2011). Auch sie benutzt „AntConc“, um Belege für verschiedene Lemmata zu kategorisieren, Kollokationen dazu zu berechnen und Frequenzverteilungen in verschiedenen zeitlich definierten Teilkorpora auszugeben. Um die Relevanz von Frequenzunterschieden zu berechnen verwendet sie statistische Signifikanztests.

Als Studentin der Computerlinguistik kann Angela Fahrni in **„Regelmässigkeiten in Kundenrezensionen auf Amazon“** (Fahrni, 2008) auf avanciertere Methoden zurückgreifen. Mit der Hilfe von eigenen Perl-Scripten erstellt sie sich ein Korpus von 39.063 Kundenrezensionen von `amazon.de`, das sie im XML-Format, angereichert mit den verfügbaren Metadaten, speichert. Zudem setzt sie den „TreeTagger“ (Schmid, 1994) ein, um die Daten mit Part-of-Speech-Tags zu taggen. Den Tagger ergänzt sie um eigene Wortklassen, um Emoticons zu annotieren. Für die Analyse verwendet sie die Konkordanz- und Recherchesoftware „XAIRA“ (Oxford University Computing Services 2011, vgl. Abbildung 2), die XML-Dokumente verarbeiten kann. Zudem benutzt sie „gCLUTO – Graphical Clustering Toolkit“ (Rasmussen/Karypis, 2004), um zu berechnen, welche sprachlichen Marker positive und negative Rezensionen gut voraussagen können.

Das **„Wörterbuch der Krise“** (Baumgärtner u. a., 2010)[8] ist ein kollaboratives Werk von den Studierenden Verena Baumgärtner, Sascha Braun, Barbara Katharina Dietz, Verena Keite, Maximilian Nowroth, Frederic Wagner und Johannes Wolf.[9] Ziel war, mit korpuslinguistischen Mitteln drei unterschiedliche Krisen-Diskurse lexikographisch anzugehen. Grundlage bilden Recherchen in öffentlichen Korpora (DeReKo IDS, o. J.; DWDS, o. J.), aber auch eigens zusammengestellte Zeitungskorpora, die vor allem mit Kollokationsanalysen bearbeitet wurden.

Eine besondere Datengrundlage verwendet Madeleine Ehrensperger für ihre Arbeit **„Geschlechts- und Altersspezifisches Sprachverhalten“** (Ehrensperger, 2006): Mittels eines Fragebogens mit politischen/gesellschaftlichen Einstellungsfragen erhält sie von 60 Versuchspersonen beiderlei Geschlechts und unterschiedlichen Alters schriftlich geäußerte Statements. Diese untersucht sie hinsichtlich der Verwendung der linguistischen Parameter Satzlänge, Ich-Aussagen, Satzklammern, Abkürzungen und Ausrufe-

[7] „AntConc“ kann keine annotierten Korpora verarbeiten. Im Fall der beschriebenen Arbeit war es aber möglich, die gewünschten Redebeiträge aus dem annotierten Korpus zu extrahieren und als nicht-annotierte Textdateien in AntConc zu laden.

[8] Online erreichbar über `http://www.bubenhofer.com/Krise/` (23. März 2011).

[9] Mitbetreuerin dieser Arbeiten war Stefaniya Ptashnyk.

und Fragezeichen, um herauszufinden, ob bezüglich dieser Parameter ein geschlechtsspezifisches Schreiben beobachtbar ist. Sie nutzt keinerlei technische Hilfsmittel, obwohl das nahegelegen hätte.

2.3 Hoffnungen und Enttäuschungen

Als Betreuer der oben kurz dargestellten Arbeiten machte ich die Erfahrung, dass die Studierenden es generell schätzten, empirisch arbeiten zu können. Ebenso attraktiv scheint es zu sein, dass diese Arbeiten sehr anwendungsbezogen und nicht primär eine theoretische Auseinandersetzung sind. Die Korpuslinguistik kann zudem bis zu einem gewissen Grad das Bedürfnis stillen, mehr oder weniger klar definierte Methoden für die Analyse von Sprachdaten anwenden zu können.

Doch diesen positiven Aspekten müssen auch Enttäuschungen der Studierenden gegenüber gestellt werden. In erster Linie treten diese ein, wenn der große technische Aufwand deutlich wird. Empirisches Arbeiten ist aufwändig und wenn gleichzeitig noch die technischen Fertigkeiten erlangt werden müssen, um die Werkzeuge anwenden zu können, kann dies zu Frustgefühlen führen.

Bei einigen Arbeiten stellt sich zudem das Problem der Operationalisierung der Hypothesen, die vor dem Hintergrund linguistischer Theoriebildung, gerade im Bereich der Diskurs- und Kulturanalyse, sehr schwierig ist. Damit verbunden ist dann die Enttäuschung über die (vermeintlich) beschränkte Aussagekraft von korpuslinguistischen Analyseresultaten. N-Gramm-Analysen oder Kollokationsberechnungen führen rasch zu großen Datenmengen, die gesichtet und kategorisiert werden müssen; die Analyse auf Knopfdruck ist eine Utopie.

3 Nutzerprofil Online-Kurs Korpuslinguistik

Meine Website „Einführung in die Korpuslinguistik. Praktische Grundlagen und Werkzeuge“ (Bubenhofer, 2006-2011) entstand im Rahmen meiner gleichnamigen Veranstaltung an der Universität Zürich (vgl. Abbildung 3). Seit 2006 gab es immer wieder kleinere Aktualisierungen und Erweiterungen, so dass das Angebot inzwischen die in Tabelle 2 dargestellten Themen umfasst.

Gemäß Zugriffsstatistik gab es im Jahr 2010 14.158 Besuche mit 43.319 Seitenaufrufen, wobei 34% über direkte Zugriffe[10], 21% über Verweise und 45% über Suchmaschinen zustande gekommen sind. Aus der Analyse der Verweise lässt sich schließen, dass das Angebot auch an einigen Universitäten in der Lehre benutzt wird.

Die Abbildungen 4 und 5 zeigen die am häufigsten und am längsten aufgerufenen Seiten der Einführung. Nicht weiter überraschend liegen grundlegende Themen wie „Einführung“, „Definitionen“ oder „Korpustypen“ ganz oben. Auch die Ausführungen zu „DeReKo/COSMAS II“ werden stark nachgefragt.

[10] Die Adresse wurde also direkt in den Browser eingegeben oder über ein gespeichertes Lesezeichen abgerufen.

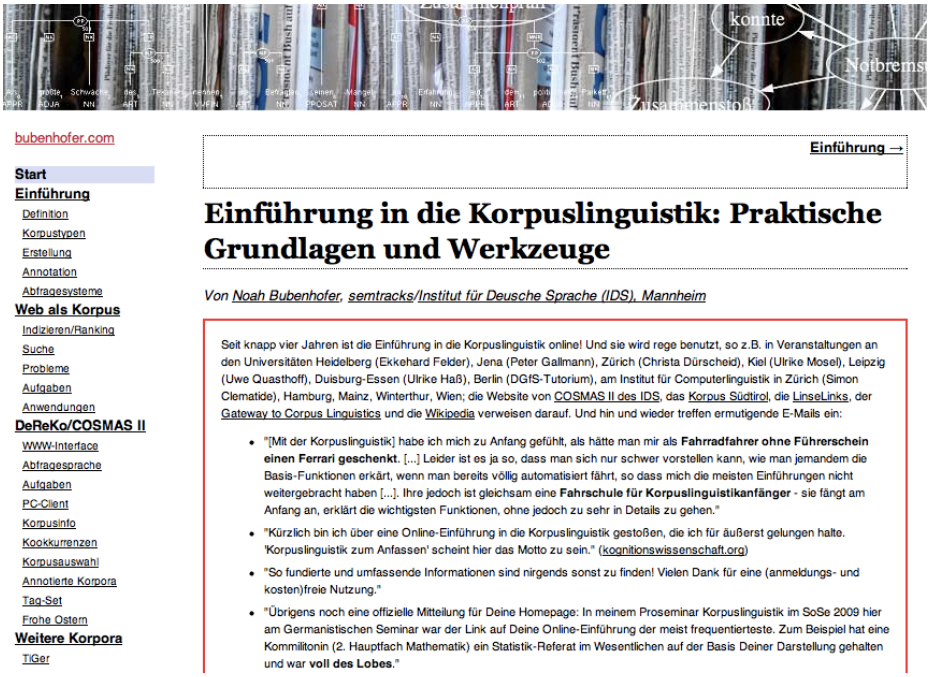

Abbildung 3: Die Startseite der „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011), erreichbar unter `www.bubenhofer.com/korpuslinguistik/`.

1. **Einführung:** Definition Korpuslinguistik, Korpustypen, Erstellung von Korpora, Annotation, Abfragesysteme
2. **Web als Korpus:** Funktionsweise von Suchmaschinen, Suchmöglichkeiten, Probleme, Anwendungen
3. **DeReKo/COSMAS II:** Informationen zur Funktions- und Verwendungsweise des DeReKo IDS (o. J.)
4. **Weitere Korpora:** TiGer (Lezius, 2002; Brants u. a., 2002), DWDS (o. J.), Wortschatz Leipzig (o. J.), Archiv für gesprochenes Deutsch des Instituts für Deutsche Sprache (Haas/Wagener, 1992)
5. **Eigenes Korpus:** Daten beschaffen, aufbereiten, analysieren, „AntConc“ (Anthony, 2010), „kfNgram“ (Fletcher, 2010)
6. **Corpus Workbench:** Die Arbeit mit der Open Corpus Workbench (CWB, 2011)
7. **Datenbank Filemaker:** Einführung, Datenbank erstellen, durchsuchen, Daten importieren, CSV-Formatierung, exportieren
8. **Anwendungen:** Forschungsprozess, Semantik, Argumentationsmuster, Diskursanalyse
9. **Statistik:** Einführung, Signifikanztests
10. **Visualisierung:** Möglichkeiten, GraphViz (Ellson u. a., 2011), Beispiele
11. **Anhang:** Software, Unix-Befehle, Reguläre Ausdrücke, Literatur, Lexikon, Impressum

Tabelle 2: Inhalt der „Einführung in die Korpuslinguistik“ (Bubenhofer, 2006-2011).

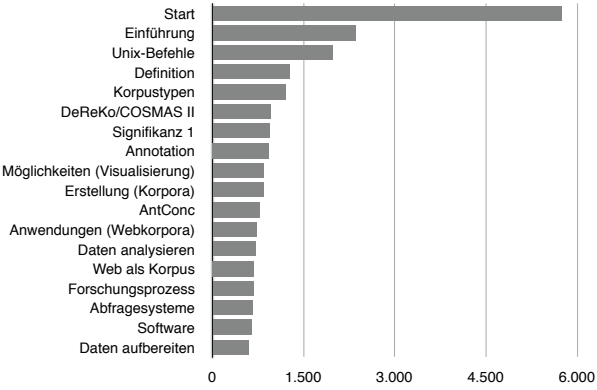

Abbildung 4: Die im Jahr 2010 am häufigsten aufgerufenen Seiten der „Einführung in die Korpuslinguistik“ (Anzahl Seitenaufrufe).

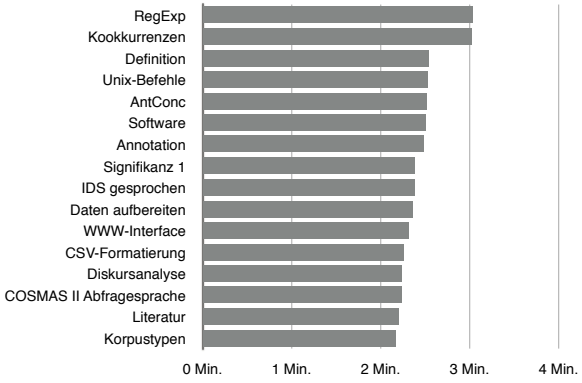

Abbildung 5: Die im Jahr 2010 am längsten aufgerufenen Seiten der „Einführung in die Korpuslinguistik“ (durchschnittliche Anzahl Minuten).

Die Seiten „Reguläre Ausdrücke“ und „Kookkurrenzen“ führen zu den höchsten Verweildauern. Ebenso weitere textlastige Ausführungen zu Software, Annotation, Statistik etc. In beiden Abbildungen ist zudem sichtbar, dass die Seite zu den Unix-Befehlen stark nachgefragt wird. Es handelt sich hierbei um eine Sammlung von grundlegenden Unix-Befehlen, die für korpuslinguistische Fragestellungen nützlich sind. Allerdings muss hierbei betont werden, dass sowohl diese Seite als auch jene zu den regulären Ausdrücken wohl nicht nur von Besuchern nachgefragt werden, die ein korpuslinguistisches Interesse mitbringen, sondern vor allem über Suchmaschinen Besucher anziehen, die aus anderen Gründen danach suchen.

Da 45% der Besuche über Suchmaschinen zugeführt werden, analysierte ich zusätzlich die Suchbegriffe, mit denen die Besucher auf die Seite gelangen (Tabelle 3). Hierbei muss natürlich beachtet werden, dass nur Suchbegriffe angezeigt werden, mit denen man auf der Website fündig werden kann. Die Liste kann also nicht aufzeigen, welche Themen im Angebot fehlen, sondern nur Hinweise darauf geben, welche Aspekte der bestehenden Inhalte auf besonders großes Interesse stoßen.

Es wird ersichtlich, dass oft nach „Themen“ oder „Anwendungsgebieten“ der Korpuslinguistik gesucht wird. Es scheint also wichtig zu sein, Anregungen für korpuslinguistisches Arbeiten zu geben. Weiter gibt es diverse Suchbegriffe, die die Korpuserstellung zum Thema haben oder gezielt Informationen zu bestimmten bestehenden Korpora suchen.

Natürlich sind auch Hinweise auf und Anleitungen von Programmen für korpuslinguistische Arbeiten gefragt. Solche Anleitungen scheinen auch dann ein Bedürfnis zu sein, wenn der Softwareanbieter eigentlich eine Dokumentation anbietet, wie z. B. bei „AntConc“, wo die offizielle Anleitung allerdings nur auf Englisch zur Verfügung steht und relativ knapp gehalten ist. Auch der Erklärungsbedarf von Methoden wie Kookkurrenzanalyse oder statistische Signifikanztests scheint groß zu sein.

Zusammenfassend kann gesagt werden, dass die Benutzer der Website einerseits Werkzeuge suchen für die Recherche in Korpora, das Erstellen und Verwalten von Korpora, die statistische Analyse und die Annotation, andererseits aber auch nach Hilfen und Anleitungen zur Bedienung dieser Werkzeuge. Und natürlich werden Inspirationsquellen gewünscht, die korpuslinguistisches Arbeiten zeigen.

4 Fazit: Möglichkeiten und Wünsche

Ich habe versucht zu zeigen, welche Inhalte eine Einführung in die Korpuslinguistik in der linguistischen Lehre umfassen kann und was für Projekte realistisch sind, in denen die Studierenden das Gelernte anwenden können. Zudem hat die Analyse des Web-Angebots meiner Einführung in die Korpuslinguistik aufgezeigt, welche Themen besonders beliebt sind.

Aus diesen Erfahrungen heraus kann ich nun als Lehrperson einige Wünsche formulieren, die in erster Linie das Angebot an korpuslinguistischem Werkzeug betrifft. Für die Lehre wäre es ein Desiderat, über Software-Module für unterschiedliche Anwendungen zu verfügen:

Inhalte	
themen korpuslinguistik	130
anwendungsgebiete korpuslinguistik	26
diskursanalyse	21
korpuslinguistik diskursanaylse	11
probleme der korpuslinguistik	5
Korpuserstellung/Korpora	
korpus definition	36
korpuslinguistik tageszeitungen	43
korpus erstellen	34
daten aufbereiten	5
erstellung ein korpus	5
wie erstelle ich einen korpus	5
textkorpus erstellen download	12
filemaker datenbank erstellen	9
korpuslinguistik copyright	25
deutschsprachige korpora	6
korpustyp	6
baumbanken	5
tiger corpus	41
cosmas ii	26
ids korpus	9
funktionen cosmas	6
dereko	5
dwds	15
lexis nexis korpus	7
Annotation	
annotation korpuslinguistik	23
annotierte korpora	12
korpuslinguistik tagging	5
pos tagger online	7
tagset	54
korpulinguistik tag sets	5
dependenz parser	5

Software	
konkordanzprogramm	33
konkordanzprogramm download	5
concordance-programme zur analyse von korpora	5
korpuslinguistik tools	7
simple concordance program	12
korpuslinguistik software	21
textdatei importieren per script filemaker	13
antconc	277
antconc anleitung	7
antconc regex	6
cluster antconc	5
t-score antconc	5
graphviz	12
graphviz beispiele	7
graphviz dot	7
graphviz gui	5
kfngram	12
filemaker	8
tigersearch	14
corpus workbench windows	7
treetagger betriebssystem	15
regexp	5
reguläre ausdrücke antconc	5
software berechnung signifikanz	5
Statistik	
kookkurrenzen	45
kookkurrenzanalyse	15
kookkurrenzprofil	14
log likelihood test	30
log-likelihood	10
llr wert	15
log likelihood tabelle	6
chi quadrat test signifikant	5
signifikanz	8
signifikanztest excel	8
kontingenztabelle signifikanz	13
darstellungsoptionen konkordanz korpuslinguistik	8

Tabelle 3: Suchbegriffe von Suchmaschinen, die zur „Einführung in die Korpuslinguistik“ führen; grob geordnet nach thematischen Bereichen (zweite Spalte: Anzahl Suchende).

- Korpuserstellung (Textaufbereitung, Web-Download etc.), Verwaltung, Annotation, Analyse und Ergebnisdarstellung.
- Die Softwaremodule sollten über einheitliche Schnittstellen verfügen, so dass man die Daten und Ergebnisse leicht mit unterschiedlichen Modulen weiterverarbeiten kann.
- Die Module sollten möglichst plattformunabhängig sein, denn die verfügbare Infrastruktur, sei es in PC-Pools an der Universität, seien es die privaten Rechner der Studierenden, ist bezüglich Betriebssystemen sehr heterogen.[11]
- Wichtig ist die einfache Bedienbarkeit über eine leicht verständliche grafische Benutzeroberfläche. Nur so können Studierende angesprochen werden, die keine große Affinität zu Computern haben.

Zum wichtigsten Ziel von korpuslinguistischen Kursen zähle ich, die Studierenden überhaupt dazu zu motivieren, korpuslinguistisch, also empirisch zu arbeiten. Dabei ist es wichtig, die Angst vor den technischen Hürden zu nehmen und die Studierenden dazu zu ermutigen, auch Methoden zu verwenden, die auf den ersten Blick kompliziert wirken. Dazu gehört z. B. auch der Einsatz von statistischen Methoden, etwa zur Überprüfung von Signifikanzunterschieden.

Dabei bewegt man sich unweigerlich auf dem schmalen Grat des Realistischen und potenziell Möglichen: Die State-of-the-Art der Korpuslinguistik ist immer fortgeschrittener als das, was in der linguistischen Lehre tatsächlich gemacht werden kann. Mit anschaulichen Beispielen, die man auf einfach bearbeitbare Teilaufgaben herunterbricht, können die Studierenden jedoch motiviert werden, sich an den Stand der Kunst heranzutasten.

Doch bei allen technischen Hürden schien es mir immer wichtig, zunächst von linguistisch fundierten Hypothesen auszugehen, diese zu operationalisieren und erst dann zu prüfen, ob das Vorhaben technisch umgesetzt werden kann. Es wäre schade, wenn man sich schon gleich zu Beginn vom technischen Aufwand abschrecken ließe.

Literatur

Anthony, Laurence (2010): *AntConc 3.2.1* <`http://www.antlab.sci.waseda.ac.jp/`>.

Baertschi, Sara (2010): Der Berg ruft. Sprachgebrauchsmuster von 1920-1945 in der Literatur des Schweizer Alpen-Clubs [Unveröffentlichte Lizentiatsarbeit, Deutsches Seminar, Universität Zürich].

Baumgärtner, Verena/Braun, Sascha/Dietz, Barbara Katharina/Keite, Verena/Nowroth, Maximilian/Wagner, Frederic/Wolf, Johannes (2010): *Wörterbuch der Krise* <`http://www.bubenhofer.com/Krise/`> [betreut von Stefaniya Ptashnyk und Noah Bubenhofer].

[11] Zwar sind Windows-Rechner sehr verbreitet, besonders an Schweizer Universitäten sind jedoch auch viele Mac-Systeme in Verwendung und in Computerlinguistik-Kontexten sind oft Linux-Systeme im Einsatz.

Brants, Sabine/Dipper, Stefanie/Hansen, Silvia/Lezius, Wolfgang/Smith, George (2002): The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Bubenhofer, Noah (2006-2011): *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge*. Elektronische Ressource <`http://www.bubenhofer.com/korpuslinguistik/`>.

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen; 4).

Bubenhofer, Noah/Dussa, Tobias/Ebling, Sarah/Klimke, Martin/Rothenhäusler, Klaus/Scharloth, Joachim/Tamekue, Suarès/Vola, Saskia/Forschergruppe semtracks (2009): „So etwas wie eine Botschaft". Korpuslinguistische Analysen der Bundestagswahl 2009. In: *Sprachreport* 4, S. 2–10.

Bubenhofer, Noah/Klimke, Martin/Scharloth, Joachim (2008): *political tracker – U.S. Presidential Campaign '08: A Semantic Matrix Analysis*. Elektronische Ressource <`http://semtracks.com/politicaltracker/`>.

Casana, Verena (2009): Homosexualität. Analyse der Paradigmengruppe *homosexuell – gleichgeschlechtlich* anhand der taz [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

CWB (2011): *The IMS Open Corpus Workbench (CWB)* <`http://cwb.sourceforge.net/`>.

DeReKo IDS (o. J.): *Das Deutsche Referenzkorpus DeReKo*. Elektronische Ressource <`http://www.ids-mannheim.de/kl/projekte/korpora/`>.

DWDS (o. J.): *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts*. Elektronische Ressource <`http://www.dwds.de`>.

Ehrensperger, Madeleine (2006): Geschlechts- und Altersspezifisches Sprachverhalten [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Ellson, John/Gansner, Emden/Hu, Yifan/Bilgin, Arif (2011): *Graphviz – Graph Visualization Software* <`http://www.graphviz.org`>.

Fahrni, Angela (2008): Regelmässigkeiten in Kundenrezensionen auf Amazon [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Fletcher, William H. (2010): *kfNgram: Information and Help* <`http://www.kwicfinder.com/kfNgram/`>.

Google Books Ngram Viewer (o. J.): *Google Books Ngram Viewer*. Elektronische Ressource <`http://ngrams.googlelabs.com/`>.

Haas, Walter/Wagener, Peter (Hgg.) (1992): *Gesamtkatalog der Tonaufnahmen des Deutschen Spracharchivs. Erarbeitet von Mitarbeiterinnen und Mitarbeitern des Instituts für Deutsche Sprache*. Tübingen: Niemeyer (Phonai; 38/39).

Lemnitzer, Lothar (2011): *Die Wortwarte* <`http://www.wortwarte.de`>.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Lezius, Wolfgang (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Phil. Diss. University of Stuttgart, Stuttgart <http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/disslezius.pdf>.

Lieberman, Erez/Michel, Jean-Baptiste/Jackson, Joe/Tang, Tina/Nowak, Martin A. (2007): Quantifying the evolutionary dynamics of language. In: *Nature* 449, S. 713–716.

Matic, Igor (2009): Konzeptuelle Metaphern der Wirtschaftskrise in der NZZ am Sonntag [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Michel, Jean-Baptiste/Shen, Yuan Kui/Aiden, Aviva Presser/Veres, Adrian/Gray, Matthew K./Team, The Google Books/Pickett, Joseph P./Hoiberg, Dale/Clancy, Dan/Norvig, Peter/Orwant, Jon/Pinker, Steven/Nowak, Martin A./Aiden, Erez Lieberman (2011): Quantitative Analysis of Culture Using Millions of Digitized Books. In: *Science* 331, H. 6014, S. 176–182 <http://www.sciencemag.org/content/331/6014/176.abstract>.

Oxford University Computing Services (2011): *All About Xaira* <http://www.oucs.ox.ac.uk/rts/xaira/>.

Papst, David (2005): *klein* und *winzig*. Eine Korpusuntersuchung zur Synonymie [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Rasmussen, Matt/Karypis, George (2004): gCLUTO: An Interactive Clustering, Visualization, and Analysis System. *Techn. Ber. 04-021*, University of Minnesota Technical Report <http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview>.

Rauber, Paul (2009): Intellektuelle im Diskurs. Zwischen Hybris und Machtkritik [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg: Winter (Kurze Einführungen in die Germanistische Linguistik; 2).

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees* <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.

Text+Berg-Korpus (2011): *Text+Berg-Korpus (Release 145)*. XML-Format [Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-2009].

Weibel, Tamara (2009): *Mieterinnen* oder *Mieter – Schweizervolk* oder *Ausländer*? Parteispezifische Personen- und Personengruppenbezeichnungen der SP und SVP im Schweizer Parlament [Unveröffentlichte Hausarbeit, Deutsches Seminar, Universität Zürich].

Wortschatz Leipzig (o. J.): *Deutscher Wortschatz Universität Leipzig*. Elektronische Ressource <http://wortschatz.uni-leipzig.de>.