Angelika Storrer

Online-Corpora zur linguistischen Analyse der deutschen Gegenwartssprache

An digitalen Textsammlungen, die im Internet oder auf CD-Rom verfügbar sind, gibt es mittlerweile keinen Mangel mehr. Das Projekt Gutenberg, in seiner deutschen Fassung bei *Spiegel Online* oder auch als DVD erhältlich (http://gutenberg.spiegel.de/info/info.htm), umfasst mittlerweile ca. 420.000 Textseiten. Die Firma *Directmedia* publiziert in der Reihe „Die digitale Bibliothek" seit den 90er-Jahren digitale Versionen von wichtigen Werken aus den Sozial- und Geisteswissenschaften, in denen man mit einem schnellen und einfach erlernbaren Suchwerkzeug recherchieren kann (http://www.digitale-bibliothek.de/). Auch viele Zeitungs- und Zeitschriftenverlage bieten online recherchierbare Archive ihrer Texte kostenlos oder kostenpflichtig zur Recherche an, teilweise auch verschlagwortet und mit anspruchvollen Suchoptionen. All diese Textsammlungen sind allerdings in erster Linie für die inhaltliche Erschließung und Dokumentation gedacht; für die sprachwissenschaftliche Analyse des Deutschen sind sie nur bedingt geeignet. Wer Belege zum verbalen Präfix „an+" sucht, möchte Formen wie „andere", „angeln" und „androgyn" ausschließen. Wer Belege zum Lexem „fahren" sucht, möchte meist auch Belege zu „fuhr" und „fährt" finden, nicht aber Formen wie „fährt (…) an", oder „fuhr (…) ab" oder homographe Formen des Verbs „führen". Wer Belege zum Verb „schicken" sucht, ist nicht an Fundstellen wie „eine schicke Bluse" interessiert. Um Lösungen für die speziellen Anforderungen der linguistischen Textanalyse bemühen sich die als „Corpuslinguistik"[1] und „Texttechnologie"[2] bezeichneten sehr aktiven Forschungsbereiche. Sie entwickeln in interdisziplinärer Zusammenarbeit mit Informatik, Computerlinguistik und Linguistik Methoden, Standards und Werkzeuge für den Aufbau und die Erschließung von Corpora gesprochener und geschriebener Sprache, die als empirische Basis für die Theoriebildung und die Überprüfung theoretischer Annahmen an authentischen Textbelegen herangezogen werden können. Die Formen der linguistischen Annotation, wie die Anreicherung von Textdaten mit Metadaten zur Explikation linguistischer Merkmale genannt wird, reichen von der Rückführung flektierter Formen auf Grundformen („Lemmatisierung" genannt) über die Morphemzerlegung und die Zuordnung von Wörtern zu syntaktischen Kategorien („Part-of-Spech-Tagging" genannt) bis zur partiellen oder vollständigen Analyse syntaktischer Strukturen von Sätzen, die in der Form sog. Baumbanken (Treebanks) über spezialisierte Such- und Recherchewerkzeuge zugänglich gemacht werden.

Der Umgang mit linguistisch aufbereiteten Corpora erfordert technische Kenntnisse und eine methodologische Vorbildung, die es ermöglicht, den Stellenwert von Belegen und die Signifikanz statistischer Berechnungen in Bezug auf eine Fragestellung richtig einzuschätzen. Es wird zunehmend eine Aufgabe der universitären Linguistikausbildung sein, Studierende an den Umgang mit Sprachcorpora heranzuführen und ihnen dabei die erforderlichen Werkzeug- und Methodenkenntnisse zu vermitteln. Hierzu müssen nicht unbedingt spezielle Veranstaltungen angeboten werden. Vielmehr lassen sich bei entsprechender technischer Ausstattung Corpusrecherchen als aktivierende Komponenten in Seminare einbinden, etwa indem Studierende kleine Untersuchungen zu Regularitäten in den Bereichen Phraseologie, Morphologie/Wortbildung, Wortschatzentwicklung und Sprachkontakt anstellen oder Merkmale gesprochener Sprache untersuchen. Im Folgenden möchte ich Ressourcen und Werkzeuge für Corpora zur deutschen Gegenwartssprache vorstellen, die kostenfrei und direkt über das Internet verfügbar sind, und die auch ohne computerlinguistische Vorbildung mit einer akzeptablen Einarbeitungszeit im Bereich der germanistischen Linguistik und der Auslandsgermanistik für die Forschung und die Lehre nutzbar gemacht werden können. Der besseren Lesbarkeit halber sind sämtliche Verweise auf WWW-Adressen in den Anhang ausgelagert.

Eine langjährige Tradition in der Zusammenstellung, Digitalisierung und linguistischen Erschließung deutscher Corpora hat das Institut für deutsche Sprache in Mannheim. Entsprechend finden sich dort zwei wichtige Corpussammlungen: Die Datenbank „Gesprochenes Deutsch" und die Corpus-Sammlung geschriebener Gegenwartssprache. In beiden Corpora kann man online recherchieren; beide Corpora bieten vielfältige Optionen zur linguistischen Recherche und Weiterverarbeitung der Ergebnisse. Die Datenbank

---

1 Einführend z.B. McEnery, T. / Wilson, A. (2001): Corpus Lingustics. Edinburgh University Press.

2 Einführend z.B. Lobin, H. / Lemnitzer, L. (2004): Texttechnologie. Perspektiven und Anwendungen. Stauffenburg.

„Gesprochenes Deutsch“ ist, wenn bestimmte technische Voraussetzungen erfüllt sind, mit Hilfe eines Web-Browsers nutzbar, in denen die vielfältigen Abfrageoptionen verständlich erklärt und anhand von Beispielabfragen veranschaulicht sind. Interessant für die Lehre ist die Alignierung (digital gespeicherte Zuordnung) von Transkripten und Tondateien, die den Zugang zu Transkripten gerade für Studierende erleichtert. Bislang sind nur Ausschnitte aus dem insgesamt am IDS verfügbaren Material öffentlich und online zugänglich; diese machen jedoch das Recherche-Potenzial der kontinuierlich weiter entwickelten Ressource deutlich und liefern genügend Material für kleinere Untersuchungen zu Merkmalen gesprochener Sprache im Rahmen eines einführenden Seminars. Weiterhin verfügt das IDS über die (nach eigenen Angaben) weltweit größte Sammlung von Corpora geschriebener deutscher Gegenwartssprache (ca. 2 Milliarden Textwörter). Der Zugriff darauf wird realisiert über eine spezielle Client-Software, *COSMAS-II*, die heruntergeladen und auf dem Rechner installiert werden kann. Wem dies gelingt, dem steht ein mächtiges Recherchewerkzeug mit vielfältigen Such- und Ausgabeoptionen auf die Mannheimer Corpora zur Verfügung – etwas Einarbeitungszeit in die Nutzung sollte man allerdings veranschlagen; deutsche Hilfetexte dazu sind online und zum Download verfügbar. Leider läuft *COSMAS-II* bislang nur mit Windows-basierten Betriebsystemen. Zum Frühjahr 2005 ist ein Werkzeug zur direkten Online-Recherche über einen Web-Browser (wie im Vorgänger-System *COSMAS-I* oder im unten beschriebenen *DWDS*-Corpus) angekündigt. Auf diese Schnittstelle werden sich Mac-Nutzer und Nutzer von Unix-basierten Systemen ebenso freuen, wie all diejenigen, die bislang mit der Client-Software nicht zurecht gekommen sind.

Wer direkt über seinen Web-Browser in einem Corpus der deutschen Gegenwartssprache recherchieren möchte, kann die Corpora nutzen, die im Zuge eines Projekts zur Erstellung eines digitalen Wörterbuchs zur deutschen Sprache des 20. Jahrhunderts an der Berlin-Brandenburgischen Akademie der Wissenschaften aufgebaut wurden. Das dort verfügbare Kerncorpus (ca. 100 Mio. Textwörter) ist außerdem nach dem Prinzip der Ausgewogenheit zusammengestellt. Das bedeutet: Die für das Corpus ausgewählten Texte sind gleichmäßig über Dekaden (im Zeitraum von 1900 bis 2000) und Textsorten (Belletristik, Fachtextprosa und Zeitungstexte) gestreut. Deshalb eignet sich dieses Corpus besonders gut für Untersuchungen zur Wortschatzentwicklung und zur Textsortenspezifik. In der „Schnupperversion“, die ohne Registrierung unmittelbar genutzt werden kann, sind allerdings nur Texte im Zeitraum von 1900 bis 1945 mit eingeschränkten Kontexten zugänglich. Die Version für registrierte Benutzer hat eine deutlich erweiterte Funktionalität und ermöglicht den Zugriff auf das gesamte Corpus über eine schnell erlernbare Nutzerschnittstelle. Für Forschungszwecke würde man sich noch bessere Möglichkeiten für das Ausdrucken und Weiterverarbeiten wünschen. Attraktiv für den Einsatz in der Lehre sind der unkomplizierte und robuste Zugang, die Visualisierung der Vorkommenshäufigkeiten zu einem Suchwort über die Dekaden hinweg und – zu diesem Zweck wurde das Corpus schließlich entwickelt – die Verlinkung der Corpusdaten mit dem digitalen Wörterbuch[3]. Dieses hat als Kernbestand die digital aufbereitete Version des *WDG* (Wörterbuch der deutschen Gegenwartssprache), dessen Artikel mit Corpusbelegen und anderen lexikologischen Informationen verknüpft sind. Diese Verlinkungsstrukturen bieten Ansatzpunkte für Recherchen zur Wortbildung oder zum Bedeutungswandel und liefern Anregungen für das Verfassen hypertextueller Wörterbuchartikel, z.B. als Übungen in Seminaren zur (computergestützten) Lexikographie.

Etwas mehr Zeit für die Einarbeitung muss man einplanen, wenn man in syntaktisch annotierten Corpora recherchieren möchte; schließlich muss man sich dafür nicht nur in das Suchwerkzeug einarbeiten, sondern auch in die Kategorien, die im jeweiligen Corpus für die syntaktische Annotation genutzt werden. Dafür eröffnen sich für die Grammatikforschung und die Lexikographie aber auch ganz neue, sehr detaillierte Recherchemöglichkeiten. Mit dem Werkzeug *TiGerSearch*, das am Institut für maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart entwickelt wurde, steht mittlerweile ein intuitiv bedienbares, gut dokumentiertes und ansprechend gestaltetes Werkzeug zu Verfügung, das für wissenschaftliche Zwecke kostenfrei auf verschiedenen Plattformen installiert werden kann (Windows, MacOS, Linux u.a.). Die Syntax der symbolischen Abfragesprache bietet sehr komplexe Suchoptionen; für Einsteiger und Nutzer ohne computerlinguistische Vorbildung bietet *TiGerSearch* aber auch eine graphische Abfragesprache, in der sich mit einfachen Abfragen an einem Beispielcorpus das Prinzip der Suche in Baumbanken erlernen lässt;

[3] Direkter Zugriff auf das Wörterbuch unter http://www.dwds.de/pages/pages_woebu/dwds_woebu.htm.

hierbei können Studierende auch gleichzeitig ihre Kenntnisse grammatischer Strukturen und Kategorien auffrischen bzw. erweitern. Für die deutsche Gegenwartssprache stehen verschiedene Baumbanken zur Verfügung: Die an der Universität des Saarlandes aufbaute *NEGR@*-Baumbank verfügt in ihrer aktuellen, zweiten Version über ca. 20.000 annotierte Sätze, die semi-automatisch erstellt und intellektuell evaluiert wurden. Die an der Universität Tübingen entwickelte Baumbank *TüBa/-D/Z* umfasst ca. 15.000 Sätze und berücksichtigt neben der Konstituentenstruktur und den grammatischen Funktionen auch topologische Felder. Die am IMS der Universität Stuttgart erstellte *TiGer*-Treebank umfasst ca. 40.000 Sätze und eignet sich wegen der engen Verbindung zum *TiGerSeach*-Werkzeug (ein Pröbchen dieses Corpus ist dem Werkzeug beigefügt) besonders gut dazu, den Umgang mit dem Werkzeug einzuüben und sich das Potenzial der Recherche in Baumbanken für die Sprachforschung zu erschließen. Es ist aber gerade eine Stärke von *TiGerSearch*, dass auch die Formate von *NEGR@* und *TüBa/-D/Z*, sowie andere Baumbank-Standards (z.B. das Format der englischen *PENN*-Treebank) unterstützt werden; die entsprechenden Corpora lassen sich problemlos mit *TiGerSearch* durchsuchen. Die drei genannten deutschen Baumbanken enthalten ausschließlich Zeitungstexte (TAZ, Frankfurter Rundschau) und sind allesamt für wissenschaftliche Zwecke kostenfrei erhältlich.

Die Beschreibungen in diesem Beitrag beziehen sich auf den Stand im Januar 2005. Da die meisten der genannten Projekte sehr aktiv sind, können sich bereits beim Erscheinen dieses ZGL-Heftes Veränderungen ergeben haben. Aktuelle Informationen finden sich unter den unten angegebenen Web-Adressen. Die von mir getroffene Auswahl beschränkte sich auf Corpora zur linguistischen Analyse der deutschen Gegenwartssprache. Wer sich für frühere Sprachstadien des Deutschen interessiert, der sei auf die Link-Sammlung des Projekts „Deutsch Diachron Digital“ verwiesen; dieses Kooperationsprojekt hat zum Ziel, ein digitales Referenzcorpus zur deutschen Sprache zu entwickeln, das von den Anfängen der Textüberlieferung bis zum 19. Jahrhundert reicht (Informationen: http://www.deutschdiachrondigital.designato.de). Wer sich für weitere Ressourcen zur Corpuslinguistik und für Corpora zu anderen Sprachen interessiert, findet stets aktuelle Informationen unter den unten aufgeführten Link-Sammlungen.

Web-Adressen (URLs) zu den genannten Online-Ressourcen:

- Überblick über die Corpusarbeit und die vorhandenen Daten am Institut für deutsche Sprache in Mannheim: http://www.ids-mannheim.de/kt/corpora.html
- Überblick über die Corpora zum gesprochenen Deutsch am Institut für deutsche Sprache in Mannheim (Deutsches Spracharchiv):
  http://dsav-oeff.ids-mannheim.de/DSAv/DSAVINFO.HTM
  http://www.ids-mannheim.de/kt/projekte/korpora/archiv.html
- Online-Zugang zur Datenbank „Gesprochenes Deutsch“ (GDG):
  http://dsav-oeff.ids-mannheim.de/DSAv/ZUGANG.HTM
- Überblick über die Corpora geschriebener Gegenwartssprache am Institut für deutsche Sprache in Mannheim: http://www.ids-mannheim.de/kt/projekte/korpora/
- Corpusrecherchewerkzeug *COSMAS-II*: http://www.ids-mannheim.de/cosmas2/
- Überblick über die Corpora an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) im Projekt *DWDS* (Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts, http://www.dwds.de):
  http://www.dwds.de/pages/pages_textba/dwds_textba.htm
- Zugänge zur Online-Schnittstelle des *DWDS*-Corpus an der BBAW:
  http://www.dwds.de/pages/pages_textba/dwds_textba_rech.htm
- Direkter Zugriff für angemeldete Nutzer: http://www.dwds.de/cgi-bin/rest/loginstart

Baumbanken:

- Das Recherchewerkzeug für Baumbanken TiGerSearch (IMS Stuttgart):
  http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
- Das *NEGR@*-Corpus (Computerlinguistik, Universität des Saarlandes):
  http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html

- Tübinger Baumbank des Deutschen Schriftsprache" (*TüBa-D/Z*, SfS Universität Tübingen): http://www.sfs.uni-tuebingen.de/de_tuebadz.shtml
- Das *TiGer*-Corpus (IMS Stuttgart): http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

Link-Sammlungen zu Corpora und zur Corpuslinguistik:

- Die *LINSE*-Rubrik zu Corpora und Corpuslinguistik enthält viele aktuelle Links zum Thema: http://www.linse.uni-essen.de/inlink/index.php?sid=793965326&t=sub_pages&cat=23
- Das *Linguistic Data Consortium* (*LDC*: http://www.ldc.upenn.edu/) organisiert die Sammlung, Verfügbarmachung und Entwicklung von Ressourcen (Daten, Werkzeuge, Standards) für Forschung, Entwicklung und Lehre in Linguistik und Sprachtechnologie. Die Liste der dort erhältlichen Corpora findet sich unter: http://www.ldc.upenn.edu/Catalog/byType.jsp
- Die *Evaluations and Language Resources Distribution Agency* (*ELDA*)**,** distribuiert Corpora und lexikalische Ressourcen mit Schwerpunkt auf europäischen Sprachen im Auftrag der übergreifenden Organisation *ELRA* (*European Language Resources Association*) und verfolgt dabei im europäischen Maßstab ähnliche Ziele wie das *LDC*. Der Katalog findet sich unter: http://www.elda.org/
- Eine Linkliste zu Baumbanken und Baumbankprojekten weltweit pflegt das IMS Stuttgart: http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml

Adresse der Verfasserin: Prof. Dr. Angelika Storrer, Universität Dortmund, Institut für deutsche Sprache und Literatur, D-44221 Dortmund, E-Mail: angelika.storrer@uni-dortmund.de