Korpusbasierte Wörterbucharbeit mit den Daten des Projekts *Deutscher Wortschatz*

Uwe Quasthoff (Leipzig)

Abstract

The corpus project *Deutscher Wortschatz* (German Vocabulary) at Leipzig University is collecting and processing textual data for 15 years. It now consists of approx. 2 billion running words in 160 million sentences. The dictionary is online available at *www.wortschatz.uni-leipzig.de* and, moreover, contains word co-occurrence data.

The pre-processing of the data used mainly language independent methods and were used for corpora in other languages, too.

The paper describes the production process for three dictionaries for which these corpus data were used: a thesaurus, a dictionary of neologisms, and a collocation dictionary. In all cases, the raw data for the dictionary entries were produced automatically, and the final entries were written only using these pre-selections. In the case of the thesaurus, the preprocessing consisted in a corpus based detection of semantically similar words. For the neologism dictionary the yearly frequency information were used and for the collocation dictionary, word co-occurrences and part of speech information were combined.

1 Einführung in die Korpusdaten

Im Leipziger Projekt *Deutscher Wortschatz* werden seit rund 15 Jahren elektronisch vorliegende Textdaten gesammelt und ausgewertet. Momentan liegen mehr als 2 Milliarden laufende Wörter in rund 160 Millionen Sätzen in deutscher Sprache vor. Das daraus erzeugte Vollformenwörterbuch ist unter *www.wortschatz.uni-leipzig.de* online zugänglich und enthält auch die Ergebnisse einer statistischen Kookkurrenzanalyse. Hier werden täglich ca. 40'000 Wörter nachgeschlagen, die Interessen der Nutzer reichen von der Suche nach Synonymen und Gebrauchsbeispielen bis hin zu Informationen über Dinge, Personen oder Ereignisse.

Das wissenschaftliche Interesse gilt jedoch hauptsächlich den automatischen Verfahren, die zur Aufbereitung und Analyse der großen Mengen Textdaten nötig sind. Eine frühe Erkenntnis bestand darin, dass sich diese Verfahren nahezu unverändert auch in anderen Sprachen anwenden lassen sollten. Seitdem steigt die Zahl der verfügbaren Sprachen ständig an, gegenwärtig (Januar 2009) liegen Daten für über 50 Sprachen vor, wobei die Textmengen stark schwanken und zwischen 5'000 Sätzen und 600 Millionen Sätzen liegen. Alle Texte werden im WWW gesammelt. Soweit verfügbar, werden drei unterschiedliche Textsorten gesammelt und in getrennten Korpora gespeichert: Zeitungstexte, allgemeine Webseiten sowie Wikipedia-Texte.

Die Datenaufbereitung erfolgt nach einem einheitlichen Schema und besteht aus den folgenden Schritten:

- Crawling: Ausgewählte HTML-Seiten werden aus dem WWW heruntergeladen.
- HTML-Stripping: Um den reinen Text zu extrahieren, werden die technisch bedingten Zusätze entfernt. Übrig bleibt Text, der allerdings noch ungewünschte Elemente enthält, etwa Menüs, Listen, Tabellen usw.
- Satzsegmentierung: Der so erhaltene Text wird zerlegt in Sätze und andere isolierte Teile, die sich zusätzlich noch in den Daten befinden.
- Putzen: In zwei Schritten wird der segmentierte Text gereinigt. Zunächst werden musterbasiert die Sätze von den Nicht-Sätzen getrennt. Dies erfolgt nach Regeln der folgenden Art:
  - Sätze beginnen mit Großbuchstaben und enden mit einem Satzzeichen. Eventuell finden sich zusätzlich An- und Ausführungszeichen am Beginn bzw. Ende.
  - Sätze enthalten sowohl absolut wie auch relativ nur eine gewissen Menge von Ziffern, Satzzeichen und anderen Sonderzeichen.
- Die Nicht-Sätze, die nicht diesen Regeln entsprechen, werden entfernt.
- Da beim Crawling immer auch Texte in anderen Sprachen eingesammelt werden, müssen im zweiten Schritt die Sätze aussortiert werden, die nicht in der gesuchten Sprache verfasst sind.
- In beiden Schritten wird recht restriktiv vorgegangen, es tritt also häufiger der Fall ein, dass aus linguistischer Sicht korrekte Sätze abgelehnt werden, als dass nicht korrekte Sätze akzeptiert werden.
- Dubletten aussortieren: Da speziell Zeitungsmeldungen in verschiedenen Zeitungen nahezu unverändert wiederholt werden, finden sich in der so entstandenen Satzliste eine große Anzahl Dubletten. Diese würden die nachfolgende statistische Analyse verfälschen und werden deshalb aussortiert.
- Nach diesem Schritt liegen die verbliebenen Sätze (mit Informationen zur jeweiligen Quelle) in alphabetischer Sortierung vor. Es ist ab diesem Zeitpunkt nicht mehr möglich, die Originaltexte vollständig zu reproduzieren, da nicht mehr alle Sätze vorhanden sind und zusätzlich die Information über die Originalreihenfolge verloren gegangen ist.
- Datenbank erzeugen: Aus der Liste von Sätzen wird im nächsten Verarbeitungsschritt eine sogenannte Textdatenbank erzeugt: Um später einen schnellen Zugriff auf die Daten zu ermöglichen, werden die Daten in eine relationale Datenbank überführt. Mit statistischen Verfahren werden außerdem weitere Daten erzeugt:
  - Frequenzangaben zu Wörtern
  - Wortkookkurrenzen: Paare von Wörtern, die statistisch gesehen auffällig oft zusam men auftreten, und zwar als unmittelbare Nachbarn (Nachbarschaftskookkurrenzen) oder gemeinsam im Satz (Satzkookkurrenzen).

Diese Verfahrensschritte laufen in einem standardisierten Verfahren ab, die als Ergebnis vorliegende Datenbank kann im Internet zum Nachschlagen angeboten werden, alternativ können die Daten lokal mit dem dafür entwickelten Korpus-Browser (cf. Richter et al. 2006) betrachtet werden.

Zu diesen jeweils aus einem Korpus erzeugten Daten liegen für die deutsche Sprache weitere Informationen über vor, wie sie auch in anderen Wörterbüchern angeboten werden: Angaben zu Wortart und Flexion, Sachgebietsangaben, Synonyme usw.

Falls verfügbar (dies ist für verschiedenen Sprachen unterschiedlich), lassen sich weitere automatische Verfahren einsetzen:

– Part-of-Speech-Tagging: Hierbei wird im Text jedem Wort seine Wortart zugeordnet. Ist ein Wort bekannt und diese Zuordnung eindeutig, dann kann das Ergebnis sofort einem Wörterbuch entnommen werden. Im Falle von Mehrdeutigkeiten (ist *einen* Verb oder unbestimmter Artikel, ist *Vogel* ein gewöhnliches Substantiv oder ein Nachname?) oder bei völlig unbekannten Wörtern wird versucht, die gesuchte Information aus dem Kontext zu entnehmen.

– Eigennamenerkennung dient dazu, im Text beschriebene Personen und Ereignisse zu erkennen. Neben der Erkennung von Personennamen sind auch geographische Namen und Firmennamen von Interesse, ebenso Zeit- und Mengenangaben.

– Trendanalyse: Speziell bei Zeitungstexten lassen sich aus dem Erstellungsdatum des Textes weitere Informationen ableiten. Je nach Häufigkeit eines Wortes lassen sich aus den täglichen, monatlichen oder jährlichen Worthäufigkeiten möglicherweise Muster oder Trends ableiten.

– Semantische Wortähnlichkeit: Inhaltlich ähnliche Wörter zeichnen sich häufig dadurch aus, dass sie auch in ähnlichen Kontexten auftreten. Betrachtet man die oben berechneten Kookkurrenzen als typische Kontexte eines Wortes, so führt ein Vergleich der Kookkurrenzmengen verschiedener Wörter zu einem Maß für deren semantische Ähnlichkeit (cf. Bordag 2008).

Im folgenden Abschnitt soll dargestellt werden, inwieweit diese Daten bei der Erstellung von Wörterbüchern genutzt werden können.

2 Unterstützung bei der Wörterbucherstellung

Wörterbucharbeit ist traditionell Handarbeit. Zwar lässt sich die Datenerfassung und Speicherung mittels Editoren und Datenbanken maschinell unterstützen, aber die Datenauswahl und Datenanordnung erlaubt nur selten maschinelle Unterstützung.

Die folgenden Beispiele zeigen jedoch, wie solche maschinellen Verfahren genutzt werden können, um Vorschläge für in Wörterbuchartikeln zu verwendende Daten zu erzeugen. Solche Daten sollen im Folgenden als Rohdaten bezeichnet werden. Die Verfahren werden möglicherweise viel zu viele Rohdaten vorschlagen, im Gegenzug werden wir aber davon ausgehen können, dass alle relevanten Vorschläge auch vorgelegt werden. Damit verändert sich die Arbeit des menschlichen Wörterbuchbearbeiters hin zu einer mehr routinemäßigen Tätigkeit, bei der vorgegebene Vorschläge angenommen oder abgelehnt und möglicherweise kleine Verschiebungen in der Einordnung vorgenommen werden. Diese Vereinfachung in der Tätigkeit sorgt erstens für eine erhebliche Beschleunigung und zweitens für eine vorher nicht da gewesene Sicherheit, alle relevanten Wörter bearbeitet zu haben. Außerdem reicht möglicherweise eine geringere Qualifikation der Bearbeiter, da sich deren jetzt eher routinemäßige Tätigkeit genauer beschreiben, organisieren und kontrollieren lässt.

Auf der Gegenseite werden neue Einsichten in die zugrundeliegenden Korpusdaten sowie in die zur Verfügung stehenden Daten benötigt, damit die automatisch erzeugten Rohdaten möglichst hohe Qualität haben. Eine Qualitätserhöhung der Rohdaten bedeutet eine Verringerung der Menge der Rohdaten und deshalb eine Reduzierung der nachfolgenden manuellen Arbeit.

3 Wörterbuchtypen

3.1 Sachgruppenwörterbuch

Das erste große Wörterbuchprojekt basierend auf den Daten des Projekts *Deutscher Wortschatz* begann im Jahr 2000 und betraf die vollständige Überarbeitung des Sachgruppenwörterbuchs *Dornseiff: Der Deutsche Wortschatz nach Sachgruppen* (cf. Dornseiff 2004). In diesem Wörterbuch sind die Wörter der deutschen Alltagssprache in knapp 1000 Sachgruppen eingeteilt, die Wörter in jeder Sachgruppe sind weiter zunächst nach Wortart und dann in semantische Gruppen untergliedert.

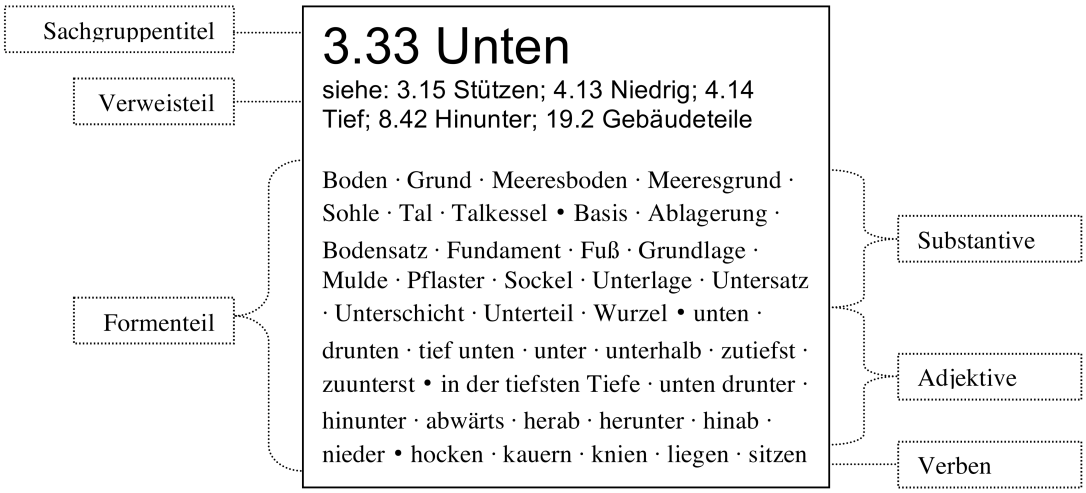

Abb. 1: Beispieleintrag nach Dornseiff 2004

Die zu überarbeitende Auflage stammte aus dem Jahr 1959 und wies nicht nur Lücken im Wortbestand der einzelnen Sachgruppen auf, sondern auch das gesamte Sachgruppensystem musste überarbeitet und in Themenbereichen wie Politik, Wissenschaft und Technik oder Sport wesentlich erweitert werden. Nachdem ein vergleichbares Vorhaben in den siebziger Jahren bereits einmal gescheitert war, standen jetzt die notwendigen Daten und Verfahren zur Verfügung. Die folgenden drei Teilaufgaben ließen sich mehr oder weniger gut maschinell unterstützen:

Aufgabe 1: Entfernen der veralteten Wörter und Wortgruppen.

Aufgabe 2: a) Finden der neu aufzunehmenden Wörter und Wortgruppen sowie
b) deren Einordnung in vorhandene Gruppen (falls sinnvoll).

Aufgabe 3: Einrichtung neuer Sachgruppen, Anpassung des Sachgruppensystems.

Während sich die Aufgaben 1 und 2a mittels der Frequenzangaben aus dem Korpus praktisch komplett lösen lassen, indem man passende Schranken für die Frequenzen angibt und sich Aufgabe 3 momentan jeder automatischen Unterstützung entzieht, stellt Aufgabe 2b die eigentliche Herausforderung dar.

Aber auch bei der frequenzbasierten Auswahl der Wörter für die Aufgaben 1 und 2a unter Benutzung eines Korpus, welches im wesentlichen aus Zeitungstexten besteht, ist Sorgfalt nötig. Das größte Problem bildet die gesprochene Sprache: Sie ist einerseits sehr reich an idiomatisch geprägten Wendungen, andererseits in großen maschinenlesbaren Sammlungen stark unterrepräsentiert. Als ausgleichende Maßnahme wurde die Schwelle für die Aufnahme von Wortgruppen im Vergleich zu Wörtern deutlich gesenkt. Ebenfalls problematisch sind veraltende Wörter. Diese sind häufig bekannt, aber ihre Verwendung in aktuellen Texten ist seltener, als ihre Bekanntheit erwarten ließe. Hier wurde davon ausgegangen, dass die zu

berücksichtigenden veraltenden Wörter in der älteren Auflage vorhanden sind. Die Schwelle für das Herausnehmen vorhandener Wörter wurde nun niedriger gesetzt als die Schwelle für die Aufnahme neuer Wörter, so dass vorhandene Wörter auch mit niedrigerer Häufigkeit im Wörterbuch belassen wurden, während neue Wörter mit vergleichbarer Häufigkeit noch nicht aufgenommen wurden.

Der nächste Schritt nach der Auswahl der neu aufzunehmenden Wörter und Wortgruppen ist deren Zuordnung zu Sachgruppen sowie die Einordnung in die angemessene semantische Gruppe. Außerdem ist zu entscheiden, ob ein Wort oder eine Wortgruppe an nur einer oder an mehreren Stellen eingeordnet werden soll.

Die folgende einfachere Formulierung desselben Problems erlaubt ein algorithmisches Herangehen: Neues Wortmaterial soll an den Stellen eingeordnet werden, an denen schon inhaltlich ähnliche Wörter der gleichen Wortart stehen. Diese Aufgabenstellung lässt sich automatisch beziehungsweise durch Verwendung vorhandener Daten lösen. Als potenziell ähnliche Wörter wurden Synonymdaten und Kookkurrenzen verwendet.

Als nächstes wurden die neu aufzunehmenden Wörter in farbiger Schrift automatisch an den so ausgewählten Stellen in das Manuskript übernommen. Anschließend wurde das so erweiterte Manuskript von Hand durchgesehen und speziell die farbigen Bereiche begutachtet: Unpassende Vorschläge wurden entfernt, ggf. wurden Verschiebungen und Teilungen von semantischen Gruppen vorgenommen. Auch wenn hier ca. 50% der automatisch erzeugten Vorschläge zurückgewiesen wurden, beschränkte sich der manuelle Aufwand auf ein einmaliges Durcharbeiten des Manuskripts.

3.2 Neologismenwörterbuch

Seit etwa 1995 liegen ausreichend viele Sätze aus verschiedenen Tageszeitungen vor, um für Wörter statistisch gesicherte jährliche Häufigkeiten zu ermitteln. Mit diesen Zahlen läßt sich feststellen, für welche Wörter diese Häufigkeiten stark schwanken, speziell welche Wörter erst ab einem bestimmten Zeitpunkt in Erscheinung traten. Aus der Tatsache, dass die jährliche Häufigkeit eines Wortes im Korpus jahrelang gleich null oder verschwindend klein war und plötzlich einen gewissen Schwellwert überschreitet, darf man schließen, dass dieses Wort in die Alltagssprache aufgenommen wurde. Das tatsächliche Entstehungsdatum für diese Wörter kann viel weiter zurückliegen und es kann nur Aufgabe eines korpusbasierten Ansatzes sein, möglichst frühe Belegstellen zu finden.

Die automatische Vorverarbeitung für das Neologismenwörterbuch (cf. Quasthoff 2007) besteht zunächst darin, Kandidaten für die Stichwörter zu ermitteln. Dazu wurden die jährlichen Anzahlen für alle Wörter aus den Jahren 1995 bis 2006 herangezogen. Gesucht werden Neologismen ab 2000, der davor liegende Zeitraum 1995–1999 dient als Vergleichszeitraum, in dem die auszuwählenden Wörter gar nicht oder extrem selten auftreten sollten. Außerdem sollen die Kandidaten für Neologismen eine gewisse Gesamthäufigkeit haben, um sie als allgemein bekannt voraussetzen zu können.

Babyklappe
GESUNDHEIT
(*Auch:* Baby-Klappe.)
Einrichtung, meist an Krankenhäusern, an der Neugeborene anonym abgegeben werden können.

Babyklappen, Wärmebettchen, in die Säuglinge von außen gelegt werden können, gibt es in Deutschland seit 2000. Sie sollen das Aussetzen oder Töten von Säuglingen verhindern und der Mutter eine anonyme Abgabe des Kindes ermöglichen. Rechtlich ist die Situation jedoch umstritten, da sich bei Findelkindern aus der Babyklappe die Herkunft nicht feststellen lässt.

Im Jahre 1709 hatte ein reicher holländischer Kaufmann eine »Babyklappe« am Waisenhaus eingerichtet. (Berliner Zeitung vom 07.03.2000) ■ *»Für Frauen in extremen Notsituationen kann die Babyklappe eine große Entlastung sein.«* (Berliner Zeitung vom 09.03.2000) ■ *Dort wurde im März 2000 die bundesweit erste Babyklappe eröffnet, nachdem 1999 zwei Neugeborene in den Müll geworfen worden sind.* (Berliner Zeitung vom 09.02.2001)

Abb. 2: Beispieleintrag aus Quasthoff 2007

Eine nach diesen Kriterien automatisch erzeugte Wortliste erhält allerdings nicht nur Kandidaten von Neologismen, sondern auch viele Eigennamen, beispielsweise von Sportlern oder Politikern, deren Popularität plötzlich zugenommen hat. In den Beobachtungszeitraum fällt auch die Rechtschreibreform: Wörter mit nur geänderter Schreibung sind natürlich ebenfalls irrelevant. Die so automatisch ermittelte Kandidatenliste muss also von Hand durchgesehen und auf die tatsächlichen Neologismen gekürzt werden.

Ist die Lemmaliste erstellt, muss für jedes Stichwort eine kurze Definition verfasst werden sowie ein kurzer Text, welcher das Wort in das zeitliche Umfeld einordnet. Außerdem sollen zwei bis drei typische Belegstellen ausgewählt werden. Hierfür wurden dem Bearbeiter für jedes Stichwort zunächst maximal 100 Belegstellen angegeben, aus denen die Beispielsätze für das Wörterbuch ausgewählt wurden. Aber auch für das Verfassen von Definition und Kurztext erwiesen sich die nicht ausgewählten Beispielsätze als extrem hilfreich, da hier häufig alle nötigen Informationen enthalten waren.

Die nachfolgende Aufstellung enthält für die Jahre 2000 bis 2008 die so ermittelte Liste von Wörtern, die in den jeweiligen Jahren davor überhaupt nicht nachweisbar waren und im ersten Jahr des Nachweises mehr als einhundert Mal auftraten.

2000: Altersvorsorgevertrag, Arbeitsagentur, Babyklappe, Bieterrennen, Breitband-Internet, brutalstmöglich, DSL-Anschluss, E-Learning, Elternzeit, Entgeltumwandlung, Festnetzsparte, Finanzportal, Green-Card-Initiative, Green-Card-Regelung, Handauszählung, Juniorprofessur, Kapitalmarktfähigkeit, Langzeitbesatzung, Metrorapid, Mittelstandsfinanzierung, Musiktausch-

börse, Separatorenfleisch, Spitzelaffäre, Studienkontenmodell, Teilzeitgesetz, UMTS-Auktion, UMTS-Geschäft, UMTS-Mobilfunklizenz

2001: ADHS, Agrarwende, Antiterrorkoalition, Antiterrorkrieg, Arzneimittelsparpaket, Biowaffenexperte, Blog, Blogger, Job-Aktiv-Gesetz, K-Frage, Kompetenzteam, Luftfahrtkrise, Markengruppe, Milzbrandfall, OBU, Patriotismusdebatte, Pisa-Ergebnis, Pisa-Test, Religionsprivileg, Riesterförderung, Riesterprodukt, Riestervertrag, Sozialforum, Squeeze-Out, Tariftreuegesetz, Terrorpilot, Verbraucherinformationsgesetz, Verbraucherministerin, Verbraucherschutzministerium, Weltsozialforum, Werbeflaute, Wettanbieter

2002: Achse des Bösen, Arbeitslosengeld I, Arbeitslosengeld II, Asbestklage, Blu-Ray, Bonusmeilenaffäre, Bundessozialministerin, Corporate-Governance-Kodex, Dankeschönspende, Fotohandy, Freiflugaffäre, Hartz-Kommission, Hartz-Papier, Hartz-Plan, Hartz-Reform, Heimatschutzministerium, Jobfloater, Lkw-Mautsystem, Lügenausschuss, Nitrofen, Nitrofenskandal, P2P-Börse, Patientenquittung, Pfandgegner, Pisa-Vergleich, Rürup-Kommission, Stammzellgesetz, Steuervergünstigungsabbaugesetz, Teuro-Debatte, Vermittlungsgutschein, Vermittlungsstatistik, Vorratsspeicherung, Waffenbericht, Wahlkampfsonderkonto, Zwangsabfindung

2003: ALG II, Allokationsplan, Alterseinkünftegesetz, Arbeitslosengeld-II-Empfänger, Bonusmaterial, Brötchentaste, Coronavirus, Defizitstrafverfahren, Dieselpartikelfilter, Doppelverfolgung, Forschungsklonen, Gesundheitskompromiss, Gesundheitskonsens, Gesundheitsmodernisierungsgesetz, Hartz-IV-Gesetz, Konventsentwurf, Kopfpauschalenmodell, Luftsicherheitsgesetz, Lungenkrankheit Sars, Mautausfall, Mautdesaster, Mautvertrag, Minijobzentrale, Offshoring, Perspektivantrag, Reformagenda 2010, Sars-Epidemie, Sars-Fall, Sechsnationengespräch, Sechsparteiengespräch, Sperrwall, Widerstandshochburg, Wiederaufbauteam

2004: Agenturbezirk, Anreizregulierung, Antiterrordatei, Desktopsuche, Ein-Euro-Job, Ein-Euro-Jobber, Exzellenzcluster, Exzellenzinitiative, Folteraffäre, Folterfoto, Gesundheitssoli, Hartz-IV-Arbeitsmarktreform, Hartz-IV-Regelung, Hauptstadtklausel, Inverssuche, Nipplegate, Ombudsrat, Optionsgesetz, Phishing, Phishing-Attacke, Phishing-Mail, Reiseschutzpass, Rosenrevolution, Rürup-Rente, Tagesbetreuungsausbaugesetz, Terminierungsentgelt, USB-Stick, Wahlfälschungsaffäre

2005: Ärztetarifvertrag, Bombenholocaust, Bundesnetzagentur, Bundestagsnachwahl, Bundesverbraucherminister, Bundesverbraucherschutzminister, Campus-Maut, Doppelkern-Prozessor, Ekelfleisch, Fanmeile, Feinstaubrichtlinie, Fernsehbeirat, Freilaufverbot, Fußballwettskandal, Gammelfleisch, Gammelfleischskandal, Gefangenenflug, Geheimflug, Globalisierungsfonds, Handyparken, Heuschreckendebatte, Hurrikanhilfe, IPTV, Jamaika-Koalition, Malware, Nachholfaktor, Neuwahlankündigung, Neuwahlcoup, Neuwahlplan, Optionsticket, Rucksackbomber, Sudoku, Tsunami-Hilfe, Umweltzone, Unterschichtenfernsehen, Urankonversion, VDSL-Netz, Visaausschuss, Vogelgrippegefahr

2006: Bahnattentat, Bahnbombenleger, Blauzungenkrankheit, Deutschpflicht, Fangteam, Grundsatzkongress, Hartz-IV-Fortentwicklungsgesetz, Hauspflicht, Integrationsgipfel, Integrationsverweigerer, Karikaturenstreit, Kofferbomber, Namenschutz, Nutztierbetrieb, Optimierungsgesetz, Partnermonat, Poloniumspur, Prekariat, Repoxygen, Vogelgrippegebiet, Welterbetitel, Wildfleischskandal, Zwergplanet

2007: Abschmelzmodell, ABS-Fonds, Anspielversion, Flatrate-Partys, Hypothekenkrise, iPhone-Besitzer, Kinderbildungsgesetz, Kohlestiftung, Konjunkturbonus, Kreditmarktkrise, Krippenausbau, Nazometer, Onlinedurchsuchung, Postmindestlohn, Subprime-Hypothek, Subprimekrise, Subprime-Markt, Subprime-Segment, Weltklimabericht

2008: Bildungsrepublik, Community-Links, Einlegerschutz, Holzklotzwurf, Konjunkturhilfe, Konjunkturspritze, Konsumgutschein, Konsumscheck, Netbook, Rettungsschirm, Schuhwerfer, sendungsbezogen, Steuerscheck, Superdelegierte, Videorubrik

ISSN 1615-3014

3.3 Kollokationswörterbuch

Während es für das Englische mehrere Kollokationswörterbücher (cf. Benson et al. 1997, Hill/Lewis 1997, Crowther et al. 2002) gibt, ist dies für das Deutsche nicht der Fall. Andererseits liefern die automatisch ermittelten Wortkookkurrenzen auf den ersten Blick die für solch ein Wörterbuch notwendigen nötigen Rohdaten. Unter Zugrundelegung des Kollokationsbegriffs von F. Hausmann (cf. Hausmann 1985) wurde das Kollokationswörterbuch (ähnlich dem Sachgruppenwörterbuch) als Wortfindungswörterbuch konzipiert, allerdings mit einem größeren Schwerpunkt auf der Nutzung durch Fremdsprachler.

Die Auswahl der Stichwörter erfolgt frequenzbasiert für die Wortklassen Substantiv (S), Verb (V) und Adjektiv/Adverb (A). Bei einer vorgesehenen Anzahl von etwa 4'500 Stichwörtern erfolgte deren Auswahl strikt frequenzbasiert mit Hilfe des Korpus. Nicht berücksichtigt werden Eigennamen, deren Ableitungen (wie *US-Präsident*) sowie verbliebene Wörter, für die sich keine Kollokationen finden ließen.

Zunächst gilt es, für jedes Stichwort die Kandidaten aus den Kookkurrenzen in Abhängigkeit von Wortart und Position auszuwählen. Für Stichworte der Wortklassen A und V sind nur Kollokationen vom Typ A interessant, diese treten typischerweise als unmittelbare linke Nachbarn auf. Für Stichworte vom Typ S gibt es Kollokationen vom Typ A und vom Typ V. Die Adjektive treten typischerweise wieder als unmittelbare linke Nachbarn des Substantivs auf. Bei Verben als Kollokationen von Substantiven kann das Substantiv in verschiedenen Rollen auftreten: Als Subjekt, als Objekt oder in Präpositionalphrasen. Alle diese haben gemeinsam, dass das Substantiv als unmittelbarer linker Nachbar des Verbs auftreten kann. Dies ist in der Regel nicht die einzige möglich Position oder auch nicht die mit der größten Häufigkeit, aber der Fall der unmittelbaren Nachbarschaft tritt häufig genug auf, um statistisch signifikant zu sein. Auf die Verwendung allgemeinerer Satzkookkurrenzen wurde verzichtet, da diese zu viel nicht verwertbares Wortmaterial liefern.

Dichter

V: dichten · herausgeben · hervorbringen · schreiben • dokumentieren · erfinden · imaginieren • hinterlassen • lehren • widmen • besingen · schwärmen • *Akk.* auszeichnen · ehren · krönen · würdigen • *Akk.* lesen · zitieren

A: anerkannt · bedeutend · bekannt · berühmt · geehrt · gefeiert · groß · namhaft · prominent · verehrt • arm · einsam · erfolglos · gescheitert · unbekannt · verkannt • schreibend · singend • melancholisch · traurig · unglücklich · verzweifelt • charismatisch · genial · kritisch • gebildet · gelehrt · professionell • alt · älter · alternd · jung · jünger · zeitgenössisch • tot · verstorben • einheimisch • verbannt • dramatisch · experimentell · expressionistisch · politisch · postmodern · romantisch

dick

A: auffallend · auffällig · außerordentlich · besonders · extrem · mächtig · richtig · total · ungewöhnlich · unglaublich • recht · ziemlich • ungleich · unterschiedlich • genauso · gleich • gleichmäßig • ausreichend • ganz · sehr • wenig

Abb. 3: Beispiel aus Kollokationswörterbuch

analysieren analysierte, hat analysiert, AKK, ige, vál

elemez, analizál VMIT
AKK Roman, Text, Satz, Musikstück, Traum, Lage, Situation, Probleme, Beziehung
<ADV> **sorgfältig** gondosan * **eingehend** behatóan, tüzetesen * **gründlich** alaposan * **wissenschaftlich** tudományosan * **exakt** pontosan, egzakt módon * átv **näher** közelebbről * **statistisch** stilisztikailag * **systematisch** szisztematikusan * **logisch** logikusan, logikailag * vál **minuziös** (!) aprólékosan * **präzis(e)** pontosan, precízen

Abb. 4: Beispieleintrag Hollós

Damit bilden Nachbarschaftskookkurrenzen unter Berücksichtigung der Wortartenpaare das Ausgangsmaterial für das Kollokationswörterbuch. Problematisch ist die Menge der zu

berücksichtigenden Kookkurrenzen. Es gibt verschiedene Maße (cf. Evert/Krenn 2001), mit deren Hilfe die Kookkurenzen zu einem Wort nach Stärke geordnet werden können. Das häufig (und auch hier) verwendete Log-Likelihood-Maß von Dunning (cf. Dunning 1993) entspricht wahrscheinlich am besten der menschlichen Assoziation von Wörtern, muss deshalb jedoch nicht das geeignetste Maß sein. Versuche zeigen aber, dass sich keines der bekannten Maße besonders gut eignet, Kollokationen aus den Kookkurrenzen auszuwählen. Auch unter den (im statistischen Sinne) schwachen Kookkurrenzen finden sich regelmäßig noch (linguistisch wertvolle) Kollokationen. Es bleibt also nur der relativ aufwändige Weg, die automatisch ermittelten und nach Wortart vorsortierten Kookkurrenzen nach Kollokationen zu durchsuchen. Dabei wurden im vorliegenden Projekt je nach Wortart und Häufigkeit 20–50% der ausgewählten Kookkurrenzen übernommen.

Die Anordnung der Kollokationen erfolgt wieder sortiert nach syntaktischen und semantischen Kriterien. Zunächst werden die Kollokationen nach Wortarten sortiert, die Verben zusätzlich nach der syntaktischen Rolle des Stichwortes. Größer verbleibende Gruppen werden nach semantischen Kriterien weiter unterteilt.

Natürlich ist die Verwendung solcher Rohdaten auch für anders konzipierte Wörterbücher möglich. Als Wörterbuchtyp eng verwandt ist ein zweisprachiges Kollokationswörterbuch, deshalb ließen sich die nach den gleichen Kriterien ausgewählten Rohdaten auch für ein Deutsch-Ungarisches syntagmatisches Lernerwörterbuch (cf. Hollós) verwenden. Bei einem Umfang von etwa 2'300 Einträgen und der Berücksichtigung des Ungarischen als Zielsprache ergaben sich natürlich andere Kriterien für die Stichwortauswahl sowie zusätzliche manuelle Bearbeitungsschritte, auf die hier nicht eingegangen werden kann.

3.4 Frequenzwörterbuch

Während klassische Frequenzwörterbücher häufig als Lernerwörterbücher für Fremdsprachler konzipiert sind und dementsprechend nur geringen Unfang haben, erlaubt der korpusbasierte Ansatz bei großen Korpora auch große, aussagekräftige Frequenzwörterbücher. Dem möglicherweise eingeschränkten Nutzerkreis eines solchen großen Frequenzwörterbuches stehen allerdings umfangreiche Verwendungsmöglichkeiten der Frequenzlisten als Ausgangsmaterial für andere Wörterbücher, für theoretische Untersuchungen oder technische Anwendungen gegenüber. Sinnvoll erscheint hier die gemeinsame Veröffentlichung des Frequenzwörterbuches (egal ob als Printwörterbuch oder in elektronischer Form, beispielsweise als pdf-Datei) zusammen mit einer Veröffentlichung der Daten in einer Form, dass sie sowohl im technischen wie im rechtlichen Sinne möglichst einfach und vielseitig genutzt werden können.

Die unterschiedlich komplexe Bearbeitung einer Wortliste ermöglicht verschiedene Arten von Frequenzwörterbüchern, verbunden mit den entsprechenden Vor- und Nachteilen. Da wegen der extrem großen Datenmengen die Bearbeitung fast ausschließlich vollautomatisch erfolgen muß und Bearbeitungsfehler teilweise unbemerkt bleiben, führt ein höherer Bearbeitungsgrad fast von selbst zu einer höheren Fehlerrate in den erzeugten Daten. Die folgende Zusammenstellung zeigt verschiedene Formen von Frequenzwörterbüchern mit aufsteigendem Bearbeitungsaufwand.

Wortformenwörterbuch mit Frequenzangaben: Dies ist mit automatischen Mitteln praktisch fehlerfrei zu erzeugen. Aufgenommen und gezählt werden alle im Korpus gefundenen Wörter in ihren verschiedenen Formen.

Frequenzwörterbuch auf Basis der Grundformen: Zusätzlich erfolgt eine Lemmatisierung, die verschiedene Vollformen zu einer Grundform zusammenfasst. Die Anzahlen der Wortformen *Haus*, *Hauses*, *Häuser* und *Häusern* werden beispielsweise addiert und bei *Haus* verzeichnet.

Problematisch sind hier Mehrdeutigkeiten und Eigennamen. Wie viele der Wortformen *einen* gehören zum Artikel *ein* und wie viele zum Verb *einen*? Der Familienname *Steinhäuser* darf nicht auf *Steinhaus* reduziert werden. Diese Unterscheidungen können zwar automatisch mittels POS-Tagging (cf. Manning/Schütze 2000) vorgenommen werden, aber dabei vorkommende (und zum Teil systematische) Fehler beinträchtigen den Wert der ermittelten Häufigkeitsangaben.

Trennung von Eigennamen: Von Interesse sein kann sowohl die Häufigkeit von Eigennamen wie auch eine Liste von Wörtern ohne die Berücksichtigung der Eigennamen. Hier bietet sich wieder die Verwendung eines POS-Taggers an, der diese Trennung aber nur unscharf vornehmen kann.

Abb. 5: Auszug Häufigkeitswörterbuch

Um einen Eindruck über den Umfang solcher korpusbasierter Frequenzwörterbücher zu gewinnen, sei folgendes Beispiel betrachtet: Aus den verschiedenen deutschsprachigen Korpora des Projekts *Deutscher Wortschatz* (d. h. Zeitungstext, beliebige Webseiten und Wikipedia-Artikel) wird eine Vollformen-Wortliste erstellt. Betrachtet man nur die Wortformen mit einer Mindestanzahl von 30, so enthält die Liste rund eine Million Einträge. Der nebenstehende Ausschnitt enthält neben den Wörtern ihre Häufigkeitsklasse: Die häufigsten Wörter *der*, *die* sowie *und* tragen die Häufigkeitsklasse 0, die Vergrößerung der Häufigkeitsklasse um eins entspricht näherungsweise der Halbierung der Worthäufigkeit. Der besseren Übersicht halber sind die Wörter der Häufigkeitsklassen 0–16 (dies sind rund die häufigsten 10% aller Wörter) fett gesetzt.

Gedruckt hätte dieses Häufigkeitswörterbuch den Umfang von etwa drei Bänden im Format DIN A4 (2500 Seiten mit je 5 Spalten und 80 Zeilen).

4 Schlussbemerkungen

Die praktischen Beispiele zeigen, dass aus Korpora extrahierte Daten für die Wörterbuchproduktion direkt einsetzbar sind und möglicherweise den Aufwand bei der Wörterbucharbeit drastisch senken können. Dazu muss sich allerdings der Wörterbuchtyp eignen. Außerdem sind Überarbeitungen vorhandener Wörterbücher in der Regel einfacher zu unterstützen als die Neuerstellung, da im ersten Fall das vorhandene Grundgerüst an Daten die automatischen Verfahren unterstützen kann.

Prinzipiell bieten sich die folgenden Daten an:

– Häufigkeitsangaben zu Wörtern, evtl. in verschiedenen, z. B. zeitlich geordneten Korpora: Diese eignen sich zur Stichwortauswahl für ein Wörterbuch oder zur Beurteilung der Gebrauchsentwicklung.

– Nutzung von Beispielsätzen: Die Vorkommen einzelner Wörter lassen sich für die Auswahl von Belegstellen verwenden, für die Erstellung von Bedeutungsbeschreibungen sowie für die quantitative Beurteilung verschiedener Wortbedeutungen.

– Wortkookkurrenzen: Statistisch auffällig oft gemeinsam auftretende Wörter können durch verschiedenartige Relationen miteinander verknüpft sein. Sie lassen sich beispielsweise benutzen bei der Untersuchung von Kollokationen oder der Klassifikation von Wörtern nach Sachgebieten.

Um jedoch die bestmöglichen Rohdaten für ein Wörterbuch aus einem Korpus zu extrahieren, ist ein gegenseitiges Verständnis und eine enge Zusammenarbeit der Beteiligten bei der Korpusauswertung und Wörterbucharbeit nötig. Hier bilden sich gegenwärtig für die Wörterbuchproduktion wichtige Methoden und Verfahren zur Bereitstellung der Rohdaten heraus. In vielen Fällen sind diese Verfahren einzelsprachenunabhängig, eignen sich also auch für andere Sprachen. Dieses Potenzial, aus großen Korpora verschiedener Sprachen mit einzelsprachenunabhängigen Verfahren die Rohdaten für Wörterbücher in mehreren Sprachen zu extrahieren, kann völlig neue Formen für die internationale Zusammenarbeit bei der Wörterbuchproduktion schaffen.

Literatur

Benson, Morton/Benson, Evelyn/Ilson, Robert (eds.) (1997): *The BBI Dictionary of English Word Combinations*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bordag, Stefan (2008): "A Comparison of Co-occurrence and Similarity Measures as Simulations of Context". In: Gelbukh, Alexander (ed.): *Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008. Haifa, Israel, February 17–23, 2008*. Berlin/Heidelberg: Springer. (= *Lecture Notes in Computer Science* 4919).

Dornseiff, Franz (2004): *Der deutsche Wortschatz nach Sachgruppen*. 8. völlig neu bearb. Auflage von Uwe Quasthoff. Berlin/New York: de Gruyter.

Dunning, Ted (1993): "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 19/1.

Evert, Stefan/Krenn, Brigitte (2001): "Methods for the qualitative evaluation of lexical association measures. Toulouse/France, 2001". *Proceedings of the 39th Annual Meeting of the ACL*: 188–195.

Hausmann, Franz Josef (1985): "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels". In: Bergenholtz, Henning/Mugdan, Joachim (eds.): *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.–30.06.1984*. Tübingen, Niemeyer: 118–129. (= *Lexicographica. Series Maior* 3).

Hill, Jimmie/Lewis, Michael (eds.) (1997): *The LTP Dictionary of Selected Collocations*. Hove: Language Teaching Publications.

Hollós, Zita (in Vorbereitung): *SZÓKAPTÁR. korpusbasiertes deutsch-ungarisches syntagmatisches Lernerwörterbuch*.

Manning, Christpher D./Schütze, Hinrich (2000): *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Crowther, Jonathan/Dignen, Sheila/Lea, Diana (2002): *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Quasthoff, Uwe (ed.) (2007): *Deutsches Neologismenwörterbuch*. Berlin/New York: de Gruyter.

Richter, Matthias et al. (2006): "Exploiting the Leipzig Corpora Collection". *Proceedings of the Information Society Language Technologies Conference (IS-LTC) 2006. Ljubljana, Slovenia*. nl.ijs.si/is-ltc06/proc/13_Richter.pdf (Stand: September 2009).