Linguistik online 3, 2/99

Korpuslinguistik und große einsprachige Wörterbücher

Uwe Quasthoff & Christian Wolff (Leipzig)

Der vorliegende Beitrag diskutiert die Problematik des Aufbaus elektronischer Wörterbücher und stellt das Projekt *Deutscher Wortschatz* vor, ein umfangreiches Vollformenwörterbuch des Deutschen, das seit 1995 am Institut für Informatik der Universität Leipzig entsteht.

1 Wörterbücher als Buch, als CD-ROM und im Internet

Traditionelles Anliegen einsprachiger Wörterbücher ist es, Informationen über typischerweise bekannte Wörter zusammenzustellen, (Flexionsangaben, etymologische Angaben, Unterscheidung mehrerer Bedeutungen, Gebrauchsbeispiele speziell in Redewendungen usw.) und unbekannte Wörter inhaltlich erschließen zu helfen. Dabei gibt es unterschiedliche Möglichkeiten bei der Reihenfolge der Darstellung ("normale" Reihenfolge in lexikographischer Sortierung, rückläufige Wörterbücher, Anordnung nach inhaltlichen Kriterien, Wortlänge usw.). Der Umfang solcher Wörterbücher liegt wegen des begrenzten Umfangs einer Druckausgabe zwischen einigen 10.000 und einigen 100.000 Stichwörtern.

Elektronische Medien wie CD-ROM und Internet schaffen hier neue Aufbereitungs- und Nutzungsformen, da mit ihnen nicht nur möglich ist, sehr große Datenbestände bei vertretbaren Kosten zu distribuieren, sondern durch automatische Erschließungsverfahren (Recherchefunktionen, Indices etc.) und Hypertextverknüpfungen über das direkte Nachschlagen einzelner Lemmata hinaus neue Nutzungsformen eröffnen. Für zahlreiche gedruckte Lexika sind in den vergangenen Jahren elektronische Fassungen erstellt worden, i. d. R. auf CD-ROM und unter weitgehend unveränderter Übernahme des Ausgangsdatenbestands. Hier spielt die sog. medienneutrale Datenhaltung eine große Rolle, da sich mit ihr eine vom Zielformat (Druck oder elektronische Version) weitgehend unabhängige Verwaltung und Bearbeitung der Substanzen möglich ist, (sog. cross-media publishing, cf. Heyer 1995, Kamps et al. 1999). Diese Wörterbücher auf CD-ROM konnten jedoch bisher die zunächst hohen Erwartungen der Verlage beim Absatz nicht erfüllen, wofür vielfältige Gründe verantwortlich gemacht werden können, u. a.:

- mangelnde Qualität (die sog. *digitizing straw into gold-fallacy*)
- heterogene Formate
- zunächst zu kleine Schnittmenge von Zielgruppe und potentiellem Nutzerkreis mit adäquater technischer Ausstattung
- keine hinreichende Integration in die traditionellen Distributionskanäle (insb. Buchhandel).

Aus wirtschaftlichen Gründen ist es damit für Verlage riskant, Arbeit in sehr große Wörterbücher zu investieren, die nicht primär für den Druck gedacht sind.
Mit der Verbreitung des World Wide Web ist sind zahlreiche Lexika und Wörterbücher für den unmittelbaren Zugriff verfügbar geworden (vgl. z. B. http://www.fask.uni-mainz.de/user/wschmidt/onldict.html). Bei den einsprachigen Wörterbücher im World Wide Web handelt es sich fast ausnahmslos um Erklärungswörterbücher oder Thesauri, also Einordnungen von Fachbegriffen in eine Begriffshierarchie. Dabei sind Umfang, Qualitätsstandard und Anspruch höchst unterschiedlich.

Aus der Erkenntnis, daß ein wirklich umfangreiches einsprachiges Wörterbuch des Deutschen nach klassischem Vorbild auch in der nächsten Zeit nicht zur Verfügung stehen wird und auch nicht mit vertretbarem Aufwand herstellen läßt, wurde nach anderen Lösungsmöglichkeiten gesucht. Die zu erfüllenden Randbedingungen waren:

- Das Wörterbuch soll über das World Wide Web frei zugänglich sein;
- sein Umfang soll hinsichtlich der Quantität wie die Binnenstruktur der Einträge möglichst umfangreich sein, d. h. über mehrere Millionen Stichwörter verfügen;
- Ausgangspunkt der Konzeption war die Entscheidung für ein Vollformenlexikon, in dem sich die tatsächliche Sprachverwendung aus elektronisch verfügbaren Quellen widerspiegelt;
- es sollte keine a priori-Festlegung auf ein bestimmtes linguistisches oder ontologisches Modell erfolgen, um die Einarbeitung einer Vielzahl heterogener Quellen zu ermöglichen;
- die typischen Inhalte und Funktionen unterschiedlicher Wörterbuch- und Lexikontypen sollen zur Verfügung stehen (Nachschlagen von Begriffen; Querverweise; morphologische, syntaktische, semantische und pragmatische Information, statistische Daten; Einarbeitung von Ontologien);
- und durch die zusätzlichen Möglichkeiten des elektronischen Mediums (automatische linguistische Analyseverfahren, Recherche, Hypertextualisierung, automatische Generierung unterschiedlich strukturierter Einträge, Visualisierung von Relationen zwischen Einträgen) ergänzt werden.
- das Wörterbuch soll für unterschiedlichste Anwendungszwecke dienen können (linguistische Analysen; Einsatz als Informationsquelle in der Sprachprodukttechnologie; Anwendungen in Information Retrieval und Suchmaschinen)
- das Projekt sollte an einem universitären Informatikinstitut angesiedelt und mit geringen personellen Ressourcen durchführbar sein.

Das so entstehende Wörterbuch will und kann kein Ersatz für bestehende, erfolgreiche Wörterbücher sein. Aber es kann Antworten auf viele Fragen geben, die bei anderen Wörterbüchern offen bleiben müssen.

2 Das Wortschatz-Lexikon

Das Wortschatz-Lexikon (http://wortschatz.uni-leipzig.de) ist ein Vollformenlexikon des Deutschen und verfügt derzeit über mehr als fünf Millionen Vollformen mit einer je unterschiedlichen Anzahl zusätzlicher Angaben.

2.1 Datenbestand

Das Wortschatz-Lexikon ist aus einer Vielzahl unterschiedlicher Quellen aufgebaut, wobei quantitativ der Volltext mehrerer Tageszeitungen im Vordergrund steht; darüber hinaus gehen aber auch Fachlexika, Fachzeitschriften und Monographien aus unterschiedlichen Wissensgebieten (u. a. Medizin, Rechtswissenschaft, Inormatik) in den Datenbestand ein.

Die Quellen werden zur Auswertung von Autoren und Verlagen zur Verfügung gestellt, allerdings mit der Einschränkung, daß die ursprünglichen Texte nicht im Zusammenhang rekonstruierbar sind, also etwa ein Zeitungsartikel nicht vollständig ausgelesen werden kann. Aus diesem Grund arbeitet das Wortschatzprojekt weitgehend satzbasiert; die in einzelne Sätze zerlegten Text sind Ausgangspunkt unterschiedlicher Analysen, z. B. zur Kollokationsermittlung.

Die Angaben zu Flexion, Morphologie, Sachgebiet, Synonyme etc., die vor allem aus strukturierten Quellen wie Lexika und durch computerlinguistische Analyseprogramme ermittelt werden, sind der jeweiligen Grundform eines Wortes zugeordnet, flektierte Formen enthalten statt dessen einen Verweis auf die entsprechende Grundform. Momentan haben diese Angaben noch nicht überall die gewünschte Qualität, da sie teilweise mit automatischen Mitteln erzeugt wurden und bei der Anzahl der Einträge eine vollständige intellektuelle Überarbeitung nicht möglich ist. Über das Inventar der Beschreibungskategorien klassischer Wörterbücher hinaus sind weitere Angaben vorhanden: Für jedes Wort werden zusätzlich

(soweit vorhanden) drei Beispielsätze angezeigt, die das Wort enthalten. Zusätzlich sind für alle Formen Häufigkeitsangaben vorhanden. Sie beruhen auf dem Vorkommen in den insgesamt etwa 13 Millionen Beispielsätzen. Aus diesen Beispielsätzen lassen sich weitere Informationen gewinnen: Um den typischen Gebrauch eines Wortes zu ermitteln, werden häufig Kollokationen untersucht. Im Wortschatz-Lexikon werden zwei Arten von Kollokationen angegeben: Berücksichtigt wird einmal statistisch signifikant häufiges gemeinsames Auftreten in einem Satz, zum anderen statistisch signifikant häufiges Auftreten als unmittelbarer rechter bzw. linker Nachbar. Mit 3.8 Millionen Kollokationsangaben zu 260.000 Wörtern ist das Wortschatz-Lexikon die einzige umfassende und online verfügbare Quelle für Kollokationen des Deutschen, für den englischen Sprachraum gibt es die COBUILD English Collocations on CD-ROM, vgl. http://titania.cobuild.collins.co.uk/collscd.html.

2.2 Beispiele

Die nachfolgenden Beispiele "Lenker", "Oxymoron" und "maschinenlesbarer" (Stand: September 1999) sollen die Strukturen und Recherchemöglichkeiten illustrieren. Sie wurden über entsprechende Anfragen an das Wortschatz-Lexikon ermittelt, die über die angegebenen URLs nachvollzogen werden können.

Wort: Lenker
(http://wortschatz.informatik.uni-leipzig.de/cgi-bin/wort_www.exe?site=1&Wort=Lenker&cs=1)

Häufigkeitsklasse: 13 (d.h. der ist ca. 2^13 mal häufiger als das gesuchte Wort)
Sachgebiet:
Technik
Nachname
Morphologie:
lenk|er
Grammatikangaben:
Wortart: Substantiv
Wortart: Eigenname
Geschlecht: männlich
Flexion: der Lenker, des Lenkers, dem Lenker, den Lenker
die Lenker, der Lenker, den Lenkern, die Lenker
Kollokationen:
Denker, Sattel, Pedale, Mittelfeld, Personenwagens, Sitzposition, abgeschleppten, Beifahrer, Hände, zusammenstieß,
Traktors, Strachan, geschwungenem, Kleinbusses, Gegenfahrbahn, Verletzungen
Beispiel(e):
Mit einem Ruck hebt er das Hinterteil aus dem Sattel, beugt den Kopf wie ein angriffslustiger Stier tief über den Lenker und tritt mit aller Kraft in die Pedale. (Quelle: FAZ 1994)
Nur wenige Wirtschaftsjournalisten murren über die unangefochtenen Monopole, etwa der Fiat-Gruppe, deren Lenker über Schachteln in der Schachtel überproportionale Wirtschaftsmacht mit allen politischen Konsequenzen ausüben. (Quelle: FAZ 1994)
Aber es sind nicht die von Mussorgskis Erzähler und Tod: Ein Pferdeschlitten kommt entgegen, der Lenker parliert über Walkie-talkie mit einem anderen. (Quelle: FAZ 1994)

Wort: Oxymoron
(http://wortschatz.informatik.uni-leipzig.de/cgi-bin/wort_www.exe?site=1&Wort=Oxymoron&cs=1)

Häufigkeitsklasse: 19 (d.h. der ist ca. 2^19 mal häufiger als das gesuchte Wort)
Sachgebiet:
Allgemeine Literaturwissenschaft
Morphologie:

oxy|mor|o|n
Grammatikangaben:
Wortart: Substantiv
Geschlecht: sächlich
Pragmatikangaben:
etym: griech.
Beispiel(e):
Die deutsche Talkshow ist keine Sendeform oder ein Genre, sondern sprachtechnisch ein Oxymoron und ingesamt ohnehin eine offenbar heillose Krankheit. (Quelle: TAZ 1991)
Vorweg: 'Konzertierte Abschreckung' ist ein Oxymoron, Logiker-Jargon für 'Widerspruch in sich selbst'. (Quelle: Sueddeutsche Zeitung 1995)
'Understanding Media' ist insofern ein Oxymoron, als Medien als Formen aller Anschauung unhintergehbar sind und allem Verständnis vorausgehen. (Quelle: Sueddeutsche Zeitung 1995)

Wort: maschinenlesbarer
(http://wortschatz.informatik.uni-leipzig.de/cgi-bin/wort_www.exe?site=1&Wort=maschinenlesbarer&cs=1)

Häufigkeitsklasse: 19 (d.h. der ist ca. 2^19 mal häufiger als das gesuchte Wort)
Morphologie:
maschin|en|lesb|ar|er
Grammatikangaben:
Wortart: Adjektiv
Stammform: maschinenlesbar
Kollokationen:
Code
Beispiel(e):
Hierzu kann ein klassifizierendes Nummernsystem zur Kennzeichnung der Recyclingeigenschaften des Produkt sowie der demotierbaren Baugruppen und Bauteile dienlich sein, das auf den Komponenten als maschinenlesbarer Code aufgebracht ist. (Quelle: VDI-Nachrichten 1990/91)
Wochen später "darf" er sich eine 10,5 x 7,4 cm kleine Plastikkarte, genannt "fälschungssicherer, maschinenlesbarer Personalausweis", abholen. (Quelle: TAZ 1987)
Gerade rechtzeitig zum Jubiläum haben die Aidlinger Computerfreaks eigens eine eigene CD-Rom - eine immense Menge maschinenlesbarer Informationen also - über sich und ihren Liebling produziert. (Quelle: Stuttgarter Zeitung 1996)

3 Angewandte Korpuslinguistik

Als Grundlage für den Aufbau des Wortschatz-Lexikons dienen Methoden und Verfahren der Korpuslinguistik, einem Teilgebiet der Computerlinguistik mit engen Verbindungen zur angewandten Informatik. Anliegen der Korpuslinguistik ist es, aus umfangreichen Datenkorpora (i. d. R. Text) Informationen, insbesondere linguistische Beschreibungskategorien, zu ermitteln. Betrachtet man Korpuslinguistik als Arbeitsgebiet der angewandten Informatik bzw. der Computerlinguistik, so liegt der Schwerpunkt auf dem Einsatz automatischer Methoden und der Bearbeitung großer Datenmengen: Einerseits erfordern statistische Methoden in der Regel gewisse Mindesthäufigkeiten für die Ermittlungen zuverlässiger Aussagen. Andererseits lassen sich große Datenmenge mit vertretbarem Aufwand nicht ohne den nachhaltigen Einsatz von Rechenleistung untersuchen. Dies soll am Beispiel von Kollokationen erläutert werden.

Kollokationen von Wörtern im Text beschreiben das Phänomen gemeinsamen Auftretens von Wörtern; dabei versucht man unter allen vorhandenen Kollokationen diejenigen zu ermitteln, die sich auf der Basis eines zur Berechnung herangezogenen Kollokationsmaßes als signifikant erweisen. Die Ermittlung signifikanter Kollokationen ist ein seit langem in der Korpuslinguistik angewandtes Analyseverfahren (cf.

Lemnitzer 1998, Kim & Choi 1999). I. d. R. werden dabei sog. 2er-Kollokationen betrachtet. Um überhaupt mehrere signifikante Kollokationen für ein Wort finden zu können, müssen mindestens 50 Vorkommen untersucht werden, wobei die Anzahl ermittelter signifikanter Kollokationen mit der Häufigkeit eines Wortes steigt. Damit kommen nur Wörter mit einer absoluten Häufigkeit von mindestens 50 für diese Untersuchung in Frage. Dabei handelt es sich aber nur um einen Bruchteil der insgesamt bekannten Wörter, derzeit rund 260.000 Wörter. Für weitere Untersuchungen ist also auch ein weiterer Ausbau der Belegmenge unumgänglich. Bei den hinreichend häufig belegten Einträgen lassen sich aber recht gute Ergebnisse, z. B. für die Ermittlung von Kohyponymen erzielen, wie die obigen Beispieleinträge und die nachfolgende Kollokationsliste für "Bier" zeigen:

Kollokationen zu *Bier*:
trinken, Wein, Liter, getrunken, trinkt, gebraut, Brauerei, Reinheitsgebot, Schnaps, Maß, Glas, Brauereien,
Hektoliter, Halbe, Flasche, Dosen, hl, brauen, tranken, Kaffee, Dose, Faß, Kasten, Cola, Sekt, Getränke, Flaschen, Münchner, Mineralwasser, alkoholfreies

Auch die Suche nach Phrasen, also aus mehreren Worten bestehenden Folgen, die signifikant häufig auftreten ("feste Wendungen"), erweist sich die Belegmenge zu als häufig zu klein. Beispielsweise findet sich im Deutschen Wortschatz mit Hilfe der Recherchefunktion "Suche im Satz" nur maximal fünf Belegstellen für

- *ohne lange zu fackeln*
- *schlagende Wetter*
- *von den Augen ablesen*
- *über den grünen Klee loben*

Damit ist der Deutsche Wortschatz nur bedingt in der Lage, typische Redewendungen zuverlässig zu ermitteln. Neben der Vermutung, daß eine ungerichtete quantitative Erweiterung der Belegstellen (z. B. um eine Größenordnung) hier Abhilfe schaffen kann, stellt sich die Frage nach der Auswahl der Texte für das Korpus. Für das Wortschatz-Lexikon gelten hier folgende Kriterien:

- auszuwertender Text muß maschinenlesbar vorliegen,
- er muß einem gewissen orthographischen Mindestniveau genügen, das nicht schlechter als das von Zeitungstext sein sollte
- Ziel ist weiter die thematische und sprachliche Vielfalt der Quellen.

Das letzte Ziel ist wohl das anspruchsvollste: Letztlich wird damit gefordert, ein Mindestmaß an Belegstellen für alle Wissensgebiete und sprachliche Niveaus zu erschließen. Man muß sich dabei bewußt sein, daß aufgrund der Unterschiedlichkeit der Quellen in Hinsicht auf Sprache, Thematik und Umfang ein gleichmäßiges Besetzen eines hypothetischen "Wissensbaums" nur näherungsweise möglich ist. Beispielsweise dürfte der eingearbeitete Zeitungstext aufgrund seines Umfangs die Kollektion dominieren, während Fachbegriffe aus wissenschaftlichen Spezialgebieten oftmals nur in Quellen geringen Umfangs (z. B. ein Fachlexikon) auftreten und daher keine hohen Häufigkeiten erreichen können. Dies schließt allerdings nicht aus, daß der von einen Fachbegriff ermittelte Eintrag eine reiche Binnenstruktur aufweist. Er mag daher zwar kaum statistische Analysen zugänglich sein, liefert aber in anderer Hinsicht wertvolle Information (Einordnung in Begriffshierarchie, Zuordnung zu Sachgebieten etc.).

4 Technische Herausforderungen

Der Umfang des Deutschen Wortschatzes beträgt einschließlich Beispielsätzen und Indexdateien derzeit etwa 20 Gigabyte Textdaten. Der Umgang mit solchen Massendaten bringt eine Reihe technischer Schwierigkeiten mit sich: Zunächst ist das Internet die einzige geeignete Plattform, ein solches Wörterbuch zugänglich zu machen. Eine CD-ROM reicht für die Datenmenge lange nicht mehr aus, DVD (digital versatile disk) mit ihrem deutlich höheren Fassungsvermögen von mehreren Gigabyte Daten wäre

prinzipiell eine Alternative, ist aber bei Nutzern nicht verbreitet und langfristig nicht ausreichend.

Zur Datenhaltung wird momentan ein Unix-Server mit einer relationalen Datenbank verwendet. Auf mehrere Tabellen verteilt enthält sie momentan ca. 150 Millionen Einträge (einschließlich dem Volltext-Index). Weit verbreitete Datenbanken aus dem PC-Bereich eignen sich nicht für die effiziente Verwaltung derartiger Datenmengen. Trotz der hohen Leistungsfähigkeit des eingesetzten Datenbankservers können Engpässe auftreten, wenn etwa mehrere Nutzer Anfragen starten, die eine vollständige sequentielle Suche in den Datenbanktabellen erfordern (z.B. eine Infixsuche in allen Einträgen, etwa nach **fff** , d.h. nach Wörtern mit drei *f* in der Mitte).

Neben der Datenverwaltung und -lieferung durch Datenbankserver und Webserver stellen sich Vielzahl konzeptueller und technischer Herausforderung im Vorfeld der Lexikonpräsentation, d.h. bei der Datenaufbereitung und -überarbeitung. Seit Beginn des Projekts hat sich eine heterogene Infrastruktur mit einer Vielzahl von Werkzeugen herausgebildet, die sich grob den folgenden Kategorien zuordnen lassen:

- Aufbereiten des Ausgangsmaterials (Satzsegmentierung; Einlesen in die Datenbank; Ermittlung neuer Einträge und Aktualisierung bestehender)
- Generierung zusätzlicher Information (morphologische Analyse; Zuordnung von Flexionsklassen)
- statistische Auswertung (v. a. Ermittlung von Kollokationen und Phrasen, Visualisierung)
- Optimierungsroutinen insbesondere zur Fehlererkennung (Tippfehler)

Daneben werden eine Reihe von Arbeitsschritten (noch) manuell durchgeführt, da sie kaum vollständig zu automatisieren sind wie etwa die Zusammenführung unterschiedlicher Ontologien (z. B. unterschiedliche Sachgebietsklassifikationen unterschiedlicher Lexika).

Ein großes Problem stellt die Analyse der großen Datenmengen an sich dar: Die große Zahl der Lexikoneinträge erlaubt zwar den Einsatz von Algorithmen mit linearer Komplexität, d. h. deren Laufzeit linear mit der Zahl der bearbeiteten Einträge wächst wie etwa Programme zur morphologische Zerlegung, die jedes Wort einzeln bearbeiten, schafft aber Probleme bei Algorithmen höherer (also z. B. quadratischer) Komplexität. Beispielsweise muß bei der Bestimmung der Kollokationen jedes in Frage kommende Wort*paar* untersucht werden, also jedes Wort mit jedem anderen verglichen werden. Dies ist aus Zeitgründen nur noch mit speziellen Algorithmen möglich, die alle notwendigen Daten komprimiert im Arbeitsspeicher halten. Bei der Kollokationsbestimmung konnte dadurch die Bearbeitungszeit von geschätzten *300 Jahren* bei der Arbeit innerhalb der Datenbank (d.h. mit vielen Festplattenzugriffen) auf wenige Stunden (mit eigener Datenstruktur im Hauptspeicher) gesenkt werden.

Die seit gut eineinhalb Jahren im World Wide Web verfügbare Abfrageschnittstelle hat bisher Zugriffszahlen mit einer konstanten Wachstumsrate von ca. 20% pro Monat; bei weiterer kontinuierlicher Steigerung kann auch die Last auf dem HTTP-Server zum Problem werden. Mittelfristig soll hier die Einführung einer verteilten Architektur, bei der die Datenbank auf einem Cluster von Workstations vorgehalten wird, Abhilfe schaffen.

5 Methodische Herausforderungen

Bei der klassischen Wörterbucherstellung erfolgt die Freigabe des Wörterbuches nach der erfolgreichen Bearbeitung sämtlicher Wörterbuchartikel zu einem festen Zeitpunkt. Dieses Vorgehen ist bei einem Umfang von mehreren Millionen Einträgen nicht möglich, da hier jeder vernünftige Zeitrahmen gesprengt würde. Allein die Endredaktion würde rund 20 Personenjahre benötigen, selbst wenn man eine Kapazität von 1000 Artikeln pro Person und Arbeitstag annimmt.

Als einzige Alternative stellt sich hier die laufende Publikation des Wörterbuchs in "unfertigem Zustand" dar, wobei gleichzeitig stetig an der Verbesserung gearbeitet wird. Dieses Vorgehen wird durch das Medium Internet unterstützt, da so jeweils der aktuelle Stand angeboten werden kann, ohne einzelne Stadien fixieren zu müssen, wie das bei Druckauflagen oder CD-ROMs geschieht. Es liegt dem Vorhaben

also eine grundsätzlich andere methodische Perspektive zugrunde: Es versteht sich als ein *Dauerprojekt*, das sich mit der quantitativen und qualitativen Optimierung der Datenbestände fortschreibt, ohne dabei einen fixierten Endpunkt aufweisen zu können oder wollen.

Für den Nutzer entsteht so der Nachteil, daß nicht sicher ist, daß bei einer späteren Abfrage die gleiche Antwort der Datenbank entsteht, wenn die Bearbeitung an dieser Stelle fortgeschritten ist. Außer offensichtlichen Fehlern, deren Beseitigung sicher ein Fortschritt ist, kann sich beispielsweise die Häufigkeit eines Wortes vergrößert haben, wenn mehr Text ausgewertet wurde. Man kann allerdings einwenden, daß das "organische Wachsen" des Projektes auch die Fortschreibung sprachlicher Wirklichkeit reflektiert bzw. sich kontinuierlich an diese annähert.

In Einzelfällen kann man solchen Nachteilen auch durch die Art der Datenauswertung abhelfen: Beispielsweise wird statt neben absoluten Häufigkeit eines Worts auch eine Häufigkeitsklasse ermittelt, die sich aus dem Verhältnis dieser Häufigkeit zur Häufigkeit des Artikels Çderë, des häufigsten Wortes im Deutschen, berechnet. Dieses Verhältnis ändert sich bei der Vergrößerung des Korpus kaum, so daß die Häufigkeitsklasse unverändert bleiben sollte.

Ein weiteres methodisches Problem ist die Qualitätssicherung für die eingelesenen Daten: Die heterogenen Quellen enthalten nicht nur unterschiedlich strukturierte Beschreibungen mit demselben Informationsgehalt (z. B. unterschiedliche Bezeichnungen für eine Sachgebietsangabe; unterschiedlich aufgebaute grammatikalische Angaben in den Eintragsstrukturen zweier Lexika), sondern auch fehlerhafte Daten in nennenswertem Umfang (z. B. Tippfehler in Zeitungstext). Um die Qualität des Wortschatz-Lexikons zu sichern bzw. zu optimieren, wird ein mehrschichtiges System der Qualitätssicherung für die Wortschatz-Einträge betrieben:

- Jedem Eintrag werden Qualitätsmarkierungen zugeordnet, die sich u. a. aus einer Überprüfung anhand als besonders hochwertig geltender Quellen ergibt.
- Anhand heuristischer Regeln können fehlerhafte Einträge ermittelt werden, wenn sie signifikant von den orthographischen Regeln abweichen.
- Für eine Untermenge des Wortschatzes erfolgt auch eine intellektuelle Nachbearbeitung, um einen Kernbestand an Einträgen mit besonders hoher Qualität zu erhalten.
- Unterschiedlich strukturierte Klassifikationssysteme (Sachgebietsangaben, Theasurusklassen, kontrollierte Beschreibungsvokabularien) werden mit geeigneten Softwarewerkzeugen intellektuell überarbeitet und zusammengeführt.

Auch die Qualitätssicherung ist dabei ein ständig fortlaufender Prozeß, da bei laufender Erweiterung des Datenbestands auch seine Optimierung fortzusetzen ist.

6 Erfahrungen

Das Wortschatz-Lexikon ist unter http://www.wortschatz.uni-leipzig.de oder http://wortschatz.uni-leipzig.de zugänglich und wird rege genutzt - momentan erfolgen täglich ca. 500 Zugriffe. Die folgenden Nutzungsinteressen lassen sich feststellen:

- Nachschlagen aus *Interesse*: Nachgeschlagen werden Wörter, die sonst schwer zu finden sind. Interesse gilt speziell den Beispielsätzen, die häufig auch einen unterhaltenden Wert besitzen. Häufig wird auch der eigene Name nachgeschlagen.
- Nachschlagen zur *inhaltlichen Erschließung*. Auch hier sind häufig die Beispielsätze nützlich, die den gesuchten Begriff erklären oder wenigstens in ein Umfeld einordnen.
- Nachschlagen des *Gebrauchs* bzw. der *Schreibweise*. Dies wird vorwiegend von Nicht-Muttersprachlern weltweit benutzt. Hier erweist sich auch eine Rechtschreibhilfe als nützlich, die bei Suche nach einem falsch geschriebenen Wort auf vorhandene, ähnlich geschriebene verweist.

Der teilweise Ersatz von Handarbeit durch Programme ermöglicht es, Angaben für eine große Anzahl von

Wörtern zu erzeugen. Die dadurch möglichen Fehler zwingen den Nutzer zu einem kritischeren Lesen als im klassischen Wörterbuch. Reaktionen der Nutzer zeigen aber, daß die Freude über gefundenes Material für seltene Wörter bei weitem überwiegt.

7 Fazit und Ausblick

Die praktischen Erfahrungen bei der Erstellung eines sehr großen Wörterbuchs zeigen, daß neue Methoden bei der Erstellung notwendig sind, die auch zu anderen Nutzungsschwerpunkten führt. Verfahren, die automatisch Angaben zu Wörtern generieren, müssen weiterentwickelt und sicherer gemacht werden.

Insgesamt ist es aber gelungen, nicht nur ein umfangreiches Vollformenlexikon der allgemeinen Nutzung zur Verfügung zu stellen, sondern dabei auch eine Aufbereitungsinfrastruktur zu schaffen, die sich relativ einfach auf andere Datenbestände übertragen läßt. Dies gilt sowohl für den Aufbau eine analog strukturierten Projekts für eine andere Datenbasis im Deutschen als auch für die Anwendung auf Texte anderer Sprachen: Derzeit entstehen eine Reihe weiterer monolingualer Datenbestände z. B. für das Englische, Niederländische und Französische, bei denen wir die bisher erarbeiteten Softwarewerkzeuge unter geringen Modifikationen wiederverwenden konnten.

Neben der Ausweitung linguistischer und statistischer Analyseverfahren und der quantitativen wie qualitativen Erweiterung des Datenbestands sollen mittelfristig folgende Ziele in Angriff genommen werden:

- Zusammenführung der parallelen monolingualen Datenbestände zu einem multilingualen Korpus, der z. B. für die maschinelle Übersetzung als Werkzeug genutzt werden kann. Für das Englische und Niederländischen liegen bereits analog strukturierte, aber derzeit noch weniger umfangreiche Korpora vor, wie die folgende Übersicht zeigt:

	deutsch	englisch	niederländisch
laufende Wörter	ca. 300 Mio.	ca. 123 Mio.	ca. 22 Mio.
Sätze	ca. 13,4 Mio.	ca. 6,4 Mio.	ca. 1,5 Mio.
Wortformen	ca. 6 Mio.	ca. 700.000	ca. 600.000

Datenbestände der monolingualen Wortschatz-Korpora

- Definition geeigneter Kodierungs- und Austauschformate unabhängig vom konkreten Speicherungsmodell auf der Basis flexibler Markupsprachen wie SGML (*standard generalized markup language*) bzw. XML (*extensible markup language*) (cf. Wolff / Quasthoff 1999)
- Erweiterung der Zugriffs- und Recherchemöglichkeiten des Systems für unterschiedliche Zielgruppen und Anwendungen.

Literatur

Heyer, Gerhard (1995): "Elements of a Natural Language Processing Technology." In: Heyer, Gerhard / Haugeneder, Hans (eds.) (1995): *Language Engineering. Essays in Theory and Practice of Applied Natural Langugae Computing.* Wiesbaden: 15-32.

Kamps, Thomas et al. (1999): "SGML für dynamische Publikationen: das Beispiel Fischer Weltalmanach." In: Möhr, Wiebke / Schmidt, Ingrid (eds.) (1999): *SGML und XML. Anwendungen und Perspektive.* Berlin etc.: 173-192.

Kim, Myoung-Cheol / ChoiI, Key-Sun (1999): "A Comparison of Collocation-Based Similarity Measures in Query Expansion." *Information Processing & Management* 35/1: 19-30.

Läuter, Martin / Quasthoff, Uwe (1999): "Kollokationen und semantisches Clustering." In: *Proc. GLDV-Jahrestagung 1999* [erscheint].

Lemnitzer, Lothar (1998): "Komplexe lexikalische Einheiten in Text und Lexikon." In: Heyer, Gerhard / Wolff, Christian (eds.): *Linguistik und neue Medien.* Wiesbaden: 85-91.

Quasthoff, Uwe (1998A). "Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values." In: *Proc. First International Conference on Language Resources & Evaluation, Granada, Mai 1998,* Bd. 2: 853-856.

Quasthoff, Uwe (1998B): "Projekt Der Deutsche Wortschatz." In: Heyer, Gerhard / Wolff, Christian (eds.). *Linguistik und neue Medien.* Wiesbaden: 93-99.

Wolff, Christian; Quasthoff, Uwe (1999): *LEMDIC. Leipzig Multilingual Dictionary Design.* Draft Technical Report, Leipzig University, CS Inst., NLP Dept., July 1999, Version 0.7.