

Projekt *Der Deutsche Wortschatz*

1. *Einleitung*
2. *Konzept der Sammlung*
3. *Erweiterte Information in der zentralen Sammlung*
4. *Nutzungsmöglichkeiten der Datenbank*
5. *Fehlerkorrektur in der Datenbank*
6. *Dienstprogramme*
7. *Ausblick*

1 Einleitung

Die Sammlung und Aufarbeitung lexikalischer Daten ist bei größeren Projekten häufig auf mehrere Mitwirkende verteilt, wichtiger Bestandteil ist aber auch die zentrale Koordinierung und Qualitätskontrolle der erfaßten Daten. Ziel des vorgestellten Projektes ist es, eine völlig neue Art einer „selbstorganisierenden“ Koordinierung zu erproben. Die Aufgabenstellung ist wegen des Wunsches nach einer großen Zahl von Mitwirkenden einfach gewählt: Ziel ist die Erfassung eines möglichst großen Teils des deutschen Fachwortschatzes. Die Erfassung erfolgt mit Hilfe einer vorgegebenen Software durch Mitarbeiter aus den verschiedensten Fachrichtungen auf freiwilliger Basis, die Qualitätskontrolle wird ebenfalls dezentral organisiert, indem bereits erfaßtes Material zur Diskussion gestellt wird und Änderungen vorgeschlagen werden können. Die Rolle der zentralen Koordinierung beschränkt sich dabei auf die (halbautomatische) Organisation des Informationsflusses und eine allgemeine Aufsichtsfunktion.

2 Konzept der Sammlung

Grundlage der vorliegenden Projekts ist eine Sammlung von Wortformen der deutschen Sprache, gesammelt auf der Grundlage umfangreicher elektronisch verfügbarer Textkorpora. Bisher liegen ca. 2,5 Millionen Wortformen vor, der ausgewertete Text bestand zum größten Teil aus Zeitungstexten, zu einem geringeren Teil auch aus Fachtexten oder speziellen Wortlisten. Dementsprechend ist der Wortschatz der geschriebenen Umgangssprache zu einem sehr großen Teil abgedeckt. Dagegen gibt es große Lücken im fachsprachlichen Bereich, die sich auch nicht ohne Änderung der Erfassungsmethode beseitigen lassen.

Deshalb soll mit dem Projekt ein völlig neuer Weg beschritten werden. Mit der Sammlung zusammen wird eine Software auf CD-ROM zur Verfügung gestellt, die es ermöglicht, aus einem vorgelegten Text die neuen, der Sammlung unbekannt Wörter zu ermitteln. Damit können dezentral Nutzer aus ihnen vorliegenden, aber möglicherweise nicht allgemein zugänglichen Texten diese neuen Wörter ermitteln, durch manuelle Kontrolle eventuelle Rechtschreibfehler beseitigen und die so entstandene

Wortliste zurücksenden, damit sie in die zentrale Sammlung aufgenommen werden kann.

Weiterhin soll der externe Nutzer die Möglichkeit haben, Wortformen zum Entfernen aus der zentralen Sammlung vorzuschlagen, um beispielsweise mit orthographischen Fehlern behaftete Einträge entfernen zu können. Ziel ist ein bewußter Umgang mit möglicherweise fehlerbehafteten Daten.

Ein weiterer Schwerpunkt besteht in der zentralen Bearbeitung der zurückgesandten Wortlisten. Auch hier soll der manuelle Aufwand minimiert werden, und es soll getestet werden, in wieweit die Verantwortung für die Qualität der Sammlung dezentral gelassen werden kann. Durch regelmäßige Updates erhalten die dezentralen Nutzer die aktualisierten zentralen Daten.

2.1 Zum dezentralen Sammeln

Das Vorgehen des Sammelns von Material durch Mitwirkende auf freiwilliger Basis ist nicht neu. Im Bereich der Lexikographie wurde beispielsweise Material für das Wörterbuch der obersächsischen Mundarten von zunächst 1600 Mitwirkenden zusammengetragen, von denen 400 zu einer längerfristigen Zusammenarbeit bereit waren (BERGMANN 1993:X).

Das folgende Beispiel mit wesentlich mehr mitwirkenden stammt aus der Biologie: Vor ca. zehn Jahren rief der britische Marienkäferforscher Michael MAJERUS seine Landsleute zum Beobachten dieser Insekten auf. 30.000 Teilnehmer sorgten für eine einmalige Feldstudie (vgl. MAJERUS 1994).

2.2 Warum die Sammlung von Vollformen?

Die Sammlung von Vollformen bringt gegenüber der Sammlung von Grundformen mehr Material in die Sammlung, das von einer gewissen Redundanz ist. Die Entscheidung zur Sammlung von Vollformen hat folgende Gründe:

- Bei Auswertung von Volltext ist die Sammlung von Vollformen relativ einfach, die zusätzliche Reduktion auf Grundformen ist eine mögliche Fehlerquelle.
- Die vorliegende Redundanz kann zur Fehlerkorrektur genutzt werden, wenn beispielsweise das orthographisch fehlerhafte Wort einzeln mehreren flektierten Formen des korrekten Wortes gegenübersteht.
- Aussagen über das Nichtvorkommen bestimmter flektierter Formen sind möglich.
- Speicherplatzprobleme treten bei der Vollformensammlung nicht auf, da als Distributionsmedium die CD-ROM zur Verfügung steht.

2.3 Zusammenwirken von dezentraler Erfassung und zentraler Verwaltung

Die folgende Übersicht zeigt die verschiedenen Arbeitsschritte sowohl bei den Nutzern als auch bei der halbautomatischen zentralen Lexikonverwaltung.

Projekt Deutscher Wortschatz

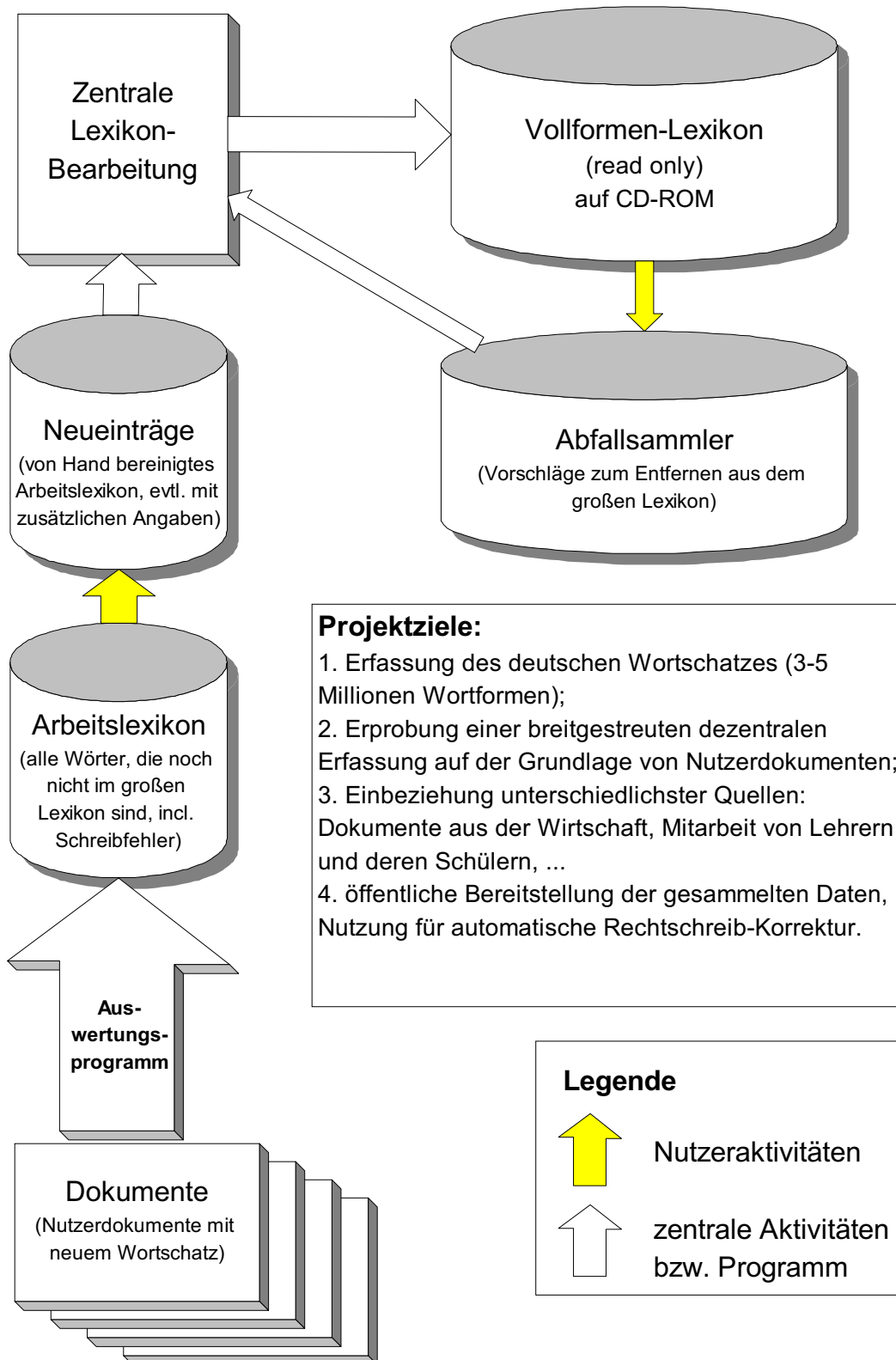


Abb. 1: Übersicht der Ablaufprozesse im Projekt Deutscher Wortschatz.

3 Erweiterte Information in der zentralen Sammlung

3.1 Struktur der zentral vorliegenden Daten aus der Textanalyse

Die zentral vorliegenden Daten enthalten mehr Informationen, als für die Erfassung neuer Wortformen notwendig ist (vgl. QUASTHOFF 1998). Die vorliegenden Daten haben folgende Form:

- **Frequenz:** Häufigkeit des Auftretens dieser Wortform. Diese so ermittelte Häufigkeit ist nur für hochfrequente Wörter zuverlässig, für niederfrequente Wörter ist sie stark abhängig von den ausgewerteten Korpora, da niemals ein repräsentativer Querschnitt „aller Texte“ vorliegen wird. Außerdem kann gegenwärtig die Häufigkeit bei den dezentral erfaßten Wortformen nicht berücksichtigt werden. Aussagekräftig ist also allenfalls die gemessene relative Häufigkeit der Wortformen.
- Wortform mit Information über Groß- / Kleinschreibung
- Weiterhin erfaßt werden ein Beispielsatz sowie Beispieltyp (z. B. Zeitungstext, Fachtext, ...). Damit lassen sich in vielen Fällen orthographische Fehler sicherer erkennen als nur mit der Wortform.

4 Nutzungsmöglichkeiten der Datenbank

4.1 (Halb-)Automatische Vervollständigung der Angaben

Im folgenden werden kurz einige Möglichkeiten skizziert, mittels automatischer Verfahren linguistische Informationen zu extrahieren.

4.1.1 Automatische Ergänzung der grammatischen Angaben, Verweis auf Grundform

Aus anderen maschinenlesbaren Wörterbüchern liegen Wortlisten mit dazugehörigen grammatischen Angaben vor. Ziel ist, aus Grammatikangaben für wenige Grundformen automatisch Grammatikangaben für möglichst viele Wörter zu erzeugen.

Es ist bis auf wenige Ausnahmen möglich,

- automatisch flektierten Formen den dazugehörigen Grundformen zuzuordnen, falls die Grundform in einer solchen zusätzlichen Liste ist,
- für Wortmengen, die aus Kandidaten für flektierte Formen einer Grundform bestehen, die Flexionsklasse der Grundform automatisch zu bestimmen sowie Verweise von den flektierten Formen auf die Grundform automatisch zu erzeugen.

4.1.2 Lexikonbasierte Kompositazerlegung

Eine lexikonbasierte Kompositazerlegung ist möglich, sobald die Bestandteile eines Kompositums im Lexikon vorhanden sind. Durch die Größe der vorliegenden Wortliste lassen sich zusätzliche Informationen wie z. B. Mengen anderer Komposita mit einem bestimmten Bestandteil an einer festen Position ermitteln. Damit lassen sich häufig Mehrdeutigkeiten bei der Kompositazerlegung auflösen, beispielsweise wird

für *Bundeswehreinheit* die Zerlegung *Bundesweh-Reinheit* unterdrückt, da es keine weiteren Komposita mit Kopf *Bundesweh-* gibt.

Die erfolgreiche Zerlegung eines Kompositums wiederum gestattet in den meisten Fällen die eindeutige Bestimmung der Flexionsklasse des Kompositums.

4.2 Praktische Anwendungen

4.2.1 Rechtschreibkontrolle

Die üblichen Verfahren zur Rechtschreibkontrolle bei Textverarbeitungsprogrammen basieren auf Wortlisten und funktionieren folgendermaßen: Ist ein Wort nicht in der Wortliste des Systems enthalten, so wird der Nutzer benachrichtigt und es werden mehr oder weniger passende Verbesserungsvorschläge gemacht. Wegen des geringen Umfangs der Wortlisten wird auch häufig bei korrekten Wortformen zurückgefragt. Dieses Verhalten kann mit einer umfangreicheren Wortliste verbessert werden. Momentan enthält die Wortliste etwa viermal so viele Einträge wie die Rechtschreibkontrollen der großen Textverarbeitungen. Bei einer angestrebten weiteren Vergrößerung des Wortschatzes und einer Überarbeitung für spezielle Zwecke der Rechtschreibkontrolle wird damit eine deutliche Überlegenheit erreicht.

4.3 Neue Abfragemöglichkeiten

Das Vorliegen lexikalischer Daten in elektronischer Form ermöglicht prinzipiell neue Abfragemöglichkeiten (vgl. BLÄSER & WERMKE 1990). Speziell mit den obengenannten Daten lassen sich folgende Abfragen realisieren:

- Welche flektierten Formen gehören zur gleichen Grundform wie eine bestimmte Wortform? In welchem Verhältnis stehen die dazugehörigen Frequenzen?
- Welche Komposita mit einem bestimmten Bestandteil (z. B. *wehr*) sind vorhanden?
- Welcher Fachbegriff (z. B. *Benutzeroberfläche* oder *Benutzungsoberfläche*) wird häufiger gebraucht?

5 Fehlerkorrektur in der Datenbank

5.1 Manuelle Korrektur durch verteilte Nutzer

Trotz großer Sorgfalt bei der automatischen Sammlung enthält die Wortliste ca. 1-2% fehlerbehaftete Einträge. Diese resultieren sowohl aus orthographischen Fehlern im Original als auch Fehlern bei der automatischen Übernahme aus dem vorliegenden maschinenlesbaren Text (z. B. Verwechslung von Bindestrich und Silbentrennung, falsche Interpretation von Sonderzeichen, abruptes Dateiende mitten im Wort, ...)

Weiterhin sind einige Fachbegriffe (z. B. *Betone* als Plural von *Beton*) nur den jeweiligen Experten bekannt, Nichtfachleute können hier Fehler vermuten. Dementsprechend muß bei der Auswertung von dezentral korrigiertem Material mit teilweise widersprüchlichen Angaben gerechnet werden. Da möglicherweise auch manuell eine korrekte Entscheidung ohne Fachleute kaum getroffen werden kann, bleibt nur die

Möglichkeit, solche Einträge als strittig zu markieren und als solche zur öffentlichen Diskussion zu stellen.

5.2 Automatische Korrekturmöglichkeiten

Durch die beim Sammeln zusätzlich erhaltenen Daten ergeben sich Möglichkeiten zur automatischen Korrektur von Fehlern in der Datensammlung wie oben beschrieben. Gleichzeitig lassen sich Informationen über die Häufigkeiten spezieller Fehler gewinnen (vgl. QUASTHOFF 1998).

6 Dienstprogramme

6.1 Textauswertung mit Satzsegmentierer

In der Wortliste sind die Wortformen nach Groß- und Kleinschreibung unterschieden. Da am Satzanfang keine Aussage über Groß- oder Kleinschreibung getroffen werden kann, ist es wichtig, diese Stellen zuverlässig zu erkennen. Deshalb werden bei der Auswertung eines Textes zunächst die Satzgrenzen ermittelt. Wörter am Satzanfang werden wegen der unklaren Groß- / Kleinschreibung nicht weiter berücksichtigt. Von den restlichen Wörtern wird geprüft, ob sie bereits in der Liste der bekannten Wörter vertreten sind. Darin nicht gefundene Wörter werden in ein Nutzerwörterbuch *Neue Vollformen* aufgenommen.

6.2 Die Wörterbuchverwaltung

Mit Hilfe der Wörterbuchverwaltung kann der Nutzer das von ihm erstellte Wörterbuch *Neue Vollformen* editieren und beispielsweise Wörter mit orthographischen Fehlern entfernen. Analog können fremdsprachige Ausdrücke oder nicht übliche Abkürzungen aussortiert werden. Weiterhin kann die Listenverwaltung zur Nachbearbeitung der Liste der bekannten Wörter verwendet werden. Diese Liste mit einem Umfang von über zwei Millionen Einträgen wird sicher immer Einträge enthalten, die offensichtlich fehlerhaft oder zumindest strittig sind. Mit dem Aussortieren dieser Einträge wird die Qualität der Grundliste ständig erhöht.

Die Nutzer sind anschließend aufgefordert, die von ihnen erzeugten Listen an die Zentralen Dienste per Diskette oder e-mail zurückzuschicken. Wünschenswert (aber nicht Bedingung) ist auch die Überlassung der ausgewerteten Texte, um für die zentrale Sammlung zusätzlich Beispielsätze zu den neuen Wörtern erzeugen zu können. Denkbar ist auch eine Sachgebietszuordnung auf der Grundlage der Texte.

6.3 Zentrale Dienste

Aufgabe der zentralen Dienste ist, die von den jeweiligen Nutzern zurückgesandten Wortlisten (also z. B. *Neue Vollformen* und *Schreibfehler*) in die zentrale Sammlung zu integrieren. Für *Neue Vollformen* werden zusätzlich Beispielsätze gespeichert, falls entsprechende Texte mitgeliefert werden. Die Behandlung fehlerhafter Einträge ist

komplizierter. Einmal kann ein Wort für fehlerhaft gehalten werden, obwohl es das nicht ist. Zum anderen wird im Falle eines häufigen Fehlers dieser wieder unbemerkt auftreten und so das fehlerhafte Wort wieder aufgenommen, anschließend wieder entfernt usw. Um dies zu vermeiden, wird eine Liste der strittigen Fälle geführt, in der alle Wertungen zu einem solchen Wort eingetragen und zu einem späteren Zeitpunkt für einen Entscheidungsvorschlag genutzt werden.

7 Ausblick

Das Projekt testet zunächst die Machbarkeit einer dezentralen selbstorganisierenden lexikalischen Sammlung. Dabei sollen Erfahrungen über die Bereitschaft zur Mitarbeit sowie die zu erwartende Qualität gesammelt werden. Bei positiven Ergebnissen ist es denkbar, im nächsten Schritt auch zusätzliche Angaben zu neuen Wörtern zu sammeln. Folgende Möglichkeiten bieten sich an:

- Ergänzung nicht vorhandener bzw. Kontrolle vorhandener Grammatik- und Sachgebietsangaben entsprechend einem vorgegebenen Schema
- Einordnen von Fachbegriffen in einen Thesaurus oder ein Klassifikationssystem analog (vgl. WEHRLE & EGGERS 1967, DORNSEIFF 1970, CHAPMAN 1993). Herstellen weiterer Beziehungen zwischen Wörtern entsprechend der lexikalischen Funktionen der *Meaning-Text Theory* (vgl. STEELE 1990).
- Seit 1998 ist die Datenbank im World Wide Web zugänglich (<http://wortschatz.uni-leipzig.de>).