Deutscher Wortschatz im Internet

U. Quasthoff
Institut für Informatik
Universität Leipzig
04109 Leipzig
quasthoff@informatik.uni-leipzig.de

Entstehung der Sammlung

Die mittlerweile vorliegende Datensammlung, die momentan sicher eine der umfangreichsten frei zugänglichen Datensammlungen zur deutschen Sprache ist, entstand Anfang der 90er Jahre mit einer Wortliste mit sporadisch vorhandenen Angaben zu Grammatik und Sachgebiet. Diese war angelegt worden, da damals ein allgemeiner Mangel an frei verfügbaren maschinenlesbaren Daten zum deutschen Wortschatz bestand. Anliegen der damals begonnenen Sammlung war und ist, verfügbare Daten zunächst zu sammeln und sie (sobald wie möglich) zu nutzen, um fehlende Angaben zu ergänzen und eventuelle Fehler zu beseitigen. Dazu bietet sich speziell die Redundanz an, die in einer so großen Sammlung zu finden ist. Hier erweist sich auch die Sammlung von Vollformen als günstig, da bei einer späteren Reduktion auf Grundformen die Menge der flektierten Formen zu einer Grundform die korrekte Erkennung dieser Grundform erleichtert. Ebenso läßt sich dann das korrekte Flexionsschema leichter erkennen bzw. überprüfen. Nachdem lange mit verschiedenen Wortlisten gearbeitet worden war, wurde 1994 auf ein relationales Datenbanksystem umgestellt, um eine einfachere Datenverwaltung und einheitliche Zugriffsmöglichkeiten zu bekommen.

Schnell stellte sich heraus, daß bei einigen so gesammelten Wörtern für den Betrachter im nachhinein nicht festgestellt werden kann, ob es sich um ein fehlerhaft geschriebenes Wort, einen Eigennamen oder vielleicht einen ihm unbekannten Fachbegriff handelt. Um solche Fragen wenigstens in der überwiegenden Mehrzahl der Fälle klären zu können, wurde ab ca. 1996 zusätzlich für jede neue Wortform ein Belegsatz gesammelt. Dazu wurde ein eigener Satzsegmentierer entwickelt (s.u.).

Das Vorhandensein der Beispielsätze wiederum ermöglicht Untersuchungen von Kollokationen. Erste Versuche sahen sehr erfolgversprechend aus, allerdings war das verwendete Material an Beispielsätzen nicht repräsentativ, da immer nur Sätze mit neuen Wörtern gesammelt wurden. Deshalb wurde 1998 dazu überge-

gangen, möglichst alle Sätze zu sammeln. Ein Volltext-Index ermöglicht effektives Suchen in diesen Sätzen.

Seit März 1998 ist die Datensammlung des Projekts Deutscher Wortschatz zu großen Teilen im Internet unter `http://wortschatz.uni-leipzig.de` zugänglich. Die Nutzerzahlen zeigen wachsendes Interesse, Anfang 1999 liegen die Zugriffszahlen bei 300–400 pro Tag.

Entsprechend ihrer Entstehung ist es nicht Anliegen der Sammlung, auf irgendeine Weise normierend auf den Sprachgebrauch einzuwirken. Zwar werden offensichtliche Fehler (z.B. orthographische Fehler oder Segmentierungsfehler, d.h. hauptsächlich Leerzeichen mitten im Wort oder fehlende Leerzeichen zwischen den Wörtern), wenn sie erkannt werden, als solche gekennzeichnet und die entsprechenden Wörter sind dann nicht mehr sichtbar, aber sonst wird in der Sammlung die deutsche Sprache dargestellt, wie sie in der Gegenwart maschinenlesbar zugänglich ist.

Vorliegendes Datenmaterial

Wortformen

Momentan liegen in der Datenbank ca. 5 Millionen Wortformen vor. Neben der absoluten Häufigkeit (d.h. der Anzahl, wie oft eine Wortform im Text gezählt wurde) ist eine Häufigkeitsklasse angegeben, welche diese Häufigkeit relativ zur absoluten Häufigkeit des häufigsten Wortes *der* mißt. Diese Häufigkeitsklasse bleibt bei der Auswertung weiterer Quellen (theoretisch, d.h. bei gleicher Zusammensetzung des Gesamtkorpus) unverändert, während sich die absoluten Häufigkeiten natürlich ändern.

Die Häufigkeitsklasse eines Wortes berechnet sich folgendermaßen: Sei n(*wort*) die absolute Häufigkeit des Wortes *wort*, n(‘*der*’) die absolute Häufigkeit des Wortes *der*. Die Häufigkeitsklasse von *wort* ist dann definiert als

$$h(wort) = \log_2 (n(\text{‘}der\text{’})/n(wort))$$

und wird auf die nächstgelegene ganze Zahl gerundet.

Die folgende Tabelle enthält die häufigsten 20 Wörter und ihre Häufigkeitsklassen. Die absoluten Zahlen stammen vom Juli 1998.

Wort	Anzahl	Häufigkeitsklasse
der	7507541	0
die	6836196	0
und	4965268	1
in	3850947	1
den	2758375	1
von	2232689	2
zu	2081975	2
das	1889278	2
mit	1843993	2
sich	1751629	2
nicht	1680592	2
des	1625102	2
ist	1603975	2
auf	1595642	2
für	1584354	2
im	1535076	2
dem	1463015	2
ein	1323984	2
eine	1229663	3
als	1081635	3

Niederfrequente Wörter haben folgende Häufigkeitsklassen:

Anzahl	Häufigkeitsklasse
12– ...	18
6–11	19
3–5	20
1–2	21

Grammatikangaben

Die Grammatikangaben unterscheiden sich für Grundformen und für flektierte Formen. Für flektierte Formen erfolgt nur ein Verweis auf die Grundform, während für die Grundform Wortklasse und Flexionstyp angegeben sind. Damit lassen sich wieder sämtliche flektierten Formen generieren, und für eine bestimmte flektierte Form lassen sich alle Ableitungen ermitteln. Weiterhin sind Eigennamen als solche markiert. Beispiele für die Verwendung als Eigennamen werden wie weiter unten beschrieben explizit angegeben.

Für ein Wort können mehrere Grammatikangaben vorhanden sein, z.B. verschiedenen Flexionsmuster für *Bank*.

Weiterhin enthält die Datenbank eine morphologische Analyse des Wortes. Diese baut auf dem Vorgehen von [WOT] auf und ermittelt Morphemgrenzen auf der Grundlage von Übergangswahrscheinlichkeiten, die aufbauend auf einer Ausgangsmenge von uns ermittelt wurden.

Sachgebiete

Die Klassifikation der Sachgebiete folgt im wesentlichen der Klassifikation der Schlagwort-Norm-Datei (SWD) der Deutschen Bibliothek, Frankfurt ([NOR]). Diese Sachgebietstitel sind hierarchisch angeordnet, so daß der *Bereich 27. Medizin* weiter untergliedert ist und z.B. die feineren Bereiche *27.9 Innere Medizin* und *27.9b Pulmonologie* enthält. Diese Struktur ermöglicht auch die Beibehaltung des Sachgebiets *Medizin* für Begriffe, deren genauere Einordnung uns momentan unbekannt ist.

Sachgebietsangaben aus anderen Quellen wurden eingearbeitet, so daß das gesamte Klassifikationsschema momentan ca. 1700 Sachgebietstitel enthält.

Beschreibung

Für einige Begriffe liegen kurze Beschreibungen vor, die keiner einheitlichen Form folgen.

Thesaurusrelationen

Häufig sind außer Sachgebietsangaben noch mehr inhaltliche Zusammenhänge zwischen Begriffen angegeben. Typischerweise findet man in Wörterbüchern zum einen Synonyme und gelegentlich Antonyme, zum anderen Hyponyme und Hyperonyme. In vielen Quellen ist für verschiedene Zwecke der Hinweis *vgl.* verwendet. Momentan sind diese Daten noch nicht befriedigend aufgearbeitet, vorgesehen ist aber eine schrittweise Erweiterung der Menge der Relationen. Zunächst sollen (ähnlich wie in WordNet, [FEL]) verschiedene Meronym-Relationen hinzugenommen werden, anschließend weitere Relationen, wie sie z.B. in der Meaning-Text-Theory ([STE]) vorgeschlagen werden. Kandidaten für Begriffspaare in solchen Relationen werden mit den weiter unten beschriebenen Kollokationen ermittelt.

Sätze

Die Beispielsätze erfüllen einen mehrfachen Zweck. Zunächst dienen sie im klassischen Sinn als Belegstellen. Weiterhin sind sie in vielen Fällen geeignet, zu entscheiden, ob es sich bei einem bestimmten Wort, welches nicht zum passiven Wortschatz des Benutzers gehört, um einen Fachbegriff, einen Eigennamen oder nur einen Schreibfehler handelt. Interessant ist, daß bei wirklich seltenen Wörtern der Autor damit rechnet, daß der Leser sie nicht kennt und deshalb ein Beispielsatz eine definitorische Erklärung enthält. Hier sei ein solcher Beispielsatz für Hunni angegeben: *Ein Hunderter ist immerhin noch ein Hunderter, wird aber, um ihn klein zu machen, seit ein paar Jahren Hunni genannt. (Quelle: Frankf. Rundschau 1992).*

Um nur möglichst wenig fehlerbehaftete Einträge zu erhalten, wurden von vornherein nur solche Quellen benutzt, die bereits selbst einen hohen Qualitätsstandard besitzen. Im wesentlichen wurden Materialien ausgewertet, die auf CD-ROM vorliegen und uns von den entsprechenden Verlagen für das Projekt zur Verfügung gestellt wurden. Durch die Zerlegung in einzelne Sätze und die unzusammenhängende Speicherung in der Datenbank konnte sichergestellt werden, daß die Originaltexte aus der Datenbank nicht rekonstruierbar sind.

Momentan stehen ca. 10 Millionen Beispielsätze zur Verfügung, die über einen Volltextindex erschlossen sind.

Kollokationen

Als Kollokationen wollen wir hier Paare von Wortformen bezeichnen, die signifikant häufig gemeinsam in einem Satz auftreten, bezogen auf die Menge der vorliegenden Beispielsätze. Unabhängig vom konkreten Verfahren werden solche Signifikanzen statistisch berechnet. Ein solches Verfahren gilt als gut, wenn es zu einem gegebenen Begriff möglichst viele Begriffe aus dem „semantischen Umfeld" liefert und nur wenige andere, „unpassende" Begriffe.

Das hier angewendete Signifikanz-Maß berechnet sich folgendermaßen:

Sei a die Anzahl der Sätze mit Wort A, b die Anzahl der Sätze mit Wort B, n die Gesamtanzahl aller Sätze und k die Anzahl der Sätze, welche die Wörter A und B enthalten. Wir setzen x=ab/n, und definieren:

Gilt (k+1)/x<0.1 (dies ist der typische Fall), so setzen wir

$$\mathrm{sig}(A,B)=x-k*\log x+\log k!\ .$$

Diese Formel leitet sich aus einem Ansatz her, der eine Poissonsche Verteilung nutzt und damit die Wahrscheinlichkeit des Ereignisses bestimmt, daß bei statistischer Unabhängigkeit die Wörter A und B insgesamt (mindestens) k-mal gemeinsam auftreten. In [LEM] sind einige weitere Verfahren zur Ermittlung von Signifikanzen vorgestellt.

Im Gegensatz zu anderen Arbeiten, bei denen Kollokationen eher exemplarisch für einige wenige Wörter ermittelt werden, sind hier die Kollokationen für alle Wörter ab einer gewissen Mindestfrequenz berechnet und abrufbar.

Außer Kollokationen auf Satzebene wurden weiterhin alle signifikanten rechten und linken Nachbarn ermittelt. Beispielsweise treten als signifikante rechte Nachbarn zum Wort *Insel* fast sämtliche Namen von Inseln auf, geordnet in einer Reihenfolge, die sich als Mischung von Bekanntheit und Nähe interpretieren läßt: *Sachalin, Rügen, Okinawa, Mindanao, Honshu, Hainan, Usedom, Shikoku, Java, Borneo, Texel, Kyushu, Lesbos, Sylt, Hokkaido, Skye, Mainau, Luzon, Sumatra, Borkum, Batam, Chios, Korfu, Elba, Tasmanien, Amrum, Rhodos, Herrenchiemsee, Cheju, Helgoland, Euböa, Flores, Sulawesi, Juist, Negros, Brac, Taipa, Mauritius, Oahu*.

Umfang der Daten

Die folgende Tabelle gibt einen Überblick über die Anzahl der verschiedenen Angaben.

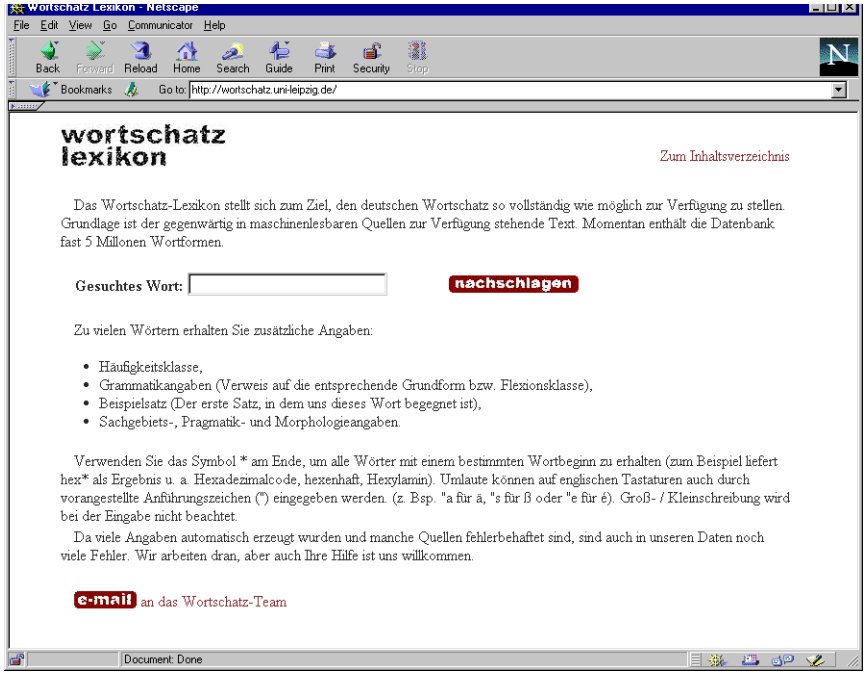
Art der Angabe	Anzahl der Einträge
Wortliste	4872716
Beispiele	7441775
Grammatik	2772369
Pragmatik	33948
Beschreibung	135914
Morphologie	3461877
Sachgebiet	1255813
Relationen	449619
Kollokationen (im Satz)	3154745
Kollokationen (Nachbarn)	709076
Volltext-Index	103730292

Möglichkeiten zur Suche über das Internet

Um die Nutzung der Datenbank von verschiedenen Plattformen aus und unabhängig vom Standort zu ermöglichen, wurde ein Zugang über das Internet ermöglicht. Verschiedene Abfragen ermöglichen unterschiedlich komplexe Aktionen.

Einfache Abfrage

Die einfache Anfrage ermöglicht das alphabetische Nachschlagen nach einzelnen Wörtern. Weiterhin können die üblichen Wildcards ? und * als Platzhalter für einen oder beliebig viele Zeichen benutzt werden. Die Benutzung dieser Wildcards ist an jeder Position möglich. Beispielsweise liefert die Suchanfrage nach *?i?i?i?i* den Eintrag *Minibikini*. Es ist allerdings zu beachten, daß eine Suche mit einer Wildcard am Anfang des Suchbegriffs eine sequentielle Suche durch die Wortliste erfordert, die mehrere Minuten dauern kann. Wird ein gesuchtes Wort (auf eine Anfrage ohne



Wildcards hin) nicht gefunden, so startet eine Suchhilfe und sucht nach Wörtern, die sich von der Eingabe auf genau eine der folgenden Arten unterscheiden darf:

- Austauschen eines Buchstabens gegen einen anderen,
- Hinzufügen eines Buchstabens,
- Weglassen eines Buchstabens,
- Vertauschen zweier benachbarter Buchstaben,
- Ersetzungen aus der folgenden Liste: ä↔ae, ö↔oe, ü↔ue, ß↔ss, f↔ph.

Besteht das Suchergebnis aus mehreren Wörtern, so wird zunächst eine Auswahlliste präsentiert, aus der ein bestimmtes Stichwort ausgewählt werden kann.

Gibt man als Suchbegriff beispielsweise *Gärtens* ein, so erhält man folgende Korrekturvorschläge: *Gärten, Gärens, Gättens, Gartens* und *Märtens*. Als Ergebnis werden die vorhandenen Angaben zu dem ausgewählten Wort angegeben, beispielhaft seien hier zwei Einträge für einen Eigennamen und ein flektiertes Verb angegeben:

```
Wort: Chemnitz:
Häufigkeitsklasse: 12 (Anzahl: 1207)
Beschreibung:
          Stadt in Deutschland (über 250000 E)
          Stadt in Sachsen
Sachgebiet:
          Nachname
          STADT
Grammatikangaben:
          Wortart: Eigenname
Beispiel:
Gegen die Konsumgenossenschaft Chemnitz ist auf Antrag
der Gewerkschaft HBV das Konkursverfahren eröffnet wor-
den. (Quelle: FAZ 1994)
Karin Chemnitz (Quelle: Telefonbuch)
Kollokationen im Satz: Zwickau, Dresden, Mosel, Leip-
zig, Sachsen, FC, sächsischen, Berlin, Riedel, Weißwas-
ser, Plauen, Drechsler
```

```
Wort: ermüdet:
Häufigkeitsklasse:     16 (Anzahl: 113)
Morphologie:           er|müd|et (+er=müd%et)
Grammatikangaben:Wortart: Verb
                       Stammform: ermüden
Relationen zu anderen Wörtern:
Vergleiche: matt, schlaff, schlapp
Synonyme:  kaputt, kraftlos, erschöpft
           für Synonyme siehe: müde, bettreif,
           ruhebedürftig, schlafbedürftig,
           schläfrig, schlaftrunken, übermüde
Beispiel:
Der Sachbuchautor Miloslav Stingl, Mitarbeiter des Mit-
teldeutschen Rundfunks, ist ein Kenner seiner Heimat-
stadt, schreibt aber einen korrekt-farblosen Stil, der
rasch ermüdet.
(Quelle: FAZ 1994)
```

Graphische Darstellung der Kollokationsmengen

Die zu einem Wort angegebenen typischen Kollokationen geben typischerweise verschiedene Zusammenhänge wieder. Unter den semantischen Zusammenhängen lassen sich folgende identifizieren:

- inhaltlich ähnliche Begriffe, sog. Kohyponyme (Schafe und Rinder zu Schweine);
- Kombinationen von Namen (Bill zu Clinton, Romeo zu Julia);
- Verb-Nomen-Kombinationen (beißen – Hund – Herrchen).

In einigen Fällen liefern die Kollokationsmengen auch Wörter in einer sprachlich ähnlichen Funktion, Beispiele dafür sind weiter unten angegeben.

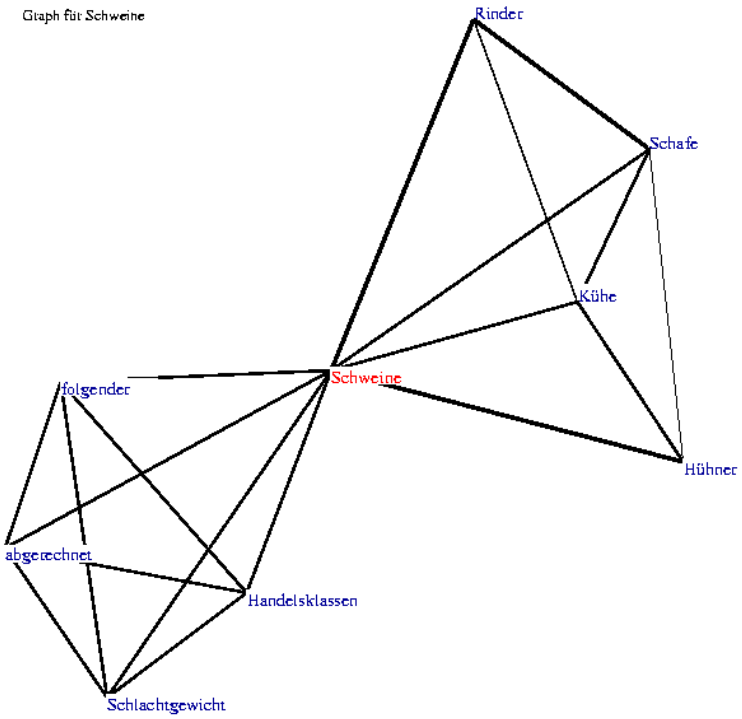
Um die Struktur zwischen den einzelnen Wörtern in einer Kollokationsmenge darzustellen, wurde eine Visualisierung als Graph gewählt, die auf dem Prinzip des Simulated Annealing beruht ([EAD, DAV]). Dazu werden die Begriffe als Knoten des Graphen dargestellt. Neben dem Ausgangsbegriff werden all die Begriffe aus der Kollokationsmenge mit aufgenommen, zwischen denen untereinander mindestens noch eine Verbindung besteht. D.h. zwei Begriffe A und B sind durch eine Kante verbunden, falls sig(A,B) einen Schwellwert übersteigt. Die Stärke der Verbindungslinie entspricht dabei der Größe dieses Signifikanzwertes, zu schwache Verbindungen werden der Übersichtlichkeit wegen nicht eingezeichnet.

Die Anordnung der Punkte erfolgt nach dem folgenden Schema: Zunächst werden die Punkte zufällig verteilt. Zwischen je zwei Punkten herrscht eine konstante Abstoßungskraft, zusätzlich herrscht zwischen zwei Punkten eine Anziehungskraft proportional sig(A,B). Ermöglicht man nun vorsichtig Bewegungen der Punkte entsprechend dieser Kräfte, so ergibt sich nach einiger Zeit eine stabile Anordnung, welche die inhaltlichen Zusammenhänge relativ gut zeigt.

Diese Kollokationsgraphen sind relativ aussagekräftig, wenn die absolute Häufigkeit des Ausgangsbegriffes eine Mindestgröße von ca. 500 übersteigt. Dies trifft gegenwärtig für ca. 26.000 Begriffe zu.

Im folgenden sind vier solche Graphen kurz beschrieben.

Kollokationen für *Schweine*

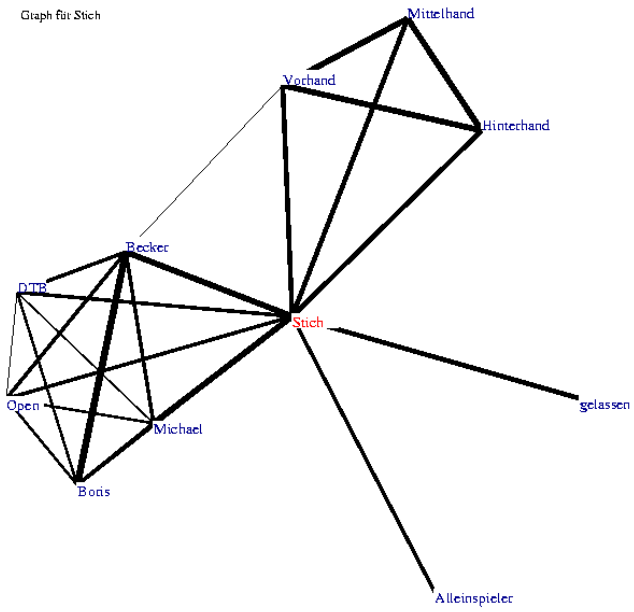


Der Graph für *Schweine* zerfällt sehr schön in zwei Teile. Im rechten Teil finden wir eine Zusammenstellung von anderen Nutztieren (ebenfalls im Plural), die untereinander stark zusammenhängen: Rinder, Hühner, Kühe, Schafe. Im linken Teil finden wir Wörter, die typische Sachverhalte zum Thema Schweine kennzeichnen, z.B.: Schweine werden nach Schlachtgewicht in Handelsklassen eingeteilt und abgerechnet.

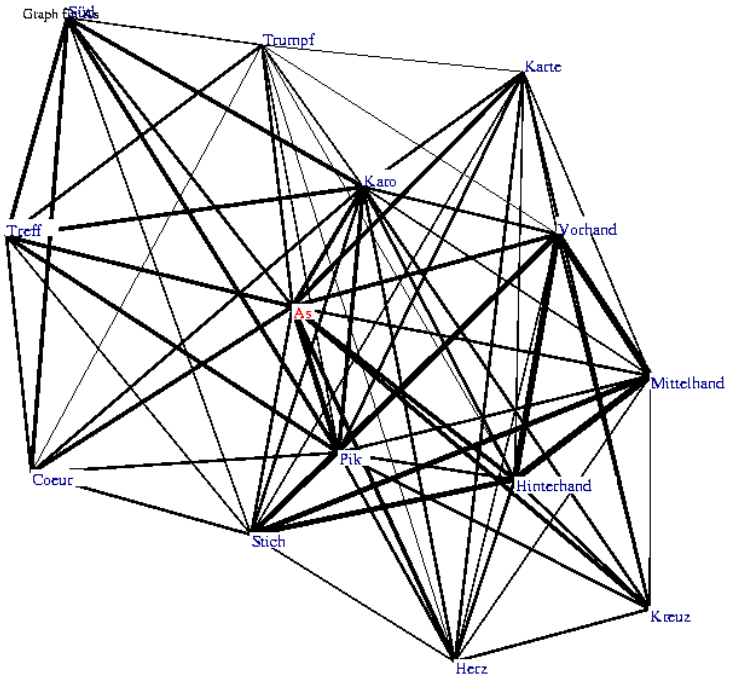
Kollokationen für *Stich*

Beim Graphen von *Stich* fallen uns zwei große Gruppen auf, die untereinander stark zusammenhängen: Einmal geht es um Tennis (Michael Stich usw.), einmal um das Skatspiel, bei dem die drei Mitspieler Vorhand, Mittelhand und Hinterhand heißen. Die zwei Teile entsprechen also verschiedenen Bedeutungen des Wortes Stich (einmal als Eigenname, speziell für den Tennisspieler Michael Stich, einmal um den Stich beim Kartenspiel). Interessant ist, daß diese beiden Gruppen auch noch über eine dünne Verbindung zwischen *Becker* und *Vorhand* zusammenhängen; diese Verbindung entsteht durch das starke Vorhandspiel von Boris Becker beim Tennis! Wir haben also auch bei *Vorhand* zwei Bedeutungen!

Die isolierte Verbindung von *Stich* und *gelassen* resultiert aus der Redewendung *im Stich gelassen.*

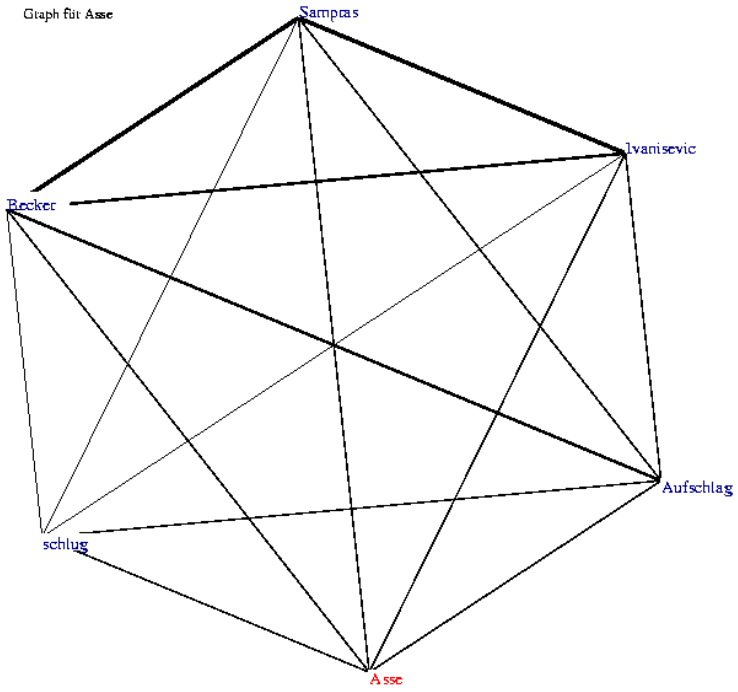


Kollokationen für *As*



Der Graph für *As* zeigt ausschließlich Begriffe aus dem Bereich Kartenspiel, und zwar auch nur Bezeichnungen für die Farben und die Mitspieler.

Kollokationen für *Asse*



Hier finden wir die Begriffe aus dem Bereich Tennis, die wir bei *As* vergeblich gesucht haben. Die Erklärung besteht einfach darin, daß in einem Bericht über Tennis offensichtlich nur eine große Anzahl von Assen erwähnenswert ist.

Diese sehr verschiedenen semantischen Umfelder für verschiedene flektierte Formen einer Grundform würden sehr verwischen, wenn wir sofort bei der Auswertung der Texte jedes Wort auf seine Grundform zurückführen würden. Erst

die Häufigkeitsangaben für die verschiedenen flektierten Formen ermöglichen das Erkennen der obigen Strukturen.

Spezielle Anfragen

Außer Anfragen, die wie oben von einem bestimmten Wort ausgehen, sind auch andere Anfragen realisierbar. Um den Nutzer nicht mit der Datenbanksprache SQL zu konfrontieren, wurden einige Anfragen vorbereitet, für die spezielle Eingabemasken vorliegen.

Folgende Anfragen werden momentan unterstützt:

- Suche nach Wörtern bestimmter, fester Länge;
- Suche nach Wörtern aus einem bestimmten Sachgebiet;
- Suche nach Wörtern mit mehreren Sachgebietsangaben;
- Suche nach frei wählbaren Zeichenketten in den Beispielsätzen;
- Suche nach frei wählbaren Zeichenketten in den Beschreibungen;
- Angabe der häufigsten Sachgebietsangaben, Grammatikangaben usw.;
- Durchschnitt zweier Kollokationsmengen.

Das letzte Beispiel ist folgendermaßen anwendbar: Beispielsweise enthält die Kollokationsmenge zu *Australien* neben anderen Ländern und australischen Persönlichkeiten auch die wichtigen australischen Städte. Weiterhin enthält die Kollokationsliste zu *Hauptstadt* die Namen vieler Hauptstädte. Der Durchschnitt beider Mengen (geordnet nach Signifikanz) enthält als erste Stadt tatsächlich *Canberra*, die australische Hauptstadt. Ähnlich kann man als Durchschnitt der Kollokationsmengen von *Rinder* und *Schweine* die Kohyponyme *Schafe, Kühe, Ziegen, Pferde, Kälber* in dieser Reihenfolge ermitteln, unterbrochen nur vom Begriff BSE.

Besteht die Ergebnismenge aus mehreren Wörtern, so sind verschiedene Anzeigemöglichkeiten auswählbar:

- alphabetisch sortiert
- rückläufig alphabetisch sortiert
- sortiert nach Häufigkeit
- sortiert nach Wortlänge
- Anzeige nur von Grundformen

Weitere Anfragen können bei Bedarf hinzugefügt werden.

Tools

Der Satzsegmentierer

Um Texte in einzelne Sätze zerlegen zu können, wurde ein eigener Satzsegmentierer entwickelt. Er hat folgende Eigenschaften:

- Eine sehr umfangreiche Liste der Abkürzungen, die mit Punkt beendet werden (als *Prof.*, aber nicht *BRD*), um solche Punkte nicht mit Satzenden zu verwechseln.
- Zusätzliche Zeichen sind als Satzende definierbar, z.B. kann eingestellt werden, ob `CRLF` als Satzende gewertet werden soll.
- Globale Ersetzungen für Zeichen sind möglich, dies ist nötig bei Sonderzeichen aus manchen Quellen.
- Die Zeichen, aus denen ein Wort bestehen darf, sind frei wählbar. Voreingestellt sind die Buchstaben a–z, ä, ö, ü, ß, é und das Zeichen - in der Wortmitte. Nicht als Wörter akzeptiert werden so *50%ig* und *Privileg(ium)*.
- Ebenso frei wählbar ist die Liste der Whitespace-Zeichen, also der Zeichen, die Wortzwischenräume kennzeichnen. Dazu kann man außer Leerzeichen, Tabulator, CRLF, und den Satzzeichen beispielsweise noch Klammern wählen. Dies sollte aber für jeden Text einzeln entschieden werden, da beispielsweise die Hinzunahme von Klammern aus *Privileg(ium)* fälschlicherweise das Wort *ium* separiert.

Weiterhin kann man dem Satzsegmentierer eine Liste bekannter Wörter mitgeben, dann erhält man als Ergebnis eine Liste der neuen Wörter mit jeweils einem Beispielsatz.

Der Satzsegmentierer steht frei zur Verfügung unter `http://wortschatz.informatik.uni-leipzig.de/Download.html` .

Umgang mit fehlerhaften Einträgen

Das Vorgehen, alle Einträge in der Datenbank per Hand zu kontrollieren und die falschen Einträge zu löschen, ist aus mindestens zwei Gründen nicht gangbar: Zunächst ist es aus Kapazitätsgründen unmöglich, alle Einträge per Hand zu bearbeiten. Weiterhin wird ein fehlerhaft geschriebenes Wort wieder automatisch aufgenommen, sobald es nach dem Löschen wieder in einer neuen Quelle gefun-

den wird. Deshalb werden in der Datenbank keine Einträge gelöscht, sondern mit Qualitätsangaben bewertet.

Die Wortformen können mit positiven und negativen Qualitätsangaben versehen werden. Positive Qualitätsangaben sind:

- Manuelle Klassifikation als korrekt.
- Autorisierung durch eine oder besser mehrere zuverlässige Quellen. Solche Quellen können sein:
- Wortlisten aus anderen Computerlexikographie-Projekten,
- Wortlisten aus Wörterbüchern,
- Wörter mit sehr großer Häufigkeit.
- Konsistenz zusätzlicher Angaben, z.B.:
- Eine vorhandene flektierte Form entspricht dem Flexionsmuster der entsprechenden Grundform.
- Das Flexionsmuster einer vorhandenen flektierten Form entspricht der morphologischen Zerlegung.

Negative Qualitätsangaben sind:

- Manuelle Klassifikation als fehlerhaft.
- Inkonsistenz zusätzlicher Angaben, z.B.
- Nicht passend zum angegebenen Flexionsmuster
- Schwierigkeiten bei der morphologischen Analyse
- Phonologisch inakzeptabel: Das Wort ist nicht aussprechbar.

Diese Kriterien ermöglichen eine teilweise automatische Überprüfung von Lexikoneinträgen.

Angabengewinnung und Fehlerbearbeitung

Angaben aus Kollokationslisten

Kollokationslisten bilden häufig einen sehr guten Ausgangspunkt, um eine Liste von Kandidaten für Wörter mit einer bestimmten Eigenschaft zu bekommen. Beispielsweise enthalten die ausgewerteten Texte gelegentlich englische Zitate, Titel von Musikstücken usw. Die darin enthaltenen englischen Wörter werden automatisch in die Datenbank aufgenommen und müssen von Hand als englisch-

sprachig markiert werden. Die Suche danach unter ca. 5 Millionen Wörtern gleicht der Suche nach der Nadel im Heuhaufen. Kollokationsmengen helfen hier, Kandidaten zu sammeln.

Beispielsweise liefert die Kollokationsliste für *the*:

> of, and, to, The, on, for, is, from, you, with, that, it, world, are, be, not, We, at, World, we, have, this, by, they, when, You, can, When, into, what, your, or, But, time, And, like, over, Breaking, only, one, but, shall, which, has, What, road, as, On, same, people, out, our, This, It, way, best, who, no, my, more, his, up, their, ...

Ähnlich lassen sich Ausdrücke in Mundart identifizieren.

Die Kollokationsliste für *ick* liefert:

> det, nich, Ick, Det, hab, is, ne, ooch, keene, wat, weeß, uff, de, ma, nu, keen, dat, aba, och, jing, jut, Nee, meen, Jöre, een, mach, inne, watt, wa, jenuch, kieke, janze, kumm, janz, tau, Mutta, janzen, hätt, sag, wieda, kleene, ha, hör, imma, un, habense, kriejen, ejal, zwee, nischt, nee, Wetta, jedacht, hebb, heff, ...

Die Kollokationsliste für *ned* liefert:

> Des, is, woaß, ma, hod, mi, wos, amoi, Bua, scho, aa, kumma, wia, woin, kriagst, hätt, gwesn, ka, koa, derf, ko, oda, muaß, san, Erd, spuin, de, gsund, liaba, woas, kan, machn, frog, skandiert, warn, nu, fei, ausm, wär, Naa, mar, koan, falschrum, sogn, Hinterlechner, hudeln, oiß, ...

Die Kollokationsliste für *een* liefert:

> de, un, dat, as, op, anner, se, vun, ok, speelt, för, düsse, is, wat, nich, mol, ick, det, na, ward, Dat, sick, deit, Se, Tied, he, hett, ut, gifft, dor, ne, to, wa, As, aba, nu, jedeen, Keerl, snackt, Nootebooms, tein, sitt, eegens, grote, uff, ...

Solche Kollokationslisten enthalten natürlich nicht nur Wörter der gesuchten Art, sind aber zur manuellen Bearbeitung viel besser geeignet als die gesamte Wortliste.

Ableitung von Angaben aus Wortenden

Bei einem hinreichend großen Grundwortschatz sind neue Wörter häufig Komposita, deren letzter Bestandteil bereits bekannt ist. Damit lassen sich aus dem Wortende häufig sowohl grammatische als auch semantische Angaben ableiten. Das soll am folgenden Beispiel demonstriert werden: Zunächst hat es den Anschein, als ob folgende Regel gilt:

> Alle Nomina auf –ung sind weiblich sind und gehorchen dem gleichen Flexionsmuster.

Aber es gibt Ausnahmen, nämlich Dung und Schwung und die dazugehörigen Komposita. Also wird unsere Regel verfeinert zu:

> Alle Nomina auf –ung, außer –dung und -wung sind weiblich sind und gehorchen dem gleichen Flexionsmuster. Die Fälle –dung und –wung müssen einzeln weiter untersucht werden.

Um dieses Vorgehen zu automatisieren, wurde die Lernstrategie von Liang ([LIA]) beim TEX-Silbentrenner so verändert, daß sie beliebige Angaben aus Wortenden lernt.

Literatur

[EAD] Eades, Peter, 1984. A Heuristic for Graph Drawing, Congressus Numerantium, vol. 42, pp. 149–160.

[DAV] Davidson R. and D. Harel: Drawing Graphs Nicely Using Simulated Annealing, ACM Transaction on Graphics, Vol. 15, No. 4, 1996, Seiten 301–331.

[FEL] Fellbaum, Chr. (ed.)WordNet: An Electronic Lexical Database, MIT Press 1998

[LEM] Lemnitzer, L.: Komplexe lexikalische Einheiten in Text und Lexikon; in: Heyer, G., Wolff, Ch.: Linguistik und neue Medien, Wiesbaden, Dt. Universitätsverlag, 1998.

[LIA] Liang, F. M.: Word hy-phen-a-tion by com-pu-ter. Technical Report STAN-CS-83-977, Stanford University, August 1983.

[NOR] Normdaten-CD-ROM, Die Deutsche Bibliothek Frankfurt / Main, 1998.

[QUA] Quasthoff, Uwe: Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values. In: Proceedings of the first International Conference on Language Resources & Evaluation, ELRA 1998, S. 853–856.

[STE] Steele, J.: Meaning-text theory: linguistics, lexicography and implications, University of Ottawa Press; Ottawa, London, Paris, 1990

[WOT] Wothke, K. (1993). Morphologically based automatic phonetic transcription, IBM Systems Journal 32, S. 485–511.