Häufigkeitsverteilung deutscher Morpheme

Roland Hausser
Universität Erlangen-Nürnberg
Abteilung Computerlinguistik (CLUE)
rrh@linguistik.uni-erlangen.de

Abstract

Bisher bezogen sich Angaben zum Wortschatz einer Sprache meist auf Wortformen und basierten auf Korpora, die möglichst balanciert und repräsentativ sein sollten. Die vorliegende Untersuchung betrachtet neben der Verteilung der Wortformen auch die der Morpheme und Allomorphe, basierend auf einer regelgesteuerten automatischen Wortformerkennung (DMM). Die Morphemverteilung in einem klassischen Korpus wird mit der in domänenspezifischen Korpora verglichen.

1. Desiderata der Korpuskonstruktion

Bereits im Jahre 1897–98 präsentierte Wilhelm Kaeding die erste umfassende Häufigkeitsuntersuchung von Wortformen für das Deutsche, und zwar als statistische Grundlage zur Verbesserung der Stenographie.[1] Mit seiner kleinen Armee von ‚Zählern' untersuchte Kaeding fast 11 Millionen laufende Wortformen (= 20 Millionen Silben) mit 250 178 verschiedenen Wortformen (Types).

Damit war die von Kaeding verwendete Textmenge mehr als zehnmal so groß und die resultierende Menge der Types mehr als doppelt so groß wie die des computerbasierten Limas-Korpus von 1973. Im Gegensatz zu heutigen Korpora handelt es sich bei Kaedings Textsammlung allerdings nicht um ein streng synchrones Korpus, denn die von ihm ausgewählten Beispiele decken den Zeitraum von ca. 1750 bis 1890 ab.

Mit der Verfügbarkeit von Computern erhielten Untersuchungen dieser Art neue Impulse, wobei Kučera und Francis 1967 für amerikanisches Englisch mit dem Brown-Korpus[2] den Anfang machten. Das Brown-Korpus umfaßt 500 Texte mit 1 014 231 laufenden Wortformen (Tokens) und 50 406 verschiedenen Wortformen (Types).

1968 folgte das LOB-Korpus[3] als Pendant zum Brown-Korpus für britisches

[1]Siehe Meier 1964.

[2]Benannt nach der Brown University in Rhode Island, an der Francis lehrte.

Englisch, ebenfalls mit 500 Texten, ca. einer Million Tokens und 50 000 Types. Beide Korpora wurden aus Texten der folgenden 15 *Genres* zusammengestellt.

		Brown	LOB
A	Presse: Reportagen	44	44
B	Presse: Kommentare	27	27
C	Presse: Rezensionen	17	17
D	Religion	17	17
E	Handwerk, Handel und Freizeit	36	38
F	Trivialliteratur	48	44
G	Literatur, Biographien, Essay	75	77
H	Regierungsdokumente etc.	30	38
J	Geistes- und naturwissenschaftliche Schriften	80	80
K	Erzählungen allgemein	29	29
L	Kriminalromane	24	24
M	Science-fiction	6	6
N	Abenteuer- und Wildwestgeschichten	29	29
P	Liebesromane	29	29
R	Humor	9	9
Gesamt		500	500

1.1 Die 15 Genres des Brown- und des LOB-Korpus

Die Zahlen besagen, wieviele Texte aus dem jeweiligen Genre in das Korpus aufgenommen wurden – wobei leichte Differenzen zwischen dem Brown- und dem LOB-Korpus festzustellen sind.

Für den Bau des Brown-Korpus formulierten Kučera und Francis 1967, S. xviii, folgende Desiderata:

> **1.** Exakte Spezifikation der verwendeten Sprachtexte, so daß sich die Benutzer einen genauen Begriff von der Zusammensetzung des Materials machen können.
>
> **2.** Vollständige Synchronizität: Nur Texte aus einem einzigen Kalenderjahr werden verwendet.

[3]Das *Lancaster-Oslo/Bergen*-Korpus wurde unter der Leitung von Geoffrey N. Leech und Stig Johannson angelegt. Siehe Hofland & Johannson 1982.

3. Die verschiedenen Genres werden in einem vorgegebenen Größenverhältnis zueinander gefüllt, wobei die individuellen Textbeispiele nach dem Zufallsprinzip ausgewählt werden *(random sampling)*.

4. Formale Spezifikation der im Korpus enthaltenen Informationen und automatischer Zugriff auf sie.

5. Genaue und vollständige Beschreibung der elementaren statistischen Eigenschaften des Korpus und seiner Komponenten (Genres), mit der Möglichkeit, die Analyse auf Erweiterungen des Korpus auszudehnen.

1.2 Desiderata der Korpuskonstruktion

Diese Ansprüche werden umgesetzt mit den mathematischen Methoden der Statistik, also den Grundgleichungen der Stochastik, Verteilungen für unabhängige und abhängige Häufigkeiten, Normalisierung, Fehlerberechnung etc. Dabei wird versucht, eine theoretische Verteilung zu finden, der die empirische Verteilung entspricht, insbesondere bei beliebig wachsender Datenmenge (Konstanz der empirischen Verteilungsverhältnisse).

Die deutsche Entsprechung zum amerikanischen Brown-Korpus (1967) und dem britischen LOB-Korpus (1968) ist das Limas-Korpus (1973).[4] Wie seine englischen Pendants besteht es aus 500 Texten von jeweils ca. 2 000 laufenden Wortformen. Insgesamt enthält das Limas-Korpus 1 062 624 laufende Wortformen. Aufgrund der reicheren Morphologie des Deutschen ist die Zahl der Types mit 110 837 verschiedenen Wortformen jedoch mehr als doppelt so groß wie bei den englischen Korpora.

2 Auswahl der Genres

Die Auswahl der Genres und die Festlegung ihrer unterschiedlichen Größe haben das Ziel, ein Korpus möglichst *repräsentativ* für die Sprache seiner Zeit zu machen und die verschiedenen Genres möglichst *balanciert* zu vertreten.[5] Intuitiv sind diese Begriffe leicht verständlich. So ist z. B. der Jahrgang einer Tageszeitung repräsentativer für eine natürliche Sprache als die Summe der Telefonbücher oder die Kontoauszüge einer Bank. Und ein Korpus, das Texte aus den verschiedenen Genres in den Verhältnissen von 1.1 enthält, ist besser balanciert als eines, das nur aus Texten eines einzigen Genres besteht.

---

[4]Siehe Hess, K., J. Brustkern & W. Lenders 1983.

[5]Siehe Bergenholz 1989, Biber 1994, Oostdijk & de Haan (Hg.) 1994.

Dennoch ist es schwierig, die für ein Korpus gewählte Zusammensetzung als repräsentativ und balanciert zu *beweisen*. Oostdijk 1988 kritisiert z. B. an 1.1:

> *[...] 'genre' is not a well-defined concept. Thus genres that have been distinguished so far have been identified on a purely intuitive basis. No empirical evidence has been provided for any of the genre distinctions that have been made [...].*

Für ein wirklich repräsentatives und balanciertes Korpus ist letztlich erforderlich, daß man weiß, welche Genres wie oft in einem gegebenen Zeitraum von der Sprachgemeinschaft gesprochen, geschrieben, gehört und gelesen wurden. Da es praktisch unmöglich ist, das Verhältnis zwischen Produktion und Rezeption sowohl gesprochener als auch geschriebener Sprache in sämtlichen Genres realistisch zu quantifizieren, ist der Bau repräsentativer und balancierter Korpora naturgemäß mehr das Ergebnis einer Kunst als einer Wissenschaft. Es beruht weitgehend auf allgemeinen ‚common sense'-Überlegungen und hängt zudem von dem intendierten Zweck des Korpus ab.

Inzwischen werden Korpora wie das Brown-, LOB- und Limas-Korpus im Umfang von jeweils 1 Million laufender Wortformen als wesentlich zu klein für die Erstellung aussagekräftiger Statistiken im Bereich der natürlichen Sprachen angesehen. Deshalb wurde für das britische Englisch das British National Corpus (BNC) mit 100 Millionen laufenden Wortformen zusammengestellt. Davon sind 89,7 Millionen aus dem Bereich geschriebener Sprache und 10,34 Millionen aus dem Bereich der gesprochener Sprache. Der Bereich der geschriebenen Sprache umfaßt 659 270 Types[6].

3 Auswertung von Korpora

Der Wert eines Korpus liegt nicht in dem Inhalt seiner Texte, sondern in seiner Eigenschaft als reale Stichprobe einer natürlichen Sprache. Je repräsentativer und balancierter diese Stichprobe ist, desto wertvoller ist das Korpus – zum Beispiel für eine realistische Berechnung der statistischen *Häufigkeitsverteilung* der Wortformen.

Auf der elementarsten Ebene wird diese statistische Auswertung als eine *frequenzbasierte* und eine *alphabetische* Wortliste dargestellt (als Beispiele siehe 4.1 und 4.3). In der frequenzbasierten Wortformenliste werden die Types in der

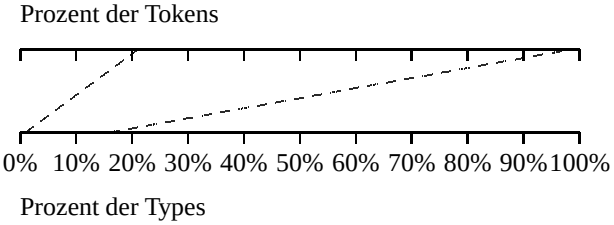
[6]Diese Zahl bezieht sich auf die reinen Oberflächen der Wortformen. Die Autoren des BNC verwenden dagegen Zahlen, die auf getaggten Wortformen beruhen. Nach letzterer Zählweise enthält das BNC 921 073 Types.

Reihenfolge ihrer (Token-) Häufigkeit aufgelistet. Der Platz einer Wortform in dieser Reihenfolge wird der *Rang* der Wortform genannt.

Am Anfang der Frequenzliste des BNC steht zum Beispiel die Wortform the, die mit 5 776 399 Vorkommen 6,4 % der laufenden Wortformen (Tokens) ausmacht. Am unteren Ende der Frequenzliste stehen die Wortformen, die jeweils nur ein einziges Mal im Korpus vorkommen. Von diesen sogenannten *Hapax Legomena*[7] gibt es 348 145 – was 52,8 % der Types im BNC ausmacht.

Wenn wir die ersten neun Ränge betrachten, so umfassen sie 0,001368 % der Types, decken aber 21,895 % der laufenden Wortformen im BNC ab. Am anderen Ende der Skala liegen die Ränge, deren Wortformen nur ein- bis neunmal im BNC vorkommen. Sie umfassen 83,6 % der Types, entsprechen aber zusammen nur 1,2 % der laufenden Wortformen.

Mit anderen Worten, 16,4 % der Types im BNC genügen, um 98,8 % der laufenden Wortformen zu erfassen. Für das verbleibende 1,2 % der laufenden Wortformen werden 83,6 % der Types im BNC benötigt. Das entspricht dem Intervall zwischen Rang 659 270 und 108 155, also insgesamt 551 115 Types. Diese Korrelation von Type- und Token-Häufigkeit findet sich ganz allgemein in ausreichend großen Korpora natürlicher Sprachen. Sie ist in 3.1 noch einmal graphisch dargestellt.



3.1 Korrelation von Type und Token-Häufigkeit

Die Tatsache, daß 0,001 % der Types fast 22 % der laufenden Wortformen in einem Korpus abdecken, und 16 % der Types über 98 % der laufenden Wortformen, wird manchmal dahingehend mißverstanden, daß kleine Lexika, wie z. B. die heutiger Spracherkennungssysteme mit nur 1 000 Wortformen, für die meisten praktischen Belange vollkommen genügen würden. Dies ist jedoch insofern ein

[7]Aus dem Griechischen, „einmal gesagte".

schwerwiegender Irrtum, als die *semantische Signifikanz*[8] mit abnehmender Frequenz einer Wortform steigt.

So nützt es dem Benutzer herzlich wenig, wenn das System zwar der, ist, mit und zu, nicht aber signifikante Hapax Legomena wie z. B. Abbremsung, Babyflaschen oder Campingplatz versteht,[9] weil es in seinem Lexikon keinen Eintrag dafür gibt. Für das BNC gilt entsprechendes: Unter seinen Hapax Legomena finden sich z. B. audiophile, butternut, customhouse, dustheap, um nur wenige Beispiele zu nennen, die alle in einem traditionellen Lexikon wie z. B. dem Webster's New Collegiate Dictionary aufgeführt sind und dort lexikalisch beschrieben werden.

Darüber hinaus gibt es viele Wörter im Webster's, die im BNC kein einziges Mal belegt sind, z. B. *aspheric*, *bipropellant*, *dynamotor* – trotz seiner Größe und des Bemühens um ein repräsentatives, balanciertes Korpus. Somit kann die Type-Liste eines großen Korpus zwar helfen, ein traditionelles Lexikon zu ergänzen. Es ist jedoch nicht zu erwarten, daß sich ein großes Korpus als ebenso vollständig wie oder vollständiger als ein traditionelles Lexikon erweist.

4 Statistisches Tagging

Da für die Korpusanalyse anfänglich noch keine Systeme der automatischen Wortformerkennung mit ausreichender Datenabdeckung existierten, konzentrierte man sich zunächst auf eine rein buchstabenbasierte statistische Analyse. Sie resultiert in Wortformenlisten, welche die Häufigkeiten in bezug auf das Gesamtkorpus und meist auch in bezug auf die einzelnen Genres angeben. Dies illustriert z. B. die Rangliste des Brown-Korpus nach Kučera und Francis, deren Anfang in 4.1 wiedergegeben ist.

Dabei bedeutet z. B. der Eintrag `9543-15-428 HE`, daß die Form HE insgesamt 9 543 mal vorkommt, und zwar in allen 15 Genres, aber nur in 428 der insgesamt 500 Einzeltextproben.

Was in 4.1 aus linguistischer Sicht fehlt, sind grammatische Informationen, insbesondere Bestimmung (i) der Wortart und der Flexionsform (Kategorisierung) und (ii) der Grundform (Lemmatisierung). Um diese Lücke wenigstens teilweise zu schließen, entwickelte N. W. Francis 1980 das System TAGGIT, eine musterbasierte Methode der Kategorisierung, die eine starke manuelle Nachbearbeitung erfor-

[8] Siehe Zip 1932.
[9] Diese Beispiele sind dem Limas-Korpus entnommen.

```
69971-15-500 THE          21341-15-500 IN
36411-15-500 OF           10595-15-500 THAT
28852-15-500 AND          10099-15-485 IS
26149-15-500 TO            9816-15-466 WAS
23237-15-500 A             9543-15-428 HE
```

4.1 Anfang der Frequenzliste im Brown-Korpus

derte. Darauf aufbauend[10] entwickelten Garside, Leech & Sampson 1987 das CLAWS1-System, das versucht, die Kategorisierung aus der Häufigkeitsverteilung der Wortformen zu erschließen. Dieses statistische *tagging* wurde u. a. entwickelt, um schnellere und bessere Ergebnisse bei großen Korpora zu erzielen als mit *pattern matching*.

Statistisches Tagging hat inzwischen große Verbreitung gefunden. Es basiert darauf, daß zunächst die Wortformen eines kleinen Teilkorpus *(core corpus)* in Handarbeit kategorisiert werden – oder eine halbautomatische Kategorisierung zumindest nachträglich sorgfältig ediert und korrigiert wird. Nach dem manuellen Tagging des Teilkorpus werden mit Hilfe von *Hidden Markov Models* (HMMs) die Wahrscheinlichkeiten der Übergänge von einem Tag zum nächsten berechnet. Dann werden die Wahrscheinlichkeiten des manuell getaggten Teilkorpus unter Verwendung eines vereinfachten Tagsets auf das Gesamtkorpus übertragen. Im BNC umfaßt dieses sogenannte *basic (C5) tagset* 61 *labels*.

AJ0	Adjective (general or positive) (e.g. good, old, beautiful)
CRD	Cardinal number (e.g., one, 3, fifty-five, 3609)
NN0	Common noun, neutral for number (e.g. aircraft, data, committee)
NN1	Singular common noun (e.g. pencil, goose, time, revelation)
NN2	Plural common noun (e.g. pencils, geese, times, revelations)
NP0	Proper noun (e. g. London, Michael, Mars, IBM)
UNC	Unclassified items
VVB	The finite base form of lexical verbs (e. g. forget, send, live,return)
VVD	The past tense form of lexical verbs (e. g. forgot, sent, lived, returned)

[10]Siehe Marshall 1987, S. 43 – 5.

[11]Die Verwendungvon HMMs für das grammatische Tagging von Korpora wird z. B. in Leech, Garside & Atwell 1983, Marshall 1983, de Rose 1988, Sharman 1990, Brown, P., V. Della Pietra et al. 1991 beschrieben. Siehe auch K. Church & Mercer 1993.

VVG	The -ing form of lexical verbs (e. g. forgetting, sending, living, returning)
VVI	The infinitive form of lexical verbs (e. g. forget, send, live, return)
VVN	The past participle form of lexical verbs (e. g. forgotten, sent, lived, returned)
VVZ	The -s form of lexical verbs (e. g. forgets, sends, lives, returns)

4.2 Teilmenge des basic *(C5)* tagset

Nachdem das gesamte Korpus auf diese Weise getaggt ist, können – statt reiner Oberflächen – getaggte Wortformen für die Häufigkeitsanalyse verwendet werden. Dabei werden Oberflächen mit verschiedenen Tags als verschiedene Types behandelt. Dies illustriert das folgende Beispiel, das als zufällige Stichprobe aus der getaggten BNC-Liste entnommen wurde, die online zur Verfügung stand.

```
1 activ nn1-np0 1          8 activating aj0-nn1 6
1 activ np0 1              47 activating aj0-vvg 22
2 activa nn1 1             3 activating nn1-vvg 3
3 activa nn1-np0 1         14 activating np0 5
4 activa np0 2             371 activating vvg 49
1 activatd nn1-vvb 1       538 activation nn1 93
21 activate np0 4          3 activation nn1-np0 3
62 activate vvb 42         2 activation-energy aj0 1
219 activate vvi 116       1 activation-inhibition aj0 1
140 activated aj0 48       1 activation-synthesis aj0 1
56 activated aj0-vvd 26    1 activation. nn0 1
52 activated aj0-vvn 34    1 activation/ unc 1
5 activated np0 3          282 activator nn1 30
85 activated vvd 56        6 activator nn1-np0 3
43 activated vvd-vvn 36    1 activator/ unc 1
312 activated vvn 144      1 activator/ unc 1
1 activatedness nn1 1      7 activator/tissue unc 1
88 activates vvz 60        61 activators nn2 18
5 activating aj0 5         1 activators np0 1
```

4.3 Alphabetische Wortformenliste (Stichprobe BNC)

Jeder Eintrag in 4.3[12] besteht erstens aus einer Zahl, welche die Häufigkeit im Gesamtkorpus angibt, zweitens aus der Oberfläche der Wortform, drittens dem Label und viertens der Zahl der Teilkorpora, in denen die Wortform in der angegeben Kategorisierung gefunden wurde. Die verschiedenen Kategorien in 4.3 wurden über ihre Umgebung (Bigramme, Trigramme) im Text statistisch errechnet. Die entsprechende Frequenzliste des BNC besteht aus denselben Einträgen, jedoch nach Häufigkeit statt nach Alphabet geordnet.

4.3 illustriert die Ergebnisse des statistischen Taggers CLAWS4, der für die Analyse des BNC entwickelt wurde und der allgemein als einer der besten statistischen Tagger angesehen wird. Die Fehlerquote[13] von CLAWS4 wird von Leech 1995 auf 1,7 % beziffert, was auf den ersten Blick als sehr gut erscheinen mag.

Man muß jedoch bedenken, daß diese Fehlerquote das Tagging der laufenden Wortformen und nicht der Types betrifft. Angesichts der Tatsache, daß die Abdeckung der letzten 1,2 % der Tokens 83,6 % der Types erfordert (siehe 3.1), kann eine Fehlerrate von 1,7 % auch ein sehr schlechtes Ergebnis bedeuten – nämlich daß über 80 % der Types nicht oder nicht korrekt analysiert werden. Diese Überlegung wird von einer genaueren Betrachtung der Stichprobe 4.3 bestätigt. Für die BNC-Stichprobe 4.3 ergibt sich nämlich eine Fehlerquote von mindestens 60 %.

Zunächst fällt auf, daß von den 38 Einträgen der Stichprobe 27 Einträge mehrfach kategorisiert sind, nämlich `activ` (2), `activa` (3), `activate` (3), `activated` (7), `activating` (6), `activation` (2), `activator` (2) und `activators` (2). Dabei wird der Druckfehler `activ` alternativ als `nn1-np0` und als `np0` klassifiziert, was linguistisch nicht sinnvoll ist. Auch die Klassifikation von `activate` als `np0` ist aus Sicht traditioneller Lexika des Englischen falsch. Der Druckfehler `activatd` wird als `nn1-vvb` kategorisiert, bei `activation.` wird das Interpunktionszeichen nicht eliminiert und der Label `nn0` vergeben, bei `activation/` und `activator/` wird der `/` nicht korrekt interpretiert und der Label `unc` (für *unclassified*) vergeben, wobei die identischen Einträge für `activator/` auch noch separat gezählt werden.

Neben einer hohen Fehlerrate wird die BNC-Statistik durch einen schwachen Präprozessor beeinflußt. Indem etwa verschiedene Zahlen als Wortformen, z. B.

---

[12]Die getaggten BNC-Listen wurden aus dem WWW genommen (Oktober 1997).

[13]Leider wird weder in Leech 1995 noch in Burnard 1995 spezifiziert, was beim Tagging des BNC als Fehler angesehen wird. Immerhin läuft seit Juni 1995 ein neues Projekt zur Verbesserung des Taggers, ‚The British National Corpus Tag Enhancement Project', dessen Ergebnisse ursprünglich im September 1996 zur Verfügung gestellt werden sollten.

```
1 0.544 crd 1
1 0.543 crd 1
1 0.541 crd 1
```

analysiert werden, resultieren 58 477 zusätzliche Types, was 6,3 % der getaggten BNC-Types entspricht. Weitere Beispiele dieser Art sind Bindestrichsequenzen und Kombinationen von Zahlen mit Maßeinheiten.

Insgesamt wird durch das statistische Labelling die Zahl der Types erheblich aufgebläht. 921 074 getaggten BNC-Types entsprechen z. B. 659 270 Oberflächen-Types. Eine geeignete Behandlung von Zahlen und Bindestrichen würde die Oberflächen-Types um weitere 83 317 auf 575 935 Types reduzieren. Insgesamt wird die Zahl der BNC-Types durch das BNC-Tagging also um mindestens 37,5 % erhöht.

Die Tagging-Analyse des BNC ist ein gutes Beispiel für die Stärken und Schwächen einer *smart solution*. Trotz offensichtlicher Verbesserungsmöglichkeiten bei dem Prä- und Postprozessor unseres konkreten Beispiels verbleiben die folgenden prinzipiellen Grenzen des statistischen Taggings:

1. Die morphosyntaktische Analyse (*Kategorisierung*) der Wortformen ist für die Verwendung durch einen regelbasierten syntaktischen Parser viel zu ungenau.

2. Die Wortformen können nicht auf ihre Grundform zurückgeführt werden (*Lemmatisierung*).

3. Die Wortformen können weder in ihre Allomorphe noch in ihre Morpheme zerlegt werden.

4. Das Gesamtbild der Häufigkeitsverteilungen in einem Korpus wird durch ein künstliches Aufblähen der Typezahl um fast 40 % verzerrt.

4.4 Nachteile des statistischen Taggings

Diese Schwächen treten bei Sprachen mit einer etwas reicheren Morphologie als der des Englischen noch um vieles deutlicher in Erscheinung. Als Vorteile des statistischen Tagging wären dagegen der verhältnismäßig geringe Aufwand und die Robustheit zu nennen, die es nahelegen, statistische Tagger zumindest zur Vorbereitung einer gründlicheren Wortformerkennung zu verwenden.

5 Automatische Wortformerkennung DMM

Die Alternative zum statistischen Tagging ist die *solid solution* einer regel- und lexikonbasierten automatischen Wortformerkennung. Ein solches System ist LA-MORPH, das an der Abteilung für Computerlinguistik der Universität Erlangen-Nürnberg (CLUE) entwickelt wurde. LA-MORPH basiert auf der Grammatiktheorie der linksassoziativen Grammatik[14] und setzt die zu analysierenden Wortformen linksassoziativ (d. h. sukzessive von links nach rechts) aus Allomorphen zusammen. Das Lexikon, in dem diese Allomorphe gespeichert sind, wird vor der Laufzeit automatisch von Allomorphregeln aus einem Grundformlexikon erzeugt.

LA-MORPH arbeitet im Rahmen des an der CLUE entwickelten Programmpakets MALAGA.[15] Größere Anwendungen von LA-MORPH sind die Systeme DMM (Deutsche Malaga-Morphologie)[16], IMM (Italienische Malaga-Morphologie)[17], KMM (Koreanische Malaga-Morphologie)[18] und EMM (Englische Malaga-Morphologie)[19]. Diese Systeme stehen auf der CLUE-Homepage zur Verfügung und können über eine Java™-Schnittstelle[20] getestet werden.

Das Grundformlexikon der DMM besteht im Moment aus circa 49 000 Einträgen in folgender Zusammensetzung:

20 300	Substantive
11 100	Adjektive
10 600	Namen und Akronyme
6 200	Verben
960	Funktionswörter, Partikeln, Suffixe, Präfixe und Präfixoide
49160	Gesamt

Aus diesen knapp 50 000 Grundformen werden regelbasiert circa 65 000 Allomorphe generiert. Aus dem Grundformeintrag für *Haus* werden beispielsweise die Allomorphe *Haus* und *Häus* erzeugt. Insgesamt ergibt sich daraus ein Allomorphiequotient von 1,32 für das Deutsche.

[14]Hausser 1992.
[15]Beutel 1997.
[16]Lorenz 1996.
[17]Wetzel 1996.
[18]Lee 1995.
[19]Leidner 1998.
[20]Knorr 1997.

Beispielsweise würde DMM bei der Wortform Häusermeer zuerst das Allomorph Häus erkennen. Über *lexical lookup* wird die kategorialen Information über das Allomorph Häus und das zugehörige Morphem Haus bestimmt. Als nächstes wird das Allomorph -er- gefunden, analysiert und über eine Regel u. a. als Fuge angehängt.

Die Zusammensetzung Häus/er wiederum läßt nur eine bestimmte Menge möglicher Fortsetzungen (also nachfolgender Regeln und zugehöriger Kategorien nächster Allomorphe) zu. Eine dieser Nachfolgeregeln konkateniert Häus/er mit dem nächsten Allomorph, Meer, das als Substantivstamm des Morphems Meer analysiert wurde. In dieser Weise zerlegt DMM die Wortformen in *Allomorphe* und liefert die entsprechenden *Morpheme*, eine genaue morphosyntaktische Analyse, sowie die *Grundform*.

Falls es mehrere Möglichkeiten gibt, eine Wortform zusammenzusetzen (z. B. Ab/treibung vs. Abt/reibung), so werden diese Analysen morphologisch disambiguiert, d. h. es wird versucht, aufgrund morphologischer Kriterien zu entscheiden, welche der Analysen korrekt ist (oder zumindest korrekter als die anderen). Die Entscheidungskriterien sind hierbei die Art der in der Analyse zusammengesetzten Allomorphe und die Art und Anzahl der Konkatenationsschritte. Außerdem werden morphologisch ambige Analysen, die syntaktisch identisch sind, verschmolzen. Durch morphologische Disambiguierung und Verschmelzung wird erreicht, daß DMM im Schnitt nur 1,05 Analysen pro Wortform liefert (anstatt circa 1,4 Analysen ohne diese Mechanismen).

6 DMM-basierte Analyse des Limas-Korpus

Die automatische Wortformerkennung des DMM-Systems ermöglichte erstmals eine umfassende regelbasierte Analyse des Limas-Korpus. Sie liefert zum einen eine detaillierte morphosyntaktische Charakterisierung der einzelnen Wortformen. Zum anderen ergänzt sie die bekannte Häufigkeitsverteilung der Wort*formen* mit den bisher unbekannten Häufigkeitsverteilungen der *Wörter* (Grundformen), *Morpheme* und *Allomorphe*.

Es folgt zunächst eine Frequenzliste des Limas-Korpus, bei der alle Wortformen, alle Morpheme und alle Allomorphe berücksichtigt werden. Erwartungsgemäß wird der Anfang dieser Frequenzlisten von Funktionswörtern bestimmt, während bei den Morphemen die *gebundenen* Morpheme dominieren.

Rang	Wortformen	Allomorphe	Morpheme
1	39911 der	110423 en	157493 _det_pron
2	39278 die	77607 e	110422 en
3	28898 und	50049 t	77606 e
4	18814 in	39911 der	47317 t
5	12478 den	39278 die	37864 ung
6	11402 von	37860 ung	34479 n
7	11349 das	34479 n	33797 _det
8	11091 zu	31172 er	28898 und_conj
9	9607 des	28949 und	26676 s
10	9466 ist	26676 s	26058 er
11	9175 mit	22103 ein	19335 sein
12	8424 sich	18837 in	18814 in_prepos
13	8108 auf	16125 es	17027 einen
14	8056 nicht	14799 zu	13641 werden
15	7473 im	12478 den	12188 auf_prepos
16	7264 eine	12188 auf	11540 an_prepos
17	7260 sie	11540 an	11402 von_prepos
18	7094 für	11402 von	11091 zu_prepos
19	6716 dem	11349 das	10268 aus_prepos
20	6708 ein	10268 aus	9715 mit_prepos

6.1 Frequenzliste (Limas-Korpus Anfang)

Um die Tabelle nicht durch hochfrequente Funktionswörter (z. B. der) im Bereich der Wortformen und hochfrequente Suffixe im Bereich der Allomorphe und Morpheme (z. B. -en) zu vernebeln und um die problematische Darstellung bestimmter Funktionswörter, z. B. der Artikel der, die, das, dem, den etc., und bestimmter Suffixe, z. B. der Pluralendungen -en, -er, -e, -n etc., als letztlich künstliche Morpheme (z. B. der oder Def-Art bzw. -en oder Plural) zu vermeiden, werden die Frequenzen des Limas-Korpus in 6.2 noch einmal für die *offenen* Wortklassen gezeigt.

Es ist offensichtlich, daß sich Wortformen, Allomorphe und Morpheme in einem Korpus wesentlich besser vergleichen lassen, wenn nur die offenen Wortklassen berücksichtigt werden.

Durch die Beschränkung auf die offenen Wortklassen verringert sich die Zahl der Tokens im Limas-Korpus von 1 059 310 (ohne Interpunktionszeichen, Klam-

Rang	Wortformen	Allomorphe	Morpheme
1	9466 ist	9467 ist	19335 sein
2	5408 werden	5408 werden	12949 werden
3	4125 wird	4126 wird	7513 haben
4	4045 sind	4045 sind	5403 können
5	2817 hat	3552 stell	3531 stellen
6	2505 war	2819 hat	2625 zeit
7	2484 kann	2791 bild	2606 geben
8	1929 haben	2625 zeit	2547 müssen
9	1592 können	2543 kann	2427 bilden
10	1451 wurde	2513 war	2419 nehmen
11	1267 hatte	2309 arbeit	2281 jahr
12	1216 muß	2281 jahr	2272 führen
13	1166 zeit	2273 führ	2183 kommen
14	890 sei	2091 setz	2167 groß
15	890 anderen	2076 ander	2091 setzen
16	835 waren	1978 teil	2079 ander
17	777 soll	1929 haben	2036 gehen
18	767 wurden	1772 neu	1963 sollen
19	755 gibt	1736 bau	1881 lassen
20	719 jahre	1724 komm	1844 sehen

6.2 Frequenzen der offenen Klassen (Limas-Korpus Anfang)

mern etc.) auf 617 952 (den Verhältnissen in 3.1 entsprechend). Die Anzahl der Types verringert sich dagegen nur um 5 186 Types von 98 138 Rängen auf 92 952 Ränge. Von diesen 5 186 Types sind 345 Funktionswörter und 2 100 Zahlen; die restlichen Types sind morphologische Hypothesen, die nicht berücksichtigt werden, weil Information über ihre morphologische Struktur nicht zuverlässig vorliegt.

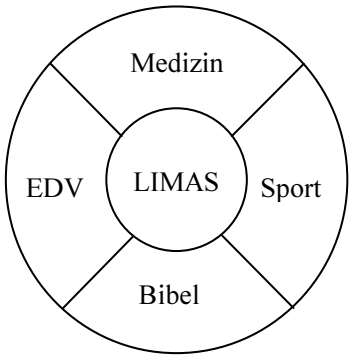
Den 617 952 Tokens bei den Wortformen stehen bei Morphemen und Allomorphen gleichermaßen 713 526 Tokens gegenüber. Die höhere Zahl der Tokens bei Morphemen und Allomorphen resultiert aus der Analyse von Komposita. So wird z. B. Häusermeer als eine Wortform, aber als zwei Morpheme (Haus und Meer) und als zwei Allomorphe (Häus und Meer) gezählt.

Den 92 952 Types der Wortformen der offenen Klassen stehen 21 969 Morpheme und 22 519 Allomorphe gegenüber. Dabei sind von den Wortformen 53 482 oder 57,5 % Hapax Legomena, von Allomorphen und Morphemen sind dagegen nur 7 078 bzw. 6 840 oder jeweils 31,4 % bzw. 31,1 % Hapax Legomena. Die regelbasierte Wortformanalyse reduziert also den Anteil der Hapax Legomena im Vergleich zur traditionellen buchstabenbasierten Methode um beinahe die Hälfte (45,4 %).

Die im Vergleich zu den Wortformen wesentlich kleinere Zahl der Hapax Legomena bei den Morphemen ergibt sich aus der regelbasierten Analyse. Als Wortformen sind z. B. Abbremsung, Babyflaschen und Campingplatz zwar Hapax Legomena im Limas-Korpus, aber ihre Bestandteile bremse, baby, flasche, camping und Platz kommen mehr als einmal vor. Da die Hapax Legomena aus statistischer Sicht als quasi unanalysierbarer Bodensatz eines Korpus betrachtet werden, hat die regelbasierte Wortformanalyse auch den Vorteil, daß sie zu einer wesentlich besseren Ausbeute führt.

7 Domänenspezifische DMM-Analysen

Um einen ersten Eindruck von der Wort- und Morphemverteilung in den niederfrequenten Bereichen des Deutschen zu gewinnen, wurde das Limas-Korpus mit vier domänenspezifischen Korpora gleicher Größe (d. h. je eine Million laufende Wortformen) ergänzt. Diese liegen in den Bereichen Medizin, Sport, EDV und Bibel. Damit ergab sich ein Gesamtkorpus mit folgender Struktur.



7.1 Struktur des CLUE-Korpus

Im CLUE-Korpus dient das Limas-Korpus als relativ hochfrequenter Zentralbereich des deutschen Wortschatzes, da es ja als repräsentatives und balanciertes Korpus konstruiert worden war. Bei den anderen Korpora stellt sich nun die Frage, inwieweit sich ihr Vokabular mit dem des Limas-Korpus und untereinanderüberschneidet.

Die domänenspezifischen Teilkorpora des CLUE-Korpus wurden folgenden Quellen entnommen:

EDV-Korpus:
Als Quelle für das EDV-Korpus dienen Texte aus den Zeitschriften *Computerzeitung*, *iX* und *c't* und den online im Internet verfügbaren Magazinen *InterNet Report* und *Autocad*.
Die Texte der *Computerzeitung* stellen die vollen Jahrgänge 1993 und 1994 dar und lagen auf CD-ROM vor. Ebenfalls auf CD-ROM liegen die Jahrgänge 1994 bis 1996 der Zeitschrift *iX* vor.
Die Texte der Zeitschriften *InterNet Report*[21] und *AUTOCAD-Magazin*[22] des *IWT-Verlags* sowie der Zeitschrift *c't*[23] wurden mit Hilfe des an der CLUE entwickelten Perl-Skriptes holwww aus dem WWW (World-Wide Web) beschafft. Diesem Skript wird ein URL (Uniform Resource Locator, „WWW-Adresse") übergeben; das Skript holt dann den Inhalt dieser WWW-Seite und speichert diesen in einer Datei. Die Beschaffung der Daten kann somit weitgehend automatisiert werden.

Sport-Korpus:
Die Quellen für das Sport-Korpus entstammen den Sportteilen der WWW-Versionen der Tageszeitung *Die Welt*. *Die Welt* verfügt über ein Archiv, in dem sich sämtliche Artikel befinden, die in der *Welt Online*[24] seit 17.Mai 1995 erschienen sind. Auf dieses Archiv kann frei zugegriffen werden. Die Artikel der Domäne Sport wurden ebenfalls mit Hilfe des oben genannten Skriptes holwww automatisch beschafft.
Bei der Auswahl der Quellen wurde darauf Wert gelegt, daß die Erscheinungsdaten der Artikel möglichst weit über alle Jahreszeiten verteilt sind, da aufgrund der Natur der Domäne starke saisonale Abweichungen zu erwarten sind.

[21] http://www.iwtnet.de/inet_report/Homepage.html
[22] http://www.iwtnet.de/autocad/Homepage.html
[23] http://www.heise.de/ct/
[24] http://www.welt.de

Medizin-Korpus:
Das Medizin-Korpus entstammt folgenden Quellen: Der Online-Version der *Ärztezeitung*[25], dem WWW-Server des Bundesministeriums für Gesundheit[26], dem Online-Forum des *Instituts für Medizin und Kommunikation*[27], dem Online-Magazin *MedizInfo*[28], dem Online-Magazin *Medizin-Forum*[29] und den WWW-Seiten der Deutschen Herzstiftung[30] und der Deutschen Krebshilfe[31]. Dazu kamen noch eine Reihe von online verfügbaren Fachzeitschriften des Springer-Verlages[32]: *Der Chirurg, Der Hautarzt, Der Internist, Der Nervenarzt, Der Radiologe, Der Schmerz, Der Unfallchirurg* und *Psychotherapeut*. Beschaffung der Texte wie beim EDV-Korpus.

Bibel-Korpus:
Das Bibel-Korpus entstammt zwei Quellen: Der Bibel in der *Elberfelder Übersetzung*[33] und der sog. *Bibel der Häretiker*[34], einer Sammlung von frühchristlichen gnostischen Handschriften. Auch diese Texte wurden automatisch mit holwww beschafft.
Aufgrund begrenzter Ressourcen wurde bei der Zusammenstellung dieser domänenspezifischen Korpora auf eine strenge Synchronizität der Texte sowie eine Randomisierung verzichtet. Dies schien zum einen vertretbar angesichts der in Abschnitt 2 dargestellten Schwierigkeiten bei der Konstruktion wirklich repräsentativer und balancierter Korpora. Zum anderen liegen die Zwecke des CLUE-Korpus (i) im Testen der DMM und (ii) in einer ersten Untersuchung (Machbarkeitsstudie) der Morphemverteilungen in speziellen Domänen.

Beim Testen der DMM am CLUE-Korpus ergaben sich folgende Erkennungsraten:

[25] http://www.aerztezeitung.de/de/htm/net/start/start.htm
[26] http://www.bmgesundheit.de/
[27] http://www.imk.ch/
[28] http://www.medizinfo.com/
[29] http://www.medizin-forum.de/aktuell/
[30] http://www.dsk.de/dhs/aktuell.htm
[31] http://www.krebshilfe.de/
[32] http://www.link.springer.de/link/service/journals/
[33] gopher://wiretap.spies.com/11/Library/Religion/Bible/German
[34] http://www.gwdg.de/~rzellwe/nhs/nhs.html

Korpus	Tokens	erk.	in %	Types	erk.	in %	to/ty
Limas	1236774	1204225	97,37	121650	104106	85,58	10,16
Bibel	1131536	1106629	97,80	37031	29932	80,83	30,56
Sport	1140121	1082154	94,92	64799	50293	77,62	17,59
EDV	1000001	899176	89,92	100208	66975	66,84	9,98
Medizin	1017646	877964	86,28	104425	66421	63,71	9,74
Total	5526079	5170149	93,56	324570	221138	68,14	17,02

7.2 Erkennungsraten der DMM

Bei den laufenden Wortformen (Tokens) liegt die Erkennungsrate der gegenwärtigen DMM zwischen 97,37 % (Limas) und 86,28 % (Medizin). Die Erkennungsrate für das gesamte CLUE-Korpus ist 93,56 %.

Bei den Types der Wortformen liegt die Erkennungsrate zwischen 85,58 % (Limas) und 63,71 % (Medizin). Für das gesamte CLUE-Korpus liegt die Type-Erkennung bei 68,14 %.[35] Bei der Einschätzung dieser Werte sollte die in 3.1 dargestellte Korrelation von Types und Tokens in Korpora im Auge behalten werden.

Es zeigt sich, daß bei den domänenspezifischen Korpora die Type-Erkennungsrate mit dem Token/Type-Verhältnis (to/ty) korreliert. Je weniger Tokens es zu einem Type gibt, je weniger oft also eine Wortform wiederholt wird, desto höher ist die Anzahl der Types im Korpus – was sich entsprechend auf die Type-Erkennungsrate auswirkt.

Eine Untersuchung der Vokabularüberschneidung[36] verschiedener Teilkorpora ergibt eine Fülle möglicher Morphemklassen, die n-Listen genannt werden, wobei n für die Namen der Teilkorpora steht. Beim CLUE-Korpus gibt es z. B. folgende n-Listen:

Morpheme, die in allen 5 Teilkorpora vorkommen
Morpheme, die jeweils in nur 4 Teilkorpora vorkommen
Morpheme, die jeweils in nur 3 Teilkorpora vorkommen
Morpheme, die jeweils in nur 2 Teilkorpora vorkommen
Morpheme, die jeweils in nur 1 Teilkorpus vorkommen

[35]Daß dieser Wert nicht dem Mittel der einzelnen Erkennungsprozente entspricht, liegt an der Vokabularüberschneidung zwischen den Teilkorpora.

[36]Siehe Schwarz 1996.

In jeder dieser n-Listen wird die Frequenz der Morpheme relativ zu den betrachteten Teilkorpora angegeben. Dies zeigt die folgende Tabelle aus Morphemen, die ausschließlich in den Domänen EDV und Medizin vorkommen, geordnet nach ihrer Gesamthäufigkeit in beiden Domäne, für die Ränge 1 – 20.

Rang	Morphem	Gesamt	EDV	Medizin
1	datei	951	950	1
2	radiologisch	184	1	183
3	interaktiv	172	167	5
4	frame	170	169	1
5	editor	165	161	4
6	spezifikation	160	159	1
7	insulin	154	19	135
8	diabetes	149	10	139
9	joint	142	8	134
10	prospektiv	128	1	127
11	infusion	114	1	113
12	macintosh	103	100	3
13	disk	99	97	2
14	neuronal	98	39	59
15	expression	94	1	93
16	environment	93	90	3
17	skript	92	90	2
18	pixel	84	73	11
19	zertifizieren	82	81	1
20	array	81	80	1

7.3 Morpheme und normierte Frequenzen zweier Domänen

Beispielsweise kommt das Morphem Datei in den Teilkorpora Limas, Sport und Bibel nicht vor, wohl aber in den Teilkorpora EDV und Medizin. Aufgrund seiner unterschiedlichen Häufigkeit (950 vs. 1) ist es für die Domäne EDV wesentlich charakteristischer als für Medizin. Entsprechend ist es mit dem Morphem radiologisch, das aufgrund seiner Häufigkeit charakteristischer für EDV ist als für Medizin.

Zum Schluß ein Vergleich der Morpheme, die in jeweils nur einem der vier domänenspezifischen Teilkorpora vorkommen.

	EDV		Medizin		Bibel		Sport	
1	prozessor	541	lymphom	547	jakobus	167	steffi	512
2	raid	206	maligne	412	ephraim	151	klinsmann	429
3	modem	156	endothel	186	sündigen	144	wimbledon	325
4	proprietär	89	paracetamol	144	zion	144	sammer	255
5	modular	45	laparoskop-	130	joab	133	villeneuve	203
6	debi	40	suppressiv	119	moab	120	berti	202
7	borchers	31	median	113	jonatan	114	sampras	184
8	explorer	29	palliativ	113	samaria	113	hoeneß	182
9	megabit	29	inzidenz	98	knechten	110	hässler	179
10	paperback	28	hypertonie	93	gilead	106	hunke	131
11	portiere	27	ruptur	93	pleroma	105	kirsten	125
12	ergonomisch	22	seehofer	91	absalom	102	derby	113
13	postleitzahl	22	poliklinik	88	esau	98	köpke	100
14	multiplex	20	mortalität	86	jerobeam	97	agassi	98
15	integer	18	psychosozial	81	pharisäer	97	doping	97
16	platine	18	zyste	81	ahab	91	ottmar	89
17	drda	17	septisch	80	elia	91	strunz	81
18	assembler	16	radiologe	76	edom	86	babbel	78
19	permission	16	fixateur	75	nebukadnezar	86	edberg	76
20	synergie	16	zerebral	73	joschafat	84	hertha	71

7.4 Domänenspezifische Unique-Vokabulare (Morpheme)

Es zeigt sich, daß die Teilkorpora mit einem hohen Token/Type-Verhältnis (Bibel und Sport) in ihrem Unique-Vokabular einen hohen Anteil an Eigennamen haben, die zudem häufig vorkommen. Die n-Listen des CLUE-Korpus stehen in ihrer Gesamtheit auf dem CLUE-Web-Server zur Verfügung.

8 Conclusio

Die von der DMM gelieferten Morpheme sind gewissermaßen ein Nebenprodukt einer regelbasierten Wortformerkennung, deren eigentliches Ziel eine präzise morphosyntaktische Analyse für das syntaktische Parsen ist. Im Bereich der Korpusanalyse zeigt sich jedoch, daß die Häufigkeitsverteilungen auf der Ebene der Morpheme ein wesentlich klareres Bild von einer Sprache geben als auf der

Ebene der Wortformen.

Neben einer theoretischen Untersuchung des deutschen Wortschatzes in verschiedenen Domänen haben die hier beschriebenen Methoden auch praktische Anwendungen. Zum einen konnte gezeigt werden, daß bei einer flektierenden Sprache wie dem Deutschen mit einer DMM-basierten Suche eine Recall/Precision-Verbesserung zwischen 42,9 % und 110,5 % erreicht werden konnten.[37] Zum anderen liegt es nahe, die domänenspezifischen n-Listen zur automatischen Klassifikation von Texten zu verwenden.

Bibliographie

Bergenholtz, H. (1976) *„Zur Morphologie deutscher Substantive, Verben und Adjektive. Probleme der Morphe, Morpheme und ihrer Beziehungen zu den Wortarten.“* In: Alfred Hoppe (Hrsg): Beihefte zur kommunikativen Grammatik. Bonn.

Bergenholtz, H. (1989) *“Korpusproblematik in der Computerlinguistik. Konstruktionsprinzipien und Repräsentativität.“* In: Hugo Steger (Hrsg.): Handbücher zur Sprach- und Kommunikationswissenschaft (Bd. IV). Berlin, New York 1989.

Beutel, B. (1997) *Malaga 4.0*, CLUE-Manual.

Brown, P., S. Della Pietra, V. Della Pietra and R. Mercer (1991) *“Word sense disambiguation using statistical methods”,* in: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, June 1991, 264-270.

Burnard, L. (ed.) (1995) *Users Reference Guide British National Corpus Version 1.0*, Oxford University Computing Services.

Church, K. & R. L. Mercer (1983) *“Introduction to the special issue on computational linguistics using large corpora.”* Computational Linguistics, Vol. 19:1, 1 – 24.

DeRose, S. (1988) *“Grammatical category disambiguation by statistical optimization.”* Computational Linguistics, 14:1, 31 – 39.

Garside, R., G. Leech und G. Sampson (1987) *The Computational Analysis of English*, Longman, London & New York.

Francis, W. N. (1980) *“A tagged corpus: Problems and prospects,”* in: S. Greenbaum, G. Leech and J. Svartvik (eds.) 1980, pp. 192 – 209.

Francis, W. N. und H. Kučera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin, Boston.

[37] Piotrowski 1998

Hausser, R. (1992) *"Complexity in Left-Associative Grammar"*, Theoretical Computer Science, Vol. 103, Elsevier.

Hausser, R. (ed.) (1996) *Linguistische Verifikation*. Dokumentation zur Ersten Morpholympics, Max Niemeyer Verlag, Tübingen.

Hess, K., J. Brustkern und W. Lenders (1983) *Maschinenlesbare deutsche Wörterbücher*, Max Niemeyer Verlag, Tübingen.

Hofland, K. und S. Johansson (1980) *Word Frequencies in British and American English*, London: Longman.

Knorr, O. (1997) *Entwicklung einer JAVA-Schnittstelle für Malaga*, CLUE-betreute Studienarbeit der Informatik.

Kučera, H. & W. N. Francis (1967) *Computational Analysis of Present-day English*, Brown University Press, Providence, Rhode Island.

Garside, R., G. Leech & G. Sampson (1987) *The Computational Analysis of English*, Longman, London & New York.

Leidner, J. (1998) *Linksassoziative morphologische Analyse des Englischen mit stochastischer Disambiguierung*, CLUE-Magisterarbeit.

Lee, K. (1995) "*Recursion Problems in Concatenation: A Case of Korean Morphology"*, Proceedings of PACLIC 10, the 10th Pacific-Asian Conference on Language, Information and Computation.

Leech, G. (1995) *"A brief user's guide to the grammatical tagging of the British National Corpus"*, Web-Seite.

Leech, G., R. Garside & E. Atwell (1983) *"The automatic grammatical tagging of the LOB Corpus"*, ICAME Journal 7, 13 – 33.

Lenders, W. und G. Willee (1986) *Linguistische Datenverarbeitung*, Westdeutscher Verlag, Opladen.

Lorenz, O. (1996) *Automatische Wortformerkennung für das Deutsche im Rahmen von Malaga*, CLUE-Magisterarbeit.

Marshall, I. (1983) *"Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus"*, Computers and the Humanities, Vol. 17, 139 – 150.

Marshall, I. (1987) *"Tag selection using probabilistic methods"*, in Garside et al. (eds.).

Meier, H. (1964) *Deutsche Sprachstatistik.* Erster Band. Hildesheim.

Oostdijk, N. (1988) *"A corpus linguistic approach to linguistic variation"*, in G. Dixon (ed.): Literary and Lingustic Computing, Vol. 3.1.

Oostdijk, N. & P. de Haan (1994) *Corpus-based Research into Language*. Editions Rodopi B. V., Amsterdam-Atlanta, GA.

Piotrowski, M. (1998) *NLP-Supported Full-text Retrieval*, CLUE-Magisterarbeit.

Sharman, R. (1990) *Hidden Markov Model Methods for Word Tagging*, Report 214. Winchester: IBM UK Scientific Centre.

Schwarz, R. (1996) *Dynamische Aktivierung domänenspezifischer Teillexika*, CLUE-Magisterarbeit.

Wetzel, C. (1996) *Erstellung einer Morphologie für Italienisch in Malaga*, CLUE-betreute Studienarbeit der Informatik.

Zierl, M. (1997) *Ein System zur effizienten Korpusspeicherung und -abfrage*, CLUE-Magisterarbeit.

Zipf, G. K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*, Oxford.