Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Lexikalische Statistik

Claudia Kunze

kunze8@googlemail.com, kunze@cl.uni-heidelberg.de

Computational Lexicography

Themen

- Quantitativer Ansatz in Literaturwissenschaft und Linguistik
- Definitionsversuch und Leitfragen
- Diafrequente Angaben in Wörterbüchern
- Statistik von Häufigkeit und Verteilung lexikalischer Einheiten in Texten
- Zipfsches Gesetz
- Studie zur morphologischen Produktivität

Entstehung quantitativer Ansätze in CL

Lexikalische Statistik: sehr altes Forschungsgebiet im Vergleich zur CL

- **Quantitative Stilistik**: literaturwiss. Ansätze zur Identifikation des Autors eines Textes unklaren Ursprungs oder zur Entwicklung quantitativer Metriken für die Stilanalyse
- **Quantitative Linguistik**: quantitative Analyse des Sprachgebrauchs im Gegensatz zur (qualitativ geprägten) Analyse der generativen Schule

Lexikalische Statistik: quantitative Analysen von Wörtern in Texten, d.h. Verteilungen von Wortfrequenzen (Zählen der Wörter in Texten, Berechnung der Verhältnisse zwischen diesen Wörtern, Vergleich von Wortfrequenzen und Verhältnissen in verschiedenen Texten).

Mögliche Leitfragen

Lexikalische Statistik hilft bei der Beantwortung z.B. folgender Fragen:

- Wie oft kommt ein Lemma *L* in einem gegebenen Korpus vor (absolute und relative Werte)?
- Wie oft kommt ein Wort (z.B. *also* oder *Tafel*) in seinen verschiedenen Funktionen oder Bedeutungen in einem Textkorpus vor?
- Wie groß ist der Anteil der Wörter, die nur einmal im Text vorkommen (die so genannten HAPAX LEGOMENA)? Wie groß ist der Anteil der 100 häufigsten Wörter an einem Text (Korpus)?
- Wie viele Wörter mit einem bestimmten Präfix oder Suffix (z.B. *Cyber-, -bar*) kommen in einem gegebenen Korpus vor? Wie viele dieser Wörter sind Hapax Legomena?
- Unterscheidet sich der Anteil an Hapax Legomena in kleinen und großen Texten? Mit anderen Worten: gibt es einen funktionalen Zusammenhang zwischen der Proportion an Hapax Legomena und der Größe des Texts (Korpus)?
- Gibt es ein Mittel, um das AKTIVITÄTSPROFIL eines Texts oder das Maß seines lexikalischen Reichtums durch die Analyse seines Wortschatzes zu erfassen und zu vergleichen?

Die ersten vier Fragen(komplexe) sind für die Computerlexikographie relevant, die beiden letzten für die Textlinguistik bzw. Texttechnologie.

Relevanz von Frequenzdaten für CL

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Bei der automatischen Verarbeitung natürlicher Sprache ist die Ambiguität sprachlicher Zeichen nach wie vor eine der größten Herausforderungen. Manche Ambiguitäten extrem selten auftretender Lesarten sind dem Menschen gar nicht bewusst (er zieht sie einfach nicht in Betracht), aber die sprachverarbeitende Software muss diese aussortieren können.

(1) Klauen wollte er nicht.

Der Beispielsatz hat zwei mögliche Interpretationen: *Er wollte nicht stehlen.* vs. *Er wollte keine Klauen (Tierfüße)*. Das Verb *klauen* kommt viel häufiger vor als das Substantiv im Plural zu *Klaue*, so dass die Häufigkeit des Auftretens eine wichtige Rolle beim lexikalischen Zugriff spielt.

Frequenzangaben in Printwörterbüchern

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Vernachlässigung von Frequenzinformation zu lexikalischen Daten in der traditionellen Lexikographie, ausgenommen im Bereich der Frequenzwörterbücher

- Muttersprachler sucht im monolingualen WB meist keine Frequenzdaten, sondern in der Regel Angaben zur Schreibung, Aussprache und Bedeutung seltener Wörter;
- Muttersprachler nutzt monolinguales WB eher zur Textrezeption und zum Schließen von Wissenslücken als zur aktiven Sprachproduktion;
- daher sind seltene Phänomene aus lexikographischer Sicht ebenso wichtig und verzeichnenswert;
- in Lernerwörterbüchern, die zur Sprachproduktion befähigen sollen, sind Frequenzangaben bedeutsamer zur Einordnung der Relevanz von Lesarten
- Fokus auf häufige Phänomene der Sprache in Lerner-WB bei der Auswahl lexikalischer Einheiten
- englische Lernerwörterbücher sind auf der Grundlage umfangreicher Korpusanalysen entwickelt worden
- Selektion und Anordnung der semantischen Lesarten für ein Lemma nach Vorkommenshäufigkeit: zuerst die häufigsten Bedeutungen eines Worts, Homonyme sollten gemäß ihrer Häufigkeit sortiert aufgelistet werden.
- Markieren der häufigsten Wörter der Sprache, z.B. im LDOCE, unterschieden nach gesprochener Sprache (S1-S3) und geschriebener Sprache (W1-W3)

Weitere Frequenzangaben in traditionellen

- in Bezug auf spezifische Merkmale der beschriebenen lexikalischen Einheiten, z.B. Schreibalternativen angeben, die ohne jegliche Bedeutungsänderung austauschbar sind
- Angabe der häufigsten, d.h. unmarkierten Form (z.B. beim gleichzeitigen Gebrauch von starken und schwachen Flexionsformen bei einigen deutschen Verben, z.B. *backen, senden*)
- LDOCE 1995: Frequenzunterschiede zwischen geschriebenem und gesprochenem Englisch: relative Vorkommenshäufigkeiten in grafischer Darstellung; nützliche Hinweise zu Gebrauchshäufigkeiten von Quasi-Synonymen
- Frequenzangaben in deutschen einsprachigen Wörterbüchern sind eher die Ausnahme und bedienen sich vager Termini wie *häufig, selten, gelegentlich auch*, die der Intuition von Lexikographen entspringen und nicht aus Korpusstudien gewonnen wurden

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Statistik von Häufigkeit und Verteilung

Betrachtung von Beispielstatistiken, die Informationen über Häufigkeit und Verteilung von lexikalischen Einheiten in Texten liefern und diese lexikalischen Einheiten damit näher bestimmen

- Index eines Textes oder Korpus: listet alle Zeichenketten (englisch: ‚Strings'), d.h. Folgen von alphanumerischen Symbolen, die (noch) nicht semantisch interpretiert wurden, auf, die im Text (Korpus) vorkommen, zusammen mit einer Zahl, die die Vorkommenshäufigkeit dieses Strings im Korpus angibt
  - Index-String ist ein so genannter TYPE
  - tatsächlich im Korpus vorkommende Strings heißen TOKEN
  - String-Token werden gezählt, der Type wird mit der Tokenhäufigkeit versehen
  - Frequenz kann als Zahl der absoluten Vorkommen (f_w für ein Wort w) vs. der relativen Vorkommen ($\frac{f(w)}{N}$, wobei N die Gesamtzahl der Token im Korpus ist) dargestellt werden
  - alphabetische vs. numerische (nach Tokenhäufigkeit) Sortierung
- Mögliche Modifikationen und Verbesserungen solcher Indizes
  - Entfernung so genannter Nichtwörter (die keine Instanzen eines lexikalsichen Zeichens sind)
  - Aufbau eines umgekehrten Index (letzter Buchstabe des Strings wird zuerst repräsentiert): *database* als *esabatad*, z.B. für morphologische Untersuchungen

Indizes, Fortsetzung

- Erstellen eines LEMMAINDEX, d.h. Wortformen werden unter ihrer gemeinsamen Basisform (LEMMA) gruppiert und zusammengezählt. Dieser ist für lexikographische Zwecke nützlicher als ein Vollformenindex, wenn auch aufwändiger zu erstellen.
- Ausblenden orthographischer Idiosynkrasien, z.B. dem Gebrauch des Bindestrichs in einem Wort: Strings *data-base* und *database* als Token desselben Typs (*database*)
- Rekonstruktion von lexikalischen Einheiten mit diskontinuierlichen Konstituenten. *Er sah das nicht ein* hat eine Instanz des Verbs *einsah* und nicht ein Vorkommen von *sah* und eines von *ein*.
- Aufbau eines n-Gramm-Index, d.h. n-Gramme von Strings oder Graphemen werden aufgelistet und gezählt. (Ein n-Gramm ist eine Folge von n linguistischen Elementen des gleichen Typs. Ein Trigramm von Graphemen z.B. ist eine Folge von drei Graphemen.)
- jede Verbesserung der Qualität eines Indexes setzt eine Art der Vorverarbeitung des Rohtexts voraus, daher Abwägung zwischen der Qualität des Index und dem Aufwand seiner Generierung;
- Index wird häufig durch Konkordanzen (Listen von Textzeilen, in denen der Schlüsselbegriff von einem Kontext von z.B. 50 Buchstaben oder 10 Wörtern links und rechts umgeben ist) ergänzt, wobei diese keine Frequenzdaten sind.

Parametrisierte Indizes

Indizes, bei denen entweder die Liste der Indexmitglieder auf Strings beschränkt ist, die ein bestimmtes Kriterium erfüllen oder bei denen kategoriale Informationen zu den quantitativen Angaben hinzugefügt werden, nennt man PARAMETRISIERTE INDIZES. Beispiele:

- Ordnung der Indizes nach der Wortart der Wörter, d.h. separate Indizes für Nomina, Verben, Adjektive usw. Ein Type wird nicht als bloßer String aufgelistet, sondern als Wort mit morpho-syntaktischer Funktion. Setzt POS-Tagging voraus.
- Index aller Konjunktionen mit ihren Satzpositionen im Sinne von [+/– satzinitial]. Einige Konjunktionen, wie*onst, allerdings, jedoch*, können im Deutschen beide Positionen einnehmen. Grammatiken unterscheiden zwischen *Konjunktionen* und *Konjunktionaladverbien*. Für NLP-Anwendungen könnte die Kenntnis der quantitativen Beziehungen zwischen diesen beiden Typen interessant sein.
- Gebrauch parametrisierter Indizes nützlich für NLP, z.B.
  - ist der String *einen* mit größerer Wahrscheinlichkeit eine Form des indefiniten Artikels als ein Numeral oder eine Verbform,
  - ist der String *kosten* mit größerer Wahrscheinlichkeit eine Wortform der lexikalischen Einheit *kosten* statt eine Form der lexikalischen Einheit *kosen*.

Summe der quantitativen Information, die man aus einem Textkorpus gewinnt, wird SPRACHMODELL genannt. Die lexikalische Ebene ist nur ein Teil hiervon, andere Teile befassen sich z. B. mit der Syntax eines gegebenen Satzes usw. Das Sprachmodell hängt stark von dem Textkorpus, aus dem es abgeleitet wurde, ab, und ist schwer über diese Daten hinaus generalisierbar. Eine Technik, die hierfür benutzt wird, nennt sich EXPECTATION MAXIMISATION.

Morphologische Produktivität

Morphologische Produktivität hat Auswirkungen sowohl auf traditionelle als auch auf Computer-Wörterbücher

- Begriff PRODUKTIVITÄT: Potenzial von Morphemen, insbesondere Affixen, an Neuwortbildungen teilzunehmen
- Morpheme, die in vielen neuen Wörtern gefunden werden, nennt man HOCHPRODUKTIV. Sie sollten als lexikalische Einheiten in Wörterbücher verzeichnet sein.
- Morpheme, die selten oder nie bei der Wortbildung vorkommen, sind keine geeigneten Kandidaten. Es kann aber sinnvoll sein, die mit ihrer Hilfe gebildeten Wörter ins WB aufzunehmen, die für den Muttersprachler ggf. nicht mehr transparent sind.

Textprogression und Wortschatzentwicklu

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Die Wortwarte (www.wortwarte.de) sammelt neue Wörter des Deutschen aus Zeitungstexten (seit dem jahr 2000 ca. 25 000 Neuwörter). Die Beobachtung, dass der Wortschatz einer Sprache täglich wächst, kann auch formaler ausgedrückt werden. Sei

- N die Anzahl der laufenden Wörter (Tokens) im Text (Korpus),
- V die Menge der Worttypen (der Wortschatz) im Korpus,
- w_i ein bestimmter Worttyp,
- $f(w_i)$ die Vorkommensfrequenz des Worttyps w_i im Korpus und
- r der RANG eines Worttyps gemäß seiner Vorkommenshäufigkeit: das Wort bzw. die Wörter, die am häufigsten vorkommen, belegen Rang 1, die zweithäufigsten Wörter belegen Rang 2 usw. Wörter, die gleich häufig vorkommen, bilden eine Menge. Jedes Element dieser Menge belegt den gleichen Rangplatz. Die Anordnung dieser Wörter ist willkürlich.

Wortschatzentwicklung und Textprogressio

Harald Baayens (2001) Arbeit zur Wort-Frequenz-Distribution:

- Beispieltext: Roman *Alice in Wonderland*
- Länge des Romans (N): 26 505 laufende Token
- Baayen unterteilt Korpus in fünf Teile von ca. 5000 laufenden Wörtern zur Darstellung der Wortschatzentwicklung bei wachsender Korpusgröße
- Tabelle 1 zeigt Teil der Wortfrequenzliste aus *Alice in Wonderland*;
- i ist ein numerischer Index; w_i ist das Wort an dieser Indexposition – die Liste ist alphabetisch nach dieser Spalte sortiert; $f(i, 26505)$ ist die Häufigkeit, mit der dieses Wort im Text vorkommt.

Ausschnitt Wortfrequenzliste *Alice in Wond*

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

i	w_i	f(i,26505)
1	a	629
2	alice	386
3	alice's	12
4	and	866
5	bank	3
6	beginning	14
7	book	7
8	but	170
9	buy	57
10	conversation	10
11	conversations	1
12	do	81
13	get	46
14	had	177
15	having	10
16	her	247
17	in	365
18	into	67

Baayens Untersuchung, Tabelle 2

- Index von Worthäufigkeiten m, numerisch sortiert und in umgekehrter Reihenfolge angeordnet, das so genannte FREQUENZSPEKTRUM des Texts.
- Für jeden Frequenzwert m wird die Anzahl der Wörter angegeben, die mit dieser Häufigkeit im Text vorkommen (die V_m, N-Spalte).
- Ergebnis: Es gibt viele Wörter, die nur einmal im Text vorkommen, und wenige, die mit einer hohen Frequenz vorkommen.
- Baayen bezeichnet das erste Phänomen als LARGE NUMBER OF RARE EVENTS (LNRE), eine recht ungewöhnliche Eigenschaft für statistisch erfasste POPULATIONEN.

Frequenzspektrum für *Alice in Wonderland*

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

m	V(m,N)
1	1176
2	402
3	233
4	154
5	99
6	57
7	65
8	52
9	32
10	36
...	...
510	1
528	1
540	1
629	1
726	1
866	1
1631	1

Beziehung: Text:Wortschatz

Statistischer Aspekt in der Beziehung zwischen einem Text und seinem Wortschatz anhand eines „Schnappschuss“(Text mit ca. 26 500 laufenden Wörtern und seiner Wortfrequenzdistributionen)

- jede Frequenzkohorte erhält einen Rang, das Wort, das 1631 Mal vorkommt den Rang 1, das Wort, das 866 Mal vorkommt den Rang 2 usw. und die 1176 Wörtern, die nur einmal im Text vorkommen (die Hapax Legomena) erhalten die höchsten Ränge
- vereinfachte Version von Zipfs Gesetz: „Die Vorkommenshäufigkeit eines Worts (bzw. der Logarithmus dieses Werts) in einem Text ist invers zu seiner Rangzahl (bzw. dem Logarithmus dieses Werts)“
- Details dazu bei Baayen 2001, S. 13-24

Textprogression und Wortschatzentwicklu

z	$f_z(z, N)$	word
1	1611	the
2	866	and
3	726	to
4	629	a
5	540	she
6	528	it
7	510	of
8	460	said
9	410	I
10	386	Alice
...	...	...

Tabelle 3:

Teil einer Wortfrequenzliste, die aus *Alice in Wonderland* abgeleitet und nach Frequenz geordnet ist

Dynamischer Aspekt der Wortschatzentwic

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Dynamischer Aspekt der Beziehung zwischen Text und Wortschatz bzw. das Verhältnis zwischen Textprogression und Wortschatzentwicklung. Der Anteil der Hapax Legomena im Wortschatz eines Textes nimmt ab, wenn die Größe des Texts, gemessen an der Anzahl der laufenden Wörter, zunimmt.

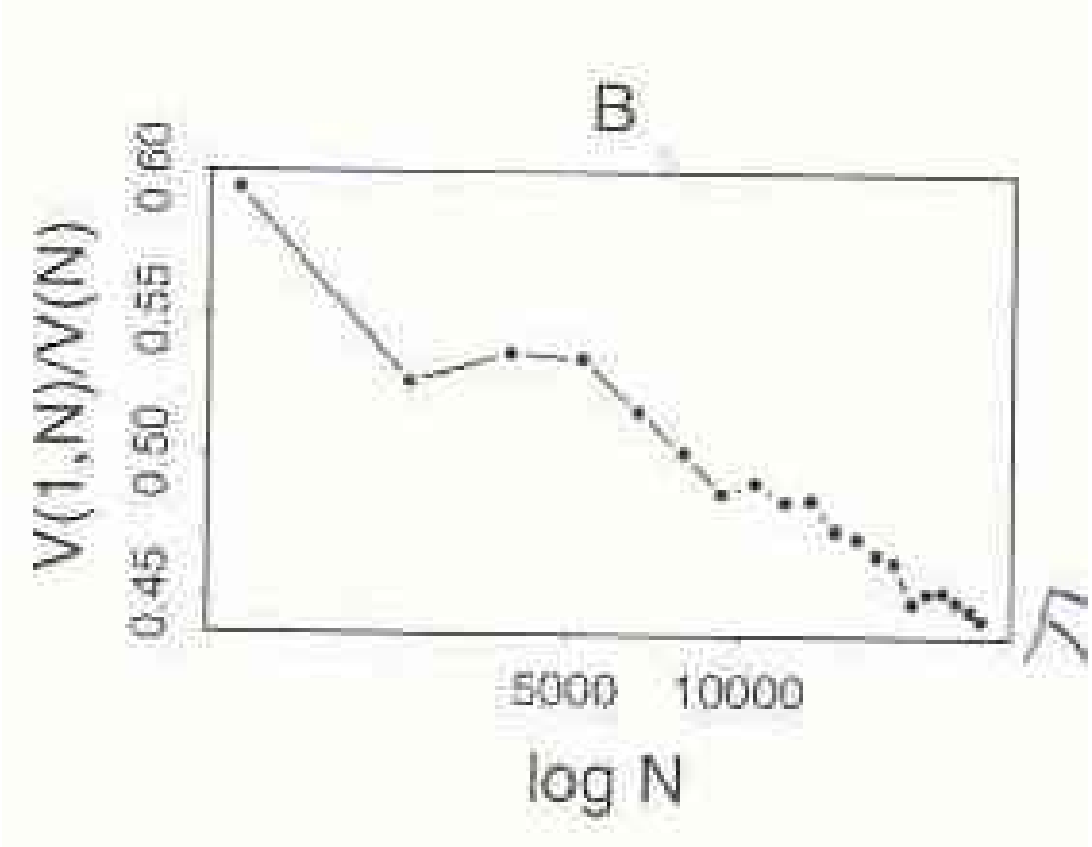

Abbildung 1: Abhängigkeit des Anteils der Hapax Legomena von der Korpusgröße

Aspekte der Textprogression

- Dieselbe Beobachtung durch Evert und Lüdeling: Anteil der Hapax Legomena in einem Beispieltext mit 229 laufenden Wörtern beträgt ca. 65%, wohingegen das Verhältnis der Hapax Legomena in einem Kontrollkorpus von ca. 36 Millionen laufenden Wörtern bei etwa 57% liegt.
- weiterer noch relevanterer Aspekt für Textprogression: absolute Zahl der Wortschatzelemente (Worttypen) steigt stetig mit der Zunahme der Textkorpusgröße. Das Diagramm in Abb. 2 zeigt die erwartete Wortschatzgröße bei unterschiedlichen Korpustextgrößen (große Punkte).

Wortschatzzunahme

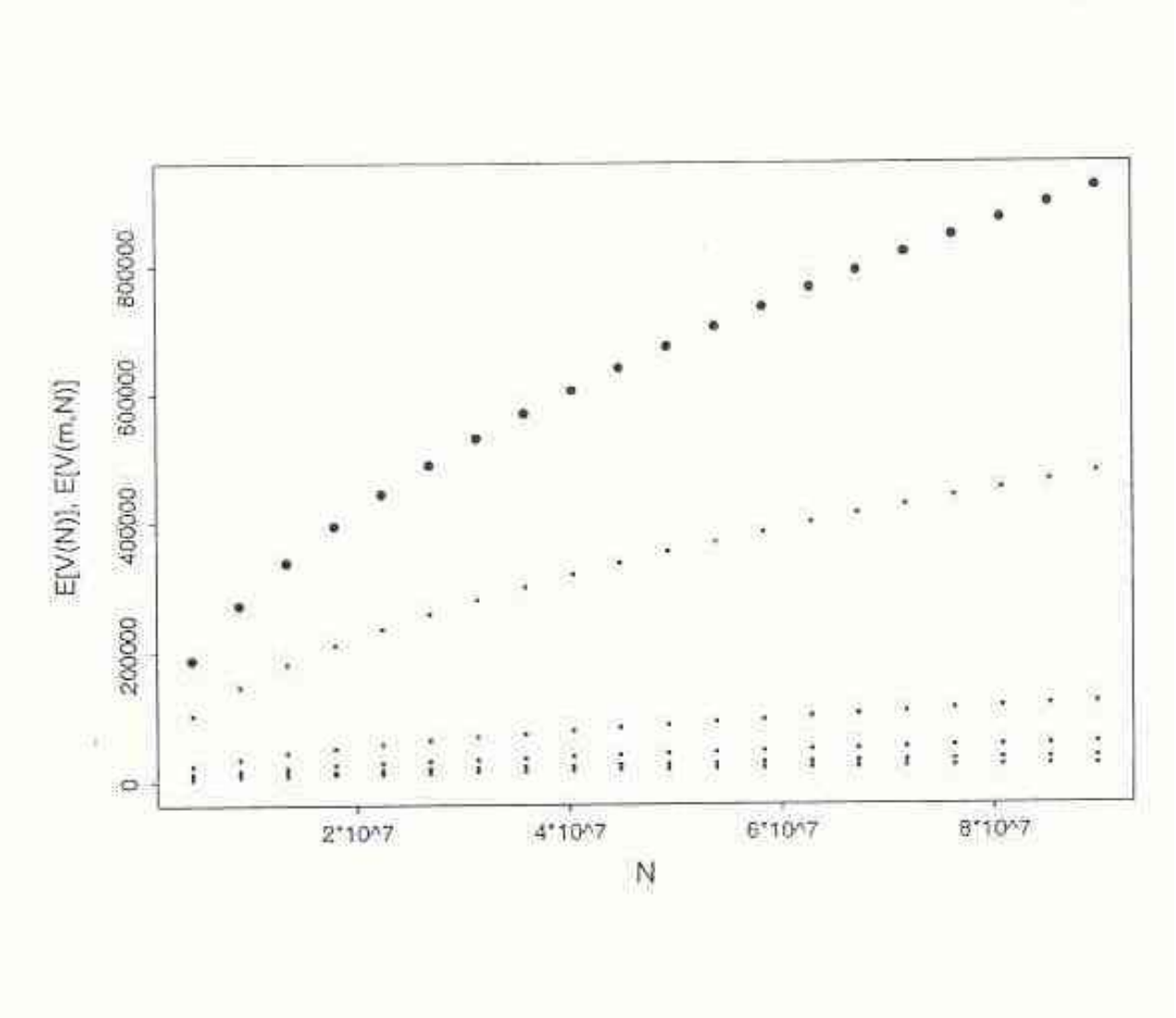

Abbildung 2: (Erwartete) Wortschatzzunahme auf der Grundlage des *British National Corpus*

Studie zur morphologischen Produktivität

- Die meisten neuen Wörter werden aus bestehenden Elementen gebildet werden, nämlich aus Stämmen und Affixen. Die Bildung eines neuen Worts durch Kombination eines möglicherweise komplexen Stamms mit einem Affix nennt man DERIVATION.
- Einige Affixe sind sehr PRODUKTIV, z.B. für das Deutsche die Affixe *-bar* und *-los*.
- Andere Affixe werden selten für die Neuwortbildung benutzt, sind also UNPRODUKTIV; z.B. im Deutschen das Suffix *-sam*.
- Erwartung: Wortformen, die mithilfe von produktiven Affixen gebildet wer den, zeigen eine relativ hohe Wachstumsrate mit zunehmender Textgröße im Vergleich zu Wörtern, die mit unproduktiven Affixen gebildet werden.
- Überprüfung der Erwartung an deutschen Korpusdaten
- Simulieren der Zunahme der Textgröße durch zwei deutsche Textkorpora:
  - Das *Mannheimer Korpus* (MK1), das vom Institut für Deutsche Sprache bereitgestellt wird; mit ca. 10 Millionen laufenden Wörtern.
  - Das *taz* Korpus – 20 Jahrgänge *die tageszeitung*, am Seminar für Sprachwissenschaft der Universität Tübingen aufbereitet, mit etwa 200 Millionen laufenden Wörtern.
- In einem einfachen Extraktionsprozess wurden alle Adjektive, die auf *-bar, -sam, -los* enden, extrahiert, zunächst vom MK1 Korpus allein, danach von beiden Korpora zusammen.

Ergebnisse der Studie

- Wir beobachten die folgenden Wachstumsraten ($f_M K1 \rightarrow f_M K1 + taz$):
  - Für *-bar*: 249 $\rightarrow$ 2041
  - Für *-los*: 222 $\rightarrow$ 1332
  - Für *-sam*: 59 $\rightarrow$ 322
- Erstaunlicherweise gibt es keinen klaren Unterschied in der Wachstumskurve, wenn man bedenkt, dass die Kurve auf einem niederen Niveau für *-sam*, auf einem mittleren Level für *-los* und auf einer etwas höheren Ebene für *-bar* beginnt.
- weiterer interessanter Parameter könnte die Proportion der Hapax Legomena sein, gemessen sowohl bezüglich MK1 als auch bezüglich des kombinierten Korpus
  - Für *-bar*: 43,78 % vs. 40,18 %
  - Für *-los*: 40,54 % vs. 48,57 %
  - Für *-sam*: 30,5 % vs. 49,06 %
- wiederum keine klare Tendenz in den Daten: Beispielsweise gibt es einen beachtlich hohen Anteil an Hapax Legomena mit dem Affix *-sam*. Eine qualitative Analyse ergab: Es gibt viele Hapax Legomena mit dem Affix *-sam*, die gebildet werden, indem ein bereits existierendes *-sam*-Wort als Stamm verwendet wird, z.B. *bühnenwirksam, medienwirksam, lakonisch-bedeutsam*. Dennoch gibt es einige echte Ausnahmen, die für Lexikographen interessant sind, z.B. *anratsam, lachsam*. Dagegen findet man mit *-los* wesentlich mehr echte Derivationen, z.B. *abstandslos, adjektivlos, zugangslos*.

Weitere Studie von Lüdeling und Evert

Untersuchungen zur Produktivität von Wortbildungselementen, wie z.B. dem Suffix *-bar*, sind populär. Die Produktivität in der Wortbildung hat einen qualitativen und einen quantitativen Aspekt, die unterschiedliche Analysemethoden erfordern. Beide Aspekte lassen sich nicht so einfach trennen.

- Ein qualitativer Aspekt der Wortbildung hängt mit der Menge der Elemente, mit denen ein bestimmtes Morphem kombiniert werden kann, zusammen. So ist z.B. der Anwendungsbereich des Suffixes *-bar* auf verbale Basen beschränkt, und hier fast ausschließlich auf die transitiven Verben. Das Suffix *-sam* hingegen tritt zusammen mit verbalen Basen (*arbeit-sam*) und mit adjektivischen Basen (*selt-sam*) auf. Der Anwendungsbereich von *-bar* und damit die Menge der hiermit bildbaren Wörter ist also beschränkter als der Anwendungsbereich von *-sam*.
- Der quantitative Aspekt der Wortbildung kann informell beschrieben werden als die Wahrscheinlichkeit, mit der man auf ein mit einem bestimmten Morphem gebildetes neues Wort trifft, nachdem man bereits eine bestimmte Anzahl von Wörtern beobachtet hat. In einer anderen Sichtweise wird der Produktivitätsindex bestimmt von der relativen Anzahl der Wörter, die bisher nur einmal in den beobachteten Daten auftauchten (eine formale Beschreibung dieses als ‚Vocabulary Growth Curve' bezeichneten Phänomens gibt Baayen (2001). In dieser Interpretation wird man nach Analyse eines Korpus der deutschen Gegenwartssprache feststellen, dass das Suffix *-bar* relativ produktiv ist.

Untersuchung der Produktivität von *-lich*

- Lüdeling und Evert (2003) untersuchen den quantitativen Aspekt der Produktivität des Suffixes *-lich*
- Textbasis: Zeitungskorpus von ca. 3 Millionen laufenden Wörtern
- Analyse der Klasse aller mit *-lich* gebildeten Wörter ergibt ein ziemlich unscharfes Bild; Autoren unterscheiden vier Klassen
  - *-lich* mit adjektivischer Basis (*grün-lich*)
  - *-lich* mit verbaler Basis (z.B. *vergess-lich*)
  - *-lich* mit nominaler Basis (z.B. *ärzt-lich*)
  - *-lich* mit phrasaler Basis (z.B. *vorweihnacht-lich*).
- Die Kombination des Suffixes mit nominaler Basis ist sehr produktiv
- die Kombination mit verbaler Basis hingegen unproduktiv
- andere Bildungsmuster: Datenmenge zu gering für eine ausreichende Bewertung
- auch unter den Nomen gibt es herausragend produktive Stämme (z.B. *X-geschicht-lich*), die eine weitere Klassifizierung der Nomen nahelegen
- qualitative Analyse kann von der quantitativen Analyse profitieren
- qualitative Analyse ist für detailliertes Bild der Produktivität von Wortbildungselementen notwendig

Schlussfolgerungen aus Studien

- automatischer Analyse von Korpora sind hinsichtlich der Anzahl und Häufigkeit von Wortbildungsmustern Grenzen gesetzt
- Fehleranfälligkeit der Analysemöglichkeiten machen eine manuelle Durchsicht der Daten (noch) erforderlich
- rein quantitative Betrachtung der Daten ist nicht angemessen. Z.B. wären alle Wörter auszufiltern, die wahrscheinlich durch Komposition entstanden sind (eine schwer zu automatisierende Angabe)
- Untersuchung zum Verhältnis von Korpusgröße und Wortschatzentwicklung sollte auf einer ausreichenden Menge von Datenpunkten beruhen (empirisch zu bestimmen)
- produktive Affixe *-los* und *-bar* sind selbst als lexikalische Einheiten aufzufassen
- Affix *-sam* bildet trotz seiner prinzipiellen Unproduktivität einige Derivationen, für die sich eine Aufnahme ins Lexikon ggf. lohnen könnte
- Produktivität nicht nach dem „Ja“ oder „Nein“, sondern nach dem „Mehr“ oder „Weniger“betrachtbar
- qualitativ-quantitative Untersuchung profitiert von Klassifikation der derivierten Wörter nach Eigenschaften der Derivationsbasis
- dadurch präzisere Aussagen zur Produktivität der Wortbildungselemente hinsichtlich verschiedener Bildungsmuster –> relevant für eine vollständige lexikographische Beschreibung