Reinhard Köhler (Hg.)
Korpuslinguistische Untersuchungen
zur quantitativen und systemtheoretischen Linguistik
2002, S. 387 - 433

Untersuchungen zur Synergetik der englischen Lexik

Kathrin Gieseking

Zusammenfassung

Reinhard Köhler (1986) hat ein synergetisches Basismodell der Lexik entwickelt und an Daten der deutschen Sprache empirisch überprüft. Dieses Modell erhebt den Anspruch, für alle natürlichen Sprachen gültig zu sein. Um diesen Anspruch zu überprüfen, sind empirische Untersuchungen an unterschiedlichen Sprachen notwendig.

Die vorliegende Arbeit beschreibt die Untersuchung der von diesem Modell vorhergesagten Zusammenhänge von sprachlichen Größen an Daten der englischen Sprache (drei Korpora und ein längerer Text). Diese Untersuchung ergab, dass alle vorhergesagten Variablenzusammenhänge im Englischen statistisch signifikant waren. Es konnte auch nachgewiesen werden, dass sich die Ergebnisse für die verschiedenen Korpora nicht signifikant unterscheiden. Es war jedoch nicht möglich, den numerischen Wert der Parameter der indirekten Variablenzusammenhänge – wie vom Modell vorhergesagt – aus den empirischen Parametern der direkten Variablenzusammenhänge zu errechnen.

Im Verlauf der Untersuchungen traten eine Reihe methodischer Fragestellungen auf, die nicht auf diese Arbeit beschränkt sind, sondern allgemein für quantitativ-synergetische Untersuchungen von Interesse sind und deshalb ausführlicher diskutiert werden.

Bei dieser Arbeit handelt es sich um eine gekürzte Fassung von Gieseking (1993).

1. Das synergetische Basismodell der Lexik von Köhler

Aufbauend auf den Grundgedanken der Synergetik, die von Hoffmann & Krott (in diesem Band) ausführlich dargestellt worden sind, hat Reinhard Köhler (1986) ein synergetisches Basismodell für ein Subsystem der Sprache, die Lexik, modelliert. Als Systemgrößen gelten in diesem Modell die sprachlichen Eigenschaften

- Länge,
- Frequenz,
- Polylexie,
- Polytextie,
- Lexikongröße und
- Phonemzahl.

Die ersten vier dieser Eigenschaften sind Größen, die zur Beschreibung einzelner Lexeme verwendet werden, während es sich bei den letzten beiden um Eigenschaften des lexikalischen Systems in seiner Gesamtheit handelt.

Für dieses lexikalische Basismodell wurden von Köhler neben dem Axiom der sprachlichen Selbstorganisation aus theoretischen Überlegungen heraus weitere Axiome postuliert, die alle zur Klasse der Systembedürfnisse gehören, also Anforderungen der Systemumgebung an das System repräsentieren, und somit Einfluß auf die Systemgrößen nehmen. Diese Bedürfnisse, die von Köhler (1990) in drei Gruppen klassifiziert wurden (Systembedürfnisse der allgemeinen Steuerung, sprachkonstituierende und sprachformende Bedürfnisse), sind ebenfalls schon von Hoffmann & Krott beschrieben worden.

Die Struktur des synergetischen Systems der Lexik wird bestimmt durch die Systembedürfnisse, die Systemgrößen und ihre Relationen zueinander. Abbildung 1 gibt einen Überblick über das Modell. Die hierbei verwendete graphenalgebraische Schreibweise wird im Anhang von Hoffmann & Krott erläutert. Diese Darstellungsweise wird der von Köhler verwendeten funktionalen Modellierung besonders gerecht, da sich aus einer so erzeugten Struktur Aussagen über die Funktion ableiten lassen. Voraussetzung für die Anwendung der linearen Graphenalgebra ist allerdings die Linearisierung des Modells. Sie wird durch eine logarithmische Transformation der Werte der Systemgrößen erreicht (angedeutet in der Abbildung durch ein vorangestelltes *L*). Die allgemeinen Prinzipien der funktionalen Modellierung haben Hoffmann & Krott schon erklärt. Deshalb werden sie im folgenden nur noch einmal kurz anhand eines konkreten Beispiels erläutert, und zwar an der Modellierung der Systemgröße *Polytextie*.

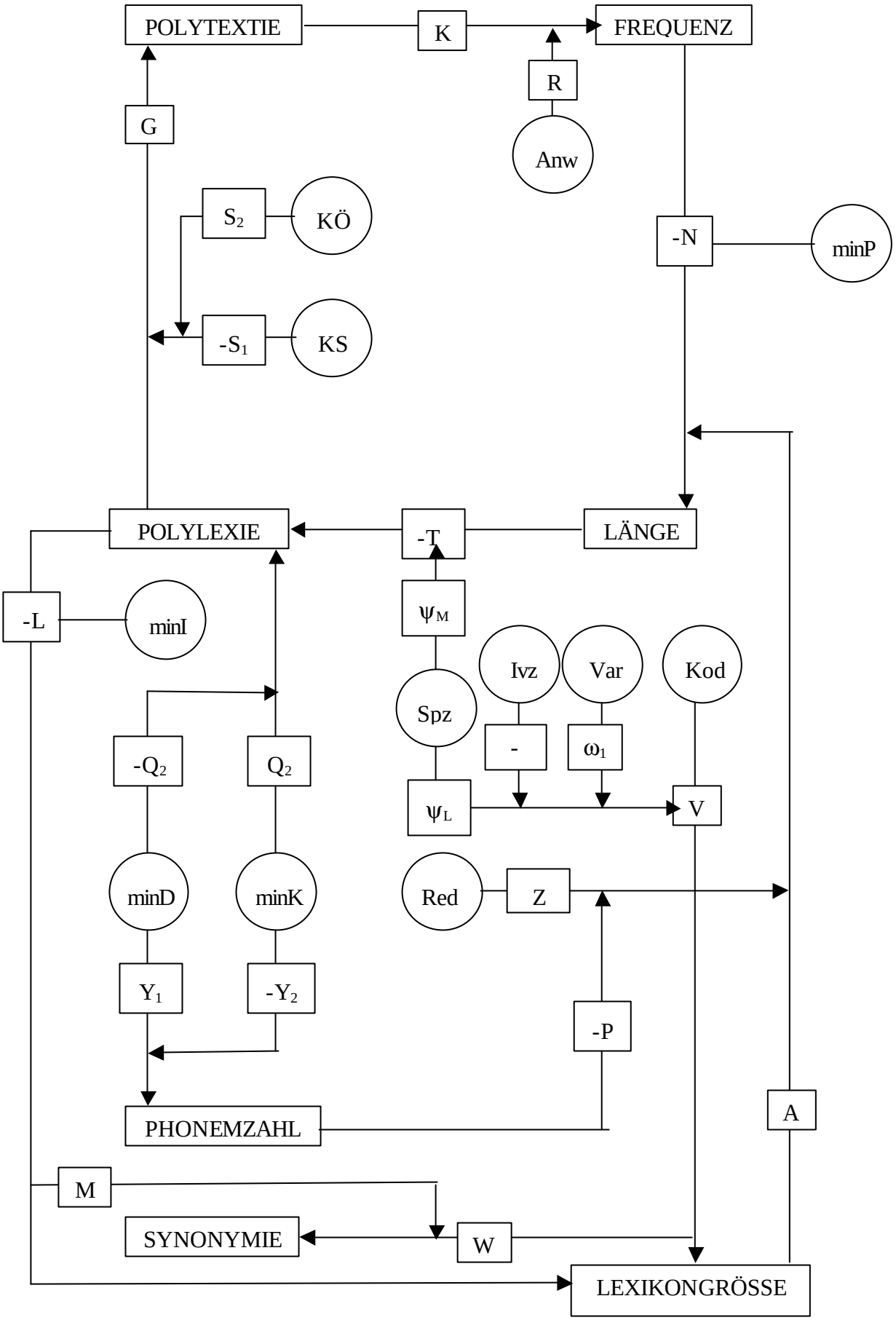

Abb. 1: Gesamtstruktur des Basismodells (nach Köhler, 1986:74)

Die Polytextie ist – wie auch die drei Systemgrößen Länge, Polylexie und Frequenz – ein Attribut jedes einzelnen Lexems. Sie bezeichnet die Eigenschaft eines Lexems, kontextunabhängig verwendbar zu sein, und wird gemessen in der Anzahl der verschiedenen Texte, in denen das Lexem verwendet wird. Die Polytextie wird beeinflußt von zwei entgegengesetzten Systembedürfnissen:

- dem Bedürfnis nach Kontextökonomie (*KÖ*) und
- dem Bedürfnis nach Kontextspezifizität (*KS*).

Neben der Summe dieser beiden Systembedürfnisse, die eine Konstante bildet, die auf jedes Lexem gleich stark wirkt, beeinflußt auch die Polylexie den Polytextiewert eines Lexems: je mehr Bedeutungen ein Wort trägt, in desto mehr (Kon-)Texten läßt es sich verwenden.

Die Hypothese, dass die Veränderung der Polytextie (*PT*) einer lexikalischen Einheit proportional zu ihrer Polylexie (*PL*) ist, läßt sich mit folgender Differentialgleichung beschreiben:

$$\frac{PT'}{PT} = \frac{G}{PL}$$

Die Veränderungsrate der abhängigen Größe *PT* wird also von einem konstanten Anteil *G* und einem variablen Anteil *PL* bestimmt. *G* repräsentiert die Wirkung der Systembedürfnisse *KÖ* und *KS* und nimmt für jede Ausprägung der unabhängigen Variablen *PL* den gleichen Wert an. *PL* stellt den Einfluß der unabhängigen Größe dar und kann verschiedene Werte annehmen.

Die Lösung dieser Gleichung lautet $PT = C \cdot PL^G$. Linearisiert führt das zu $\ln(PT) = c + G \ln(PL)$ mit $c = \ln(C)$. *G* stellt in dieser Gleichung den Proportionalitätsfaktor dar, der beschreibt, in welchem Maße die Polylexie die Polytextie beeinflußt. *C* ist der summierte Einfluß der beiden hier wirksamen Systembedürfnisse, also $\ln(PT) = (s_2 \cdot KÖ - s_1 \cdot KS) + G \cdot \ln(PL)$.

Differentialgleichungen der obigen Form lassen sich für alle direkten Abhängigkeiten zwischen den vier strukturellen Systemgrößen, also den Größen, die die Eigenschaften jedes einzelnen Lexems beschreiben, aufstellen.

Die funktionale Modellierung erlaubt es außerdem, durch einfaches Einsetzen der Gleichungen ineinander aus der Modellstruktur auch indirekte Abhängigkeiten zwischen den Systemgrößen zu deduzieren. Es lassen sich sogar die theoretischen Parameterwerte der indirekten Abhängigkeiten aus den empirischen Parameterwerten der direkten Abhängigkeiten berechnen. Die so gewonnenen theoretischen Parameterwerte können dann mit den empirischen Parameterwerten für die indirekten Abhängigkeiten verglichen werden und auf diesem Weg zu einer Aussage über die Adäquatheit der Deduktion verhelfen.

Die beiden verbleibenden Systemgrößen, Lexikongröße und Phonemzahl, hat Köhler (1986) als empirische Parameter aufgefaßt. Sie ergeben sich weitgehend aus Größen, die außerhalb des Modells liegen, und können in bezug auf die Abhängigkeiten zwischen den vier strukturellen Systemgrößen beim gegenwärtigen Entwicklungsstand des Modells nur als Konstanten betrachtet werden.

2. Die empirische Überprüfung des Modells am Englischen

Köhler (1986) hat das von ihm entwickelte lexikalische Basismodell an Daten der deutschen Sprache empirisch überprüft[1]. Die Untersuchungen bestätigten – unter Verwendung des *F*-Tests – alle vorhergesagten direkten und indirekten Zusammenhänge zwischen den Systemgrößen.

Da das Basismodell nicht nur Aussagen fürs Deutsche macht, sondern prinzipiell für alle Sprachen Gültigkeit beansprucht, ist es notwendig, die aus ihm abgeleiteten Aussagen an weiteren Sprachen zu überprüfen. In der vorliegenden Arbeit wurde diese Überprüfung an Daten der englischen Sprache vorgenommen. Fürs Englische stehen in ausreichendem Maße maschinenlesbare Texte ebenso wie ein maschinenlesbares Lexikon zur Verfügung.

2.1 Verwendete Daten

Folgende Texte und Korpora wurden für die Überprüfung des Modells analysiert:

1. *Das SUSANNE-Korpus* (*SUS*)

Das SUSANNE-Korpus[2] besteht aus 64 der 500 Texte des bekannten Brown-Korpus. Sie enthalten je ca. 2000 Lexeme und entstammen verschiedenen Themenbereichen.

Viele der Texte sind keine abgeschlossenen Texte, sondern lediglich Textfragmente. Dies ist für die Analyse nachteilig, weil die lexikalische Struktur von ganzen Texten sich von der Struktur von Textfragmenten unterscheidet und dadurch Verfälschungen auftreten können. Trotzdem wurde dieses Korpus verwendet, weil

---

[1] Die empirische Überprüfung erfolgte an einer Stichprobe von 13000 Lemmata (bzw. für die Polylexie 1325 Lemmata) des LIMAS-Korpus. Das LIMAS-Korpus enthält 500 Texte/Textfragmente mit einer Länge von jeweils etwa 2000 Wörtern.

[2] Das SUSANNE-Korpus entstand unter der Leitung von Geoffrey Sampson (Universität Sussex) mit dem Ziel, eine umfassende Taxonomie und ein Notationsschema für die Grammatik des Englischen zu entwickeln, die an den Anforderungen maschineller Sprachverarbeitung orientiert ist.

- es eine große strukturelle Ähnlichkeit zu dem von Köhler untersuchten Textkorpus aufweist und damit die Ergebnisse der Analyse Vergleichswerte zu Köhlers Daten liefern können,
- die Bestimmung eines Polytextiewerts möglich ist,
- alle Wörter dieses Korpus mit Wortklassenbezeichnern versehen sind und deshalb mit diesem Korpus die Analysen auch nach Wortklassen getrennt durchgeführt werden können.

2. *6048 Artikel des Wall Street Journals*[3] *von 1989* (*WSJ*)
Die gut 6000 Artikel besitzen recht unterschiedliche Längen und sind – anders als beim SUSANNE-Korpus – abgeschlossene Texte. Dieses Korpus ist insofern ein homogenes Korpus (wie Altmann, 1992 es fordert) als alle Texte der gleichen Textsorte (Zeitungsartikel) angehören und im gleichen Jahr entstanden sind. Das Korpus wurde zum Zweck der Kreuzvalidierung in zwei Teile (WSJ1 und WSJ2) mit je ca. 3000 Artikeln aufgeteilt, die getrennt analysiert wurden.

3. *The Time Machine von H.G. Wells* (*1895*)[4] (*TM*)
Sowohl das SUSANNE-Korpus als auch die beiden WSJ-Korpora bestehen aus einer Vielzahl kleiner Texte (bzw. Textfragmente). Um auch einen langen, abgeschlossenen Text zu untersuchen, wurde zusätzlich die Erzählung *The Time Machine* analysiert.

Aus allen Texten wurden Ziffern bzw. Wörter, die Ziffern enthalten, ausgeschlossen. Aus dem SUSANNE-Korpus wurden zusätzlich Abkürzungen und Eigennamen entfernt, weil es hier durch die Wortklassenmarkierungen weitgehend möglich war.

Lexikon
Neben den verwendeten Texten bzw. Korpora stand das maschinenlesbare *Collins English Dictionary* (CED) von 1978 zur Verfügung[5]. Neben der Anzahl der Bedeutungen pro Wort konnten in ihm auch die Wortlänge in Silben und in Phonemen gemessen werden.

---

[3] Die Artikel des *Wall Street Journals* entstammen einer Textsammlung der *Data Collection Initiative* (DCI). Die DCI ist eine Initiative der *Association for Computational Linguistics* (ACL).

[4] Dieser Text entstammt der Sammlung des ‚*Project Gutenberg*' am Illinois Benedictine College.

[5] Das CED entstammt ebenfalls der Sammlung der ACL/DCI.

2.2 Meßverfahren

Es wurde versucht, für jede Art der Abhängigkeit zwischen den Systemgrößen die größtmögliche Menge an Daten zu gewinnen. Das bedeutet, dass nicht für jede Systemgröße die gleiche Menge an Daten zur Verfügung stand. Nach welchen Kriterien die Daten für die einzelnen Systemgrößen erhoben wurden, wird im folgenden näher beschrieben.

Länge
Die Länge einer lexikalischen Einheit wurde auf drei verschiedene Arten gemessen:

(a) graphemische Länge (*LG*)
Die graphemische Länge ist eine Systemgröße, die für alle untersuchten Lexeme vorliegt. Um diesen Wert zu erhalten, wurden lediglich die Buchstaben pro Wort gezählt. Als Wort galt jede Buchstabenfolge zwischen Leerzeichen oder äquivalenten Trennern.

An dieser Stelle muß auf ein Problem hingewiesen werden, das sich sowohl auf alle Ausprägungen der Größe ‚Länge' als auch auf alle anderen Systemgrößen bezieht. Im Englischen werden Komposita – anders als im Deutschen – oft dadurch gebildet, dass zwei oder mehr Lexeme nebeneinander gestellt, aber nicht durch Zusammenschreibung oder Einfügung eines Bindestriches als Komposita gekennzeichnet sind. Da es sich beim größten Teil der Komposita um Verbindungen handelt, deren einzelne Elemente auch isoliert bedeutungstragend sind (z.B. *ballroom dancing*, *population explosion*) war es hier mit vertretbarem Aufwand nicht möglich, diese Komposita automatisch zu erkennen. Deshalb konnten zur Berechnung aller Systemgrößen nur die einzelnen Elemente der Komposita herangezogen werden. Das bedeutet, dass von vornherein ein gewisses Fehlermaß angenommen werden muß.

(b) Silbenzahl (*LS*)
Die Silbenzahl eines Lexems wurde dem CED entnommen. Diese Größe liegt also nur für die Lexeme vor, die im CED eingetragen sind.

(c) Phonemzahl (*LP*)
Die Phonemzahl eines Lexems wurde ebenfalls dem CED entnommen. Gezählt wurden die Elemente der phonetischen Umschrift im CED, die dem *International Phonetic Alphabet* (IPA) entspricht.

Zusätzlich zu den Phonemzahl für die Lexeme, die im Lexikon enthalten sind, wurde die Phonemzahl auch für einige durch einfache Ableitungen zu bildende Lexeme berechnet (z.B. wurde die Phonemzahl von ‚*gets*' durch die Addition

der im Lexikon gefundenen Länge für ‚*get*' (=3) mit der Länge des Suffixes ‚*-s*' (=1) berechnet.

Polytextie (*PT*)
Die Polytextie wurde bestimmt durch die Anzahl der verschiedenen Texte, in denen ein Lexem auftrat. Dieser Wert konnte für das SUSANNE-Korpus und die WSJ-Korpora bestimmt werden, nicht aber für *The Time Machine*, weil es sich dabei ja nur um einen einzigen Text handelt. Bei der Analyse der beiden erstgenannten Korpora stand der Polytextiewert – wie auch der Wert für die graphemische Länge – für alle untersuchten Wörter zur Verfügung.

Die hier verwendete Operationalisierung des Polytextiewertes (Polytextie eines Wortes gleich Anzahl der Texte, in denen ein Wort auftritt) ist natürlich nicht optimal. Besser weil differenzierter wäre sicherlich die Zählung der tatsächlichen *Kontexte*, in denen ein Wort vorkommt. Es ist jedoch in diesem Fall äußerst schwierig, *Kontext* zu definieren. Kontexte können semantischer oder syntaktischer Natur, enger oder weiter gefaßt sein. Nicht zuletzt stellt bei großen Datenmengen die Zählung und Analyse der Polytextie mit den zur Verfügung stehenden Mitteln ein unlösbares Problem dar, wenn eine sehr differenzierte Definition für Polytextie zugrundegelegt wird. Aus diesen Gründen wird die eingangs erwähnte Operationalisierung verwendet und als Annäherung an die Zahl der verschiedenen Kontexte, in denen ein Lexem auftreten kann, betrachtet.

Polylexie (*PL*)
Zur Bestimmung des Polylexiewerts eines Lexems wurden die unterschiedlichen Bedeutungen, die diesem Lexem im CED zugewiesen wurden, gezählt. Als unterschiedlich galten dabei die mit verschiedenen Nummern versehenen Einträge zu einem Lexem, nicht aber die mit Buchstaben markierten Unterbedeutungen (*get* 1. und *get* 2. kennzeichnen also gemäß dieser Definition zwei verschiedene Bedeutungen, *get* 1.a. und *get* 1.b. nur eine).

Frequenz (*F*)
Für die Frequenz einer Wortform wurden die Vorkommen dieser Form in jedem einzelnen Textkorpus gezählt. Wie graphemische Länge und Polytextie liegt auch dieser Wert für jedes untersuchte Wort vor.

Noch einmal zusammenfassend führten die beschriebenen Meßverfahren dazu, dass

- die Werte *LG*, *F* und *PT* für alle Lexeme vorliegen
- die Werte *LS* und *PL* für alle Lexeme, die im Lexikon eingetragen sind, vorliegen

- der Wert *LP* für alle Lexeme, die im Lexikon vorhanden waren, sowie für einige einfach zu erschließende flektierte Formen vorliegt.

Bei den Ergebnistabellen zu den Abhängigkeiten (s.u.) ist jeweils die zur Bildung der Mittelwerte verwendete Anzahl von Wertepaaren angegeben.

2.3 Statistische Methoden

Bei der empirischen Überprüfung des Modells sollen folgende Fragen mittels statistischer Untersuchungen geklärt werden:

- Sind die vom Modell vorhergesagten Zusammenhänge zwischen den Systemgrößen signifikant?
- Wie groß ist der Anteil der Varianz in den Daten, der durch das lexikalische Basismodell erklärt wird?
- Unterscheiden sich die theoretischen Parameter der indirekten und doppelt indirekten Abhängigkeiten signifikant von deren empirischen Parametern?
- Unterscheiden sich die Ergebnisse der Analyse verschiedener Korpora signifikant?

Zur Beantwortung dieser Fragen muß zunächst für jede Abhängigkeit eine Regressionsanalyse durchgeführt werden, um die theoretische Funktionsgleichung für diese Abhängigkeit zu berechnen. Anschließend wird überprüft, ob die sich ergebende Regressionskurve eine signifikante Steigung aufweist sowie der Determinationskoeffizient r^2 ermittelt, der den Anteil der durch das Modell erklärten Varianz in den Daten angibt.

Da es sich beim lexikalischen Basismodell um ein nicht-lineares, aber linearisierbares Modell handelt, besteht grundsätzlich die Möglichkeit, eine lineare oder eine nicht-lineare Regression durchzuführen. Beide Verfahren bieten Vor- und Nachteile.

Zunächst erscheint es naheliegend, ein nicht-lineares Modell mit Hilfe einer nicht-linearen Regression anzupassen. Das hat den großen Vorteil, dass die Ergebnisse dieser Analyse auch für das nicht-lineare Modell interpretiert werden dürfen. Leider eignet sich die nicht-lineare Regression jedoch nur als Anpassungsmechanismus. Das bedeutet,

- dass die Parameter iterativ geschätzt werden und deshalb ein durch nicht-lineare Regression gewonnener Determinationskoeffizient nicht als die durch das Modell erklärte Varianz der abhängigen Werte betrachtet werden darf, sondern lediglich als ein Maß für die Güte der Anpassung.

- dass keine Signifikanztests durchgeführt werden können, da die statistische Testtheorie für Resultate aus nicht-linearer Regression keine zur Verfügung stellt. Solche Tests sind aber notwendig, um
  - zu bestimmen, ob der Zusammenhang zwischen zwei Systemgrößen signifikant ist,
  - zu untersuchen, ob sich theoretische und empirische Parameter signifikant unterscheiden,
  - zu überprüfen, ob sich die durch Analyse verschiedener Korpora erzeugten Regressionskoeffizienten signifikant unterscheiden.

Wie schon in Abschnitt 1 gezeigt, lassen sich die Abhängigkeiten in Köhlers lexikalischem Basismodell durch logarithmische Transformation leicht linearisieren und lauten dann allgemein $\ln(y) = \ln(A) + B \cdot \ln(x)$.

Diese Linearisierung hat den Nachteil, dass so erzielte Parameterwerte für das zugrundeliegende nicht-lineare Modell zwar als Anhaltspunkte dienen können, aber nur für das linearisierte Modell interpretiert werden dürfen. Da die Parameter bei der linearen Regression mit der Methode der kleinsten Fehlerquadrate berechnet werden, darf man r^2 allerdings als den Anteil der durch das (lineare) Modell erklärten Varianz werten. Außerdem ermöglicht die Linearisierung den Einsatz statistischer Signifikanztests, deren Einsatz zur Überprüfung mehrerer Hypothesen notwendig ist.

Ausgehend von diesen Überlegungen und unter Berücksichtigung der Tatsache, dass die Logarithmierung eine monotone Transformation ist und die beste Lösung im nicht-linearen Modell deshalb der besten Lösung im linearisierten Modell entspricht, wurden die theoretischen Funktionsgleichungen schließlich durch *lineare Regression* ermittelt.

Ob zwischen zwei Variablen ein signifikanter Zusammenhang besteht, wurde mit der Nullhypothese H_0: $b = 0$ überprüft, die besagt, dass der theoretisch zu erwartende Wert für den Regressionskoeffizienten Null ist, also kein signifikanter Zusammenhang zwischen den Variablen besteht. Zum Test der Nullhypothese wurde nach folgender Gleichung ein *t*-Wert berechnet:

$$t = \frac{B - b}{s_B} \quad mit \;\; b = 0.$$

s_B bezeichnet den Standardfehler des Regressionskoeffizienten und berechnet sich folgendermaßen:

$$s_B = \frac{s_{Y.X}}{\sqrt{\sum x^2}} = \frac{\sqrt{\frac{\sum (Y - Y')^2}{n-2}}}{\sqrt{\sum x^2}}$$

Bleibt der *t*-Wert unter dem kritischen Wert von *t* (Tabellenwert), wird die Nullhypothese nicht verworfen und kein (linearer) Zusammenhang zwischen den beiden Variablen postuliert. Ist der *t*-Wert mindestens gleich dem Tabellenwert, wird ein signifikanter (linearer) Zusammenhang zwischen den Variablen angenommen.

Der Determinationskoeffizient, also das Maß der durch das Modell erklärten Varianz, berechnet sich nach folgender Gleichung:

$$r^2 = 1 - \frac{\sum(Y_i - Y_i')^2}{\sum(Y_i - \overline{Y})^2}$$

Zusätzlich wurde für jede Abhängigkeit eine nicht-lineare Regressionsgleichung nach dem Levenberg-Marquardt-Algorithmus berechnet, damit mit dem so bestimmten Determinationskoeffizienten auch ein Anhaltspunkt für die Güte des nicht-linearen Modells zur Verfügung steht.

Bei der Bildung der Wertepaare bei der empirischen Überprüfung seines Basismodells hat Köhler jeder Ausprägung der unabhängigen Variablen den Mittelwert aller entsprechenden abhängigen Variablen zugewiesen. Diese Analyse der Mittelwerte der abhängigen Variablen hat einen Nachteil: Jeder Mittelwert zählt gleich, egal ob er das Mittel aus Tausenden einzelner Werte oder aus nur einem einzigen Wert darstellt.

Abbildung 2, die den Zusammenhang zwischen den Variablen Silbenzahl und Polylexie im WSJ2-Korpus zeigt, soll dies deutlich machen.

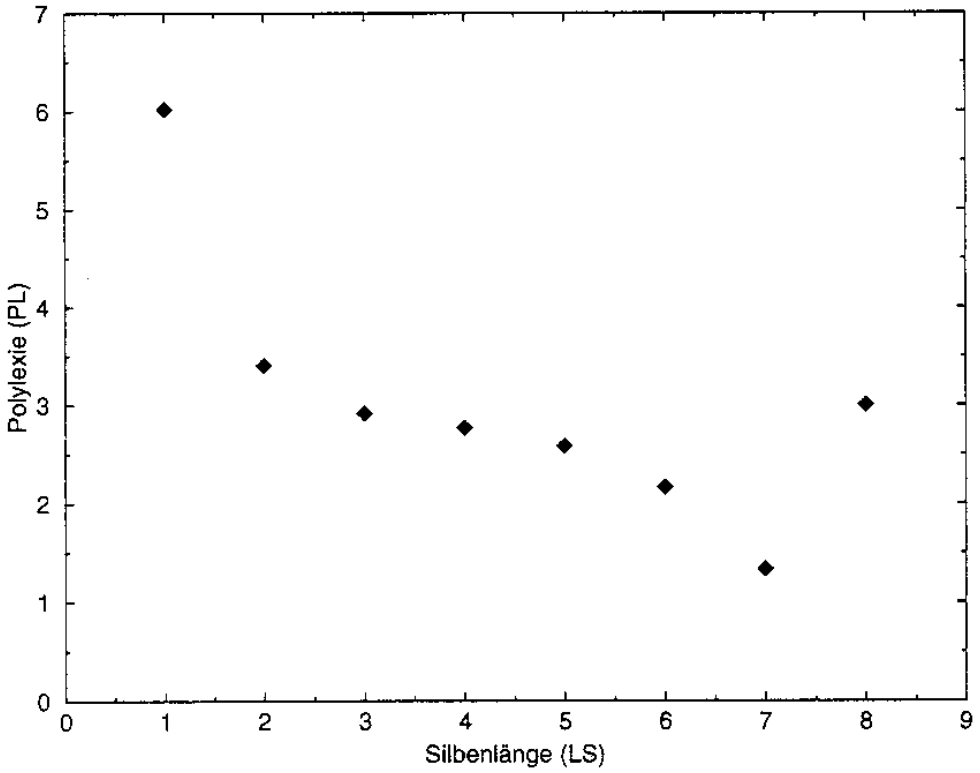

Abb. 2: Die Abhängigkeit der Polylexie von der Silbenzahl (WSJ, Teil 2)

Während der Wert für Polylexie an der Stelle *LS* = 1 sich aus 3027 Werten errechnet, repräsentiert er an der Stelle *LS* = 8 nur den Wert für *ein* Wort, das nur *einmal* im gesamten WSJ2-Korpus auftritt. (Es ist das Wort *counter-revolutionary*, für das im CED drei Bedeutungen eingetragen sind.) Für die Bestimmung des Determinationskoeffizienten für diese Abhängigkeit zählen jedoch beide Werte gleich. Auf diese Weise kann es allein durch dieses eine (relativ unwichtige) Wort zu nicht unbeträchtlichen Verschiebungen der Regressionsgeraden und Veränderungen des Determinationskoeffizienten kommen.

Hier bietet sich das Verfahren der Gewichtung der einzelnen Wertepaare (X, $\overline{Y}$) an. Dabei bekommt jedes Wertepaar als Gewicht die Anzahl der Einzelwerte zugewiesen, aus denen sich $\overline{Y}$ berechnet, und unter Verwendung dieser Gewichtung wird die Regression durchgeführt. Diese Methode verhindert weitgehend, dass Einzelwerte wie im obigen Beispiel den Determinationskoeffizienten beträchtlich beeinflussen können. Eine so erhaltene Regressionsgleichung sorgt auch dafür, dass die theoretische Kurve eng an den Mittelwertdaten in dem Bereich anliegt, der den Großteil aller untersuchten Lexeme repräsentiert, und kann so sehr gute Vorhersagen in diesem Bereich machen. Ihr Nachteil besteht darin, dass die Kurve sich in spärlich besetzten Regionen manchmal beträchtlich von den Daten entfernt und es in diesem Bereich zu völlig falschen Vorhersagen kommen kann. Oft ist es jedoch so, dass gerade isoliert stehende Mittelwerte (die häufig nur das ‚Mittel' aus einem einzigen Lexem sind) sprachlich wichtige (weil z.B. sehr häufige) Wörter repräsentieren, für die das Modell eine gute Vorhersage machen sollte.

Im Kontrast dazu liefert die Regression mit ungewichteten Wertepaaren eine etwas schlechtere Anpassung in dem Bereich, der den Großteil der Rohdaten erfaßt, aber eine vergleichsweise gute Anpassung in dünn besetzten Regionen.

Bei den im folgenden beschriebenen Untersuchungen fiel die Entscheidung *zugunsten einer Gewichtung* der Mittelwerte, um eine bessere Anpassung derjenigen Mittelwerte zu erzielen, die viele Einzelwerte repräsentieren. Außerdem ließen sich Abweichungen der Daten von der Kurve, die sich bei Gewichtung der Daten ergab, besser mit den theoretischen Modellannahmen vereinbaren.

Für die beschriebenen statistischen Maße hat diese Entscheidung zur Konsequenz, dass bei der Berechnung des Determinationskoeffizienten die quadrierten Abweichungen entsprechend der Anzahl der Datenpunkte, die den Mittelwert bilden, gewichtet werden, und beim *t*-Test als Anzahl der Freiheitsgrade die Summe aller Gewichte, also die Anzahl der Rohdatenpunkte angenommen werden muß.

Unter Anwendung der beschriebenen Operationalisierungen und statistischen Verfahren wurden alle zwölf möglichen Abhängigkeiten zwischen den vier Systemgrößen empirisch überprüft. Durch Auffächerung der Systemgröße Länge in graphemische, phonemische und silbische Länge entstanden also 24 verschiedene Abhängigkeiten. Es kann an dieser Stelle vorweggenommen werden, dass alle

untersuchten Variablenzusammenhänge, die mit der oben beschriebenen Nullhypothese H_0: $b = 0$ überprüft wurden, statistisch signifikant waren. Die Zusammenhänge werden deshalb im folgenden als gegeben angesehen und auf diesen Test wird nicht weiter Bezug genommen. Die statistischen Tests, die zur Beantwortung der letzten beiden der oben gestellten Fragen erforderlich sind, werden in den Abschnitten 2.7 und 2.8 dargestellt.

Zunächst werden die Untersuchungen der Variablenzusammenhänge, die in Köhlers lexikalischem Basismodell direkte Abhängigkeiten darstellen, beschrieben, und anschließend in kürzerer Form die indirekten und die doppelt indirekten Abhängigkeiten im Modell.

Zur Illustration der Funktionen werden – wenn nicht anders erwähnt – die Ergebnisse des SUSANNE-Korpus herangezogen.

2.4 Direkte Abhängigkeiten

Die Abhängigkeit der Polytextie von der Polylexie

Aus der Struktur des lexikalischen Basismodells läßt sich die Hypothese ableiten, dass die Polylexie einer lexikalischen Einheit ihre mittlere Polytextie bestimmt. In der entsprechenden Funktionsgleichung steht die Konstante *A* für die Summe der divergierenden Systembedürfnisse Kontextökonomie und Kontext-spezifizität und die Konstante *B* für den Einfluß der Polylexie.

Tabelle 1 zeigt die Ergebnisse der Regression mit gewichteten Datenpunkten. Die Spalten der Tabelle geben für jedes Korpus die Parameter *A* und *B*, den Determinationskoeffizienten aus der linearen Regression (r^2 (*LR*)), den Determiniationskoeffizienten aus der nicht-linearen Regression (r^2 (*NLR*)), die Anzahl der Mittelwerte (# MW) sowie die Anzahl der verschiedenen Lexeme, aus denen die Mittelwerte gebildet wurden (# Types) an.

In der dazugehörigen Abbildung 3 wird der Effekt der Gewichtung gut sichtbar[6]. Trotz großer Streuung, die bei ungewichteten Daten zu einer fast waagerechten Regressionskurve mit nicht-signifikanter Steigung führen würde, kommt es hier zu einer Regressionskurve mit deutlicher Steigung und zu recht guten Werten für den Determinationskoeffizienten.

Tabelle 1. Anpassung der Funktion $PT = A \cdot PL^B$

6 Für alle Abbildungen werden die logarithmierten Daten und die darauf berechneten Regressionsgeraden der linearen Regression der besseren Anschaulichkeit wegen zurücktransformiert.

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	2.5431	0.3753	0.82	0.78	55	6953
WSJ1	11.6220	0.7189	0.92	0.79	55	14520
WSJ2	11.7294	0.7013	0.91	0.79	55	14311

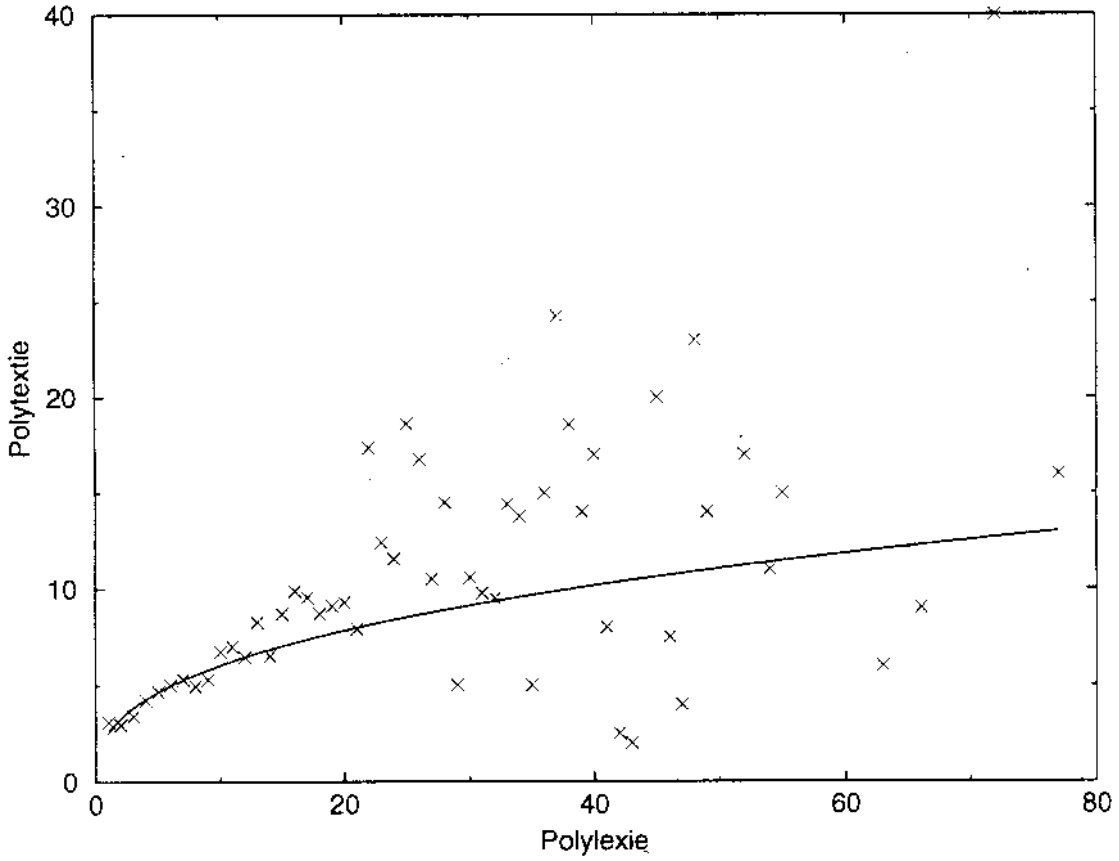

Abb. 3: die Abhängigkeit der Polytextie von der Polylexie

Die graphische Darstellung zeigt, dass die Anpassung der theoretischen Kurve an die empirischen Werte im Bereich der niedrigen Polylexiewerte (bis etwa 20) sehr gut ist, und es erst im Bereich der Polylexiewerte ab 20 zu einer relativ großen Streuung kommt. Da die Datenpunkte, die die Polylexiewerte von 1 bis 20 beschreiben, schon 98,4% der Daten erfassen, sind sowohl die signifikante Steigung der Regressionskurve als auch die guten Werte für die Determinationskoeffizienten, die durch die Datengewichtung erzielt wurden, zu rechtfertigen.

Wenn man die große Streuung im oberen Bereich der Daten genauer ansieht, stellt man fest, dass ab etwa $PL = 20$ nur noch sehr wenig Mittelwertbildung stattfindet, also viele Datenpunkte einzelne Lexeme repräsentieren. Eine nähere

Betrachtung dieser Lexeme mit einer Polylexie > 20 ergab, dass es sich bei diesen Lexemen (im SUSANNE-Korpus 112 Stück) um sehr allgemein verwendbare Wörter wie z.B. *break*, *close*, *drive*, *fix*, *form*, *free*, *get*, *good*, *hand*, *high*, *house*, *run*, *take*, *work* etc. handelt. Gerade bei solchen Wörtern ist es oft schwierig, verschiedene Bedeutungsnuancen sauber zu unterscheiden und zu klassifizieren. Deshalb ist möglicherweise ein lexikographisches Problem mitverantwortlich für die große Streuung im oberen Polylexie-Bereich.

Die Abhängigkeit der Polylexie von der Länge

Laut Modell bestimmt die Länge eines Lexems seine durchschnittliche Polylexie. Die Konstante *A* repräsentiert in der Funktionsgleichung den Einfluß der gegeneinanderwirkenden Systembedürfnisse nach Minimierung des Kodierungsaufwands und Minimierung des Dekodierungsaufwands, Konstante *B* steht für den Grad der Synthetizität der untersuchten Sprache und den Einfluß des Bedürfnisses nach Spezifikation.

Wieder wurde eine Regression mit gewichteten Fällen durchgeführt. In Variablenzusammenhängen, in denen die Länge die unabhängige Variable darstellt, ist die Gewichtung der Fälle besonders wichtig, weil sonst die wenigen Lexeme mit großen Längen, deren Rolle in der Sprache relativ unbedeutend ist, einen zu starken Einfluß auf die theoretische Verteilung hätten (vgl. das Beispiel in 2.3).

Das Verfahren führte für die graphemische und die phonemische Länge zu den Ergebnissen in den Tabellen 2 und 3. Abbildung 4 zeigt die Datenpunkte für die graphemische Länge und die Polylexie. (Die gleiche Abhängigkeit unter Messung der Länge in Phonemen ergab ein sehr ähnliches Bild.)

Tabelle 2
Anpassung der Funktion $PL = A \cdot LG^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	21.3660	-0.7951	0.83	0.70	18	6953
WSJ1	12.6890	-0.6679	0.80	0.67	18	14520
WSJ2	12.7496	-0.6667	0.79	0.66	20	14311
TM	28.2276	-0.8774	0.85	0.72	16	2937

Tabelle 3
Anpassung der Funktion $PL = A \cdot LP^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	16.5172	-0.7183	0.85	0.76	18	6929
WSJ1	11.1563	-0.6462	0.86	0.79	17	14166
WSJ2	11.2380	-0.6470	0.85	0.80	17	13955
TM	19.7253	-0.7516	0.85	0.73	14	2924

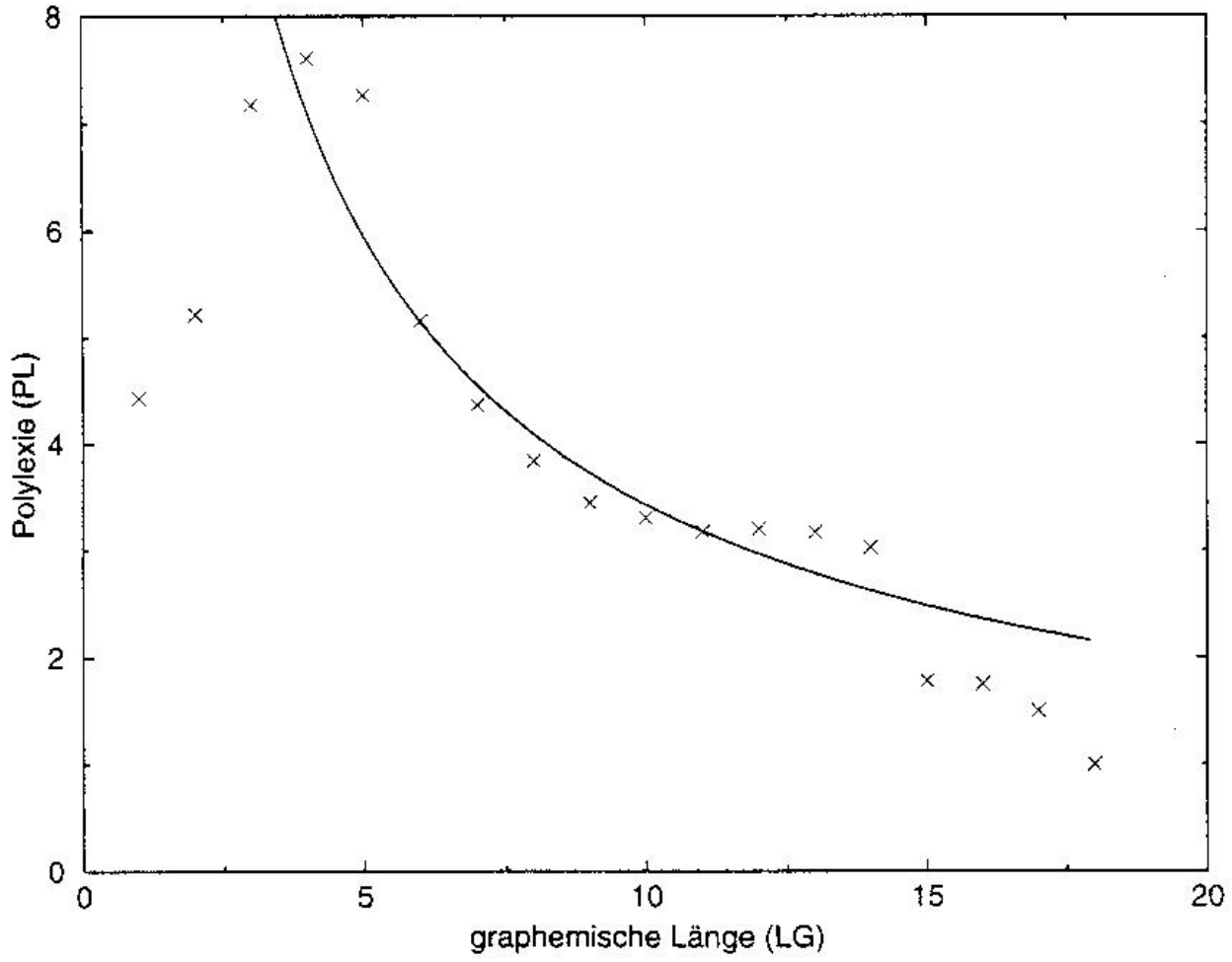

Abb. 4: Die Abhängigkeit der Polylexie von der graphemischen Länge

Besonders auffällig ist bei dieser Abhängigkeit, dass die empirischen Daten nicht zu einer monotonen Kurve führen, wie es vom Modell vorhergesagt wird. Stattdessen steigt die Kurve bis *LG* = 4 (Abb. 4) bzw. *LP* = 3 (o. Abb.) an, um dann wieder abzufallen. Deshalb erreichen natürlich die Determinationskoeffizienten für *LG* und *LP* trotz des guten optischen Eindrucks des zweiten Teils der Kurve keine besonders hohen Werte.

Wie kommt es zu dieser nicht-monotonen empirischen Kurve, die dem Modell in seiner einfachsten Form widerspricht? Eine erste Vermutung war, dass es sich bei den Lexemen, die für ihre geringe Länge eine zu kleine Anzahl von Bedeutungen haben, um Funktionswörter handelt. Diese werden in Lexika oft nicht in so viele Bedeutungsnuancen unterschieden, wie das für Inhaltswörter der Fall ist. Pronomina wie ‚*he*', ‚*she*', ‚*it*' beispielsweise besitzen trotz einer potentiell unendlichen Referenzmenge im Lexikon (notgedrungen) nur eine kleine Anzahl von Bedeutungen.

Ein weiteres kritisches Moment bilden die Initialen und Abkürzungen. Sie sind in größerem Umfang in den WSJ-Korpora enthalten, und einige wenige konnten nicht automatisch aus dem SUSANNE-Korpus herausgefiltert werden, waren also zunächst noch an der Bildung der Mittelwerte beteiligt.

Zur Klärung der Gründe für das beobachtete Phänomen der Nicht-Monotonie wurden die Lexeme des SUSANNE-Korpus betrachtet, deren LG 3 bzw. LP # 2 ist. Für die graphemische Länge ergaben die Untersuchungen, dass bei $LG = 1$ nach Ausschluß der Initialen der durchschnittliche Polylexiewert 5.0 (statt 4.43) betrug, also weiterhin unter dem Polylexiewert für $LG = 2$ lag. Die Lexeme mit $LG = 2$ waren fast ausschließlich Funktionswörter, in diesem Fall könnte also die Vermutung, dass der zu niedrige Polylexiewert durch ein lexikographisches Phänomen verursacht wird, zutreffen. Der Wert für $LG = 3$ jedoch ergibt sich fast ausschließlich aus Inhaltswörtern und ist trotzdem kleiner als der Wert für $LG = 4$. Da man den Polylexiewert für $LG = 3$ als gesichert ansehen darf, erscheinen auch die Polylexiewerte für $LG < 3$ als plausibel, obwohl sie (wegen der geringen Anzahl der Lexeme mit $LG = 1$ und den fast ausschließlich aus Funktionswörtern bestehenden Lexemen mit $LG = 2$) meßtechnisch problematisch sind.

Eine nähere Analyse der Daten der phonemischen Länge führt zu ähnlichen Ergebnissen. Teilweise verstärkte sich hier der nicht-monotone Effekt durch den Ausschluß der Funktionswörter sogar noch.

Dass Initialen und Abkürzungen wenig Einfluß auf den Verlauf der Kurve haben, zeigen auch die Werte, die *The Time Machine* liefert. Der Text enthält weder Initialen noch Abkürzungen, und trotzdem weist die entsprechende Kurve (o. Abb.) das gleiche nicht-monotone Verhalten auf. Schließlich wurde noch einmal das SUSANNE-Korpus nach Wortklassen getrennt analysiert. Sowohl die Substantive als auch die Verben zeigten das gleiche nicht-monotone Verhalten wie die Gesamtmenge aller Lexeme.

Bei der Abhängigkeit der Polylexie von der Silbenzahl (Tabelle 4 und Abbildung 5) tritt dieser Nicht-Monotonie-Effekt nicht auf. Dies hat zum einen sicher damit zu tun, dass praktisch alle Lexeme mit $LG < 4$ bzw. $LP < 3$ in die Menge der einsilbigen Wörter eingehen, die so groß ist, dass die verhältnismäßig wenigen sehr kurzen Lexeme mit geringen Polylexiewerten den Mittelwert der Polylexie für $LS = 1$ kaum beeinflussen können. Wenn man sich die Begründung des

Zusammenhangs zwischen Polylexie und Länge wieder ins Gedächtnis ruft, erscheint aber eine andere Erklärung dafür, dass die Regression im Fall der Silbenzahl zu einer ausgezeichneten Anpassung der theoretischen Kurve an die empirischen Daten führt, plausibler.

Tabelle 4
Anpassung der Funktion $PL = A \cdot LS^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	7.4745	-0.6200	0.98	0.99	7	6953
WSJ1	5.4838	-0.5780	0.92	0.95	7	14520
WSJ2	5.5295	-0.5794	0.92	0.95	8	14311
TM	8.6244	-0.6621	0.99	0.997	6	2937

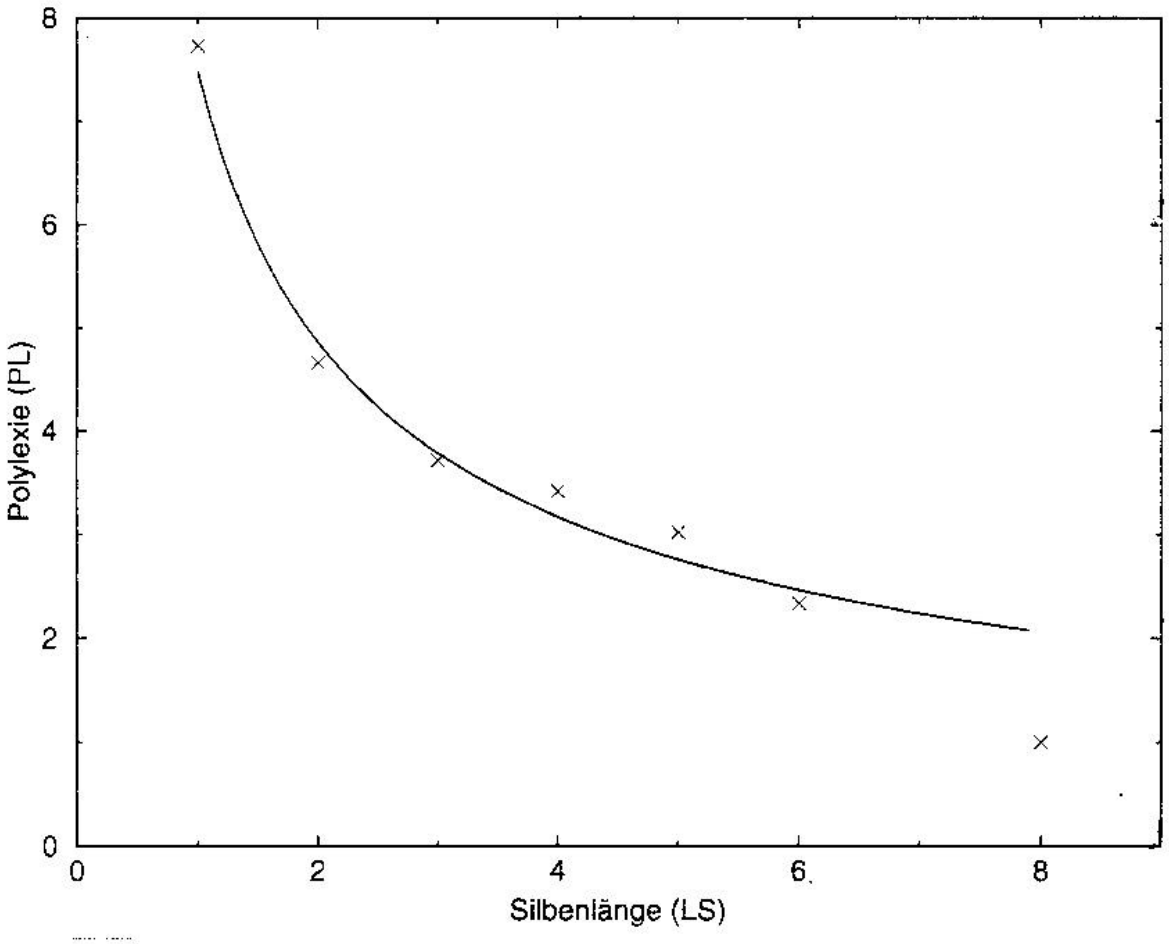

Abb. 5: Die Abhängigkeit der Polylexie von der Silbenzahl

Natürliche Sprachen können verschiedene Mittel zur Spezifikation einer Bedeutung einsetzen. Sofern eine Sprache nicht ausschließlich analytisch ausgelegt ist, macht sie auch von morphologischen Mitteln Gebrauch. Das bedeutet, dass sie, um die Bedeutung eines Lexems zu spezifizieren, das Lexem um ein oder mehrere Morpheme erweitert (z.B. *boat* → *steamboat, hold* → *uphold*). Aufgrund dieser Tatsache sagt das Modell genaugenommen einen monotonen Zusammenhang zwischen Polylexie und der in *Morphemen* gemessenen Lexemlänge voraus.

Von den drei hier verwendeten Längenmaßen stellt die Silbenzahl die beste Näherung an die Morphemzahl dar. Deshalb ist es eigentlich nicht erstaunlich, dass die Anpassung bei der Silbenzahl zu den besten Ergebnissen führt, und die Parameter, die die Silbenzahl hier liefert, sollten als die relevanten betrachtet werden.

Der Parameter *B* in der Funktionsgleichung (er entspricht dem Proportionalitätsfaktor -*T* im Basismodell) wird vom Modell als ein Maß für den Grad der Synthetizität der Sprache angesehen. Er beträgt hier in der Funktionsgleichung $PL = A \cdot L^B$ je nach Korpus für die Silbenlänge zwischen -0.5780 und -0.6621 (bzw. um die Vergleichbarkeit mit den empirischen Daten von Köhler zu ermöglichen: bei nicht-gewichteten logarithmierten Daten beträgt *B* für die graphemische Länge zwischen -0.3969 und -0.5328), ist also vom Betrag her deutlich kleiner als der von Köhler für das deutsche Korpus errechnete Synthetizitätsfaktor von -0.828. Das wird der Tatsache gerecht, dass das Englische eine analytischere Sprache ist als das Deutsche.

Die Abhängigkeit der Länge von der Frequenz

Eine weitere aus dem Köhlerschen Modell ableitbare direkte Abhängigkeit ist die Abhängigkeit der Länge eines Lexems von seiner Frequenz. In der resultierenden Funktionsgleichung steht die Konstante *A* für den kombinierten Effekt von Lexikongröße, Phonemzahl und dem Bedürfnis nach Übertragungssicherheit, die Konstante *B* für den Einfluß der Frequenz.

Bei diesem Variablenzusammenhang wird die Problematik der Gewichtung der Mittelwerte besonders offensichtlich. Zwar erhält man durch die Gewichtung eine sehr gute Anpassung an die Daten im unteren Frequenzbereich, die den Großteil der Wertepaare insgesamt repräsentieren (Abb. 6), dafür liegt die Kurve aber im hochfrequenten Bereich deutlich über den Daten (durchgezogene Regressionskurve in Abb. 7). Gerade hier repräsentieren die Datenpunkte Lexeme, die in der Sprache aufgrund ihrer Häufigkeit eine besondere Bedeutung haben.

Es ist unbefriedigend, wenn für diese Lexeme keine guten Vorhersagen gemacht werden können. Eine Regression auf ungewichteten Daten (gestrichelte Regressionskurve in Abb. 7) führt zu besseren Vorhersagen im hochfrequenten Bereich, aber nur um den Preis, dass die Vorhersagen im unteren Frequenzbereich deutlich schlechter werden. Nun mag jedes einzelne Wort in diesem Bereich

eine verhältnismäßig unwichtige Rolle in der Sprache spielen, in der Gesamtheit jedoch repräsentieren die Mittelwerte des unteren Frequenzbereichs (bis $F = 10$) ca. 89% aller Lexeme des SUSANNE-Korpus. Allein der *LG*-Mittelwert für $F = 1$ setzt sich aus ca. 45% der untersuchten Lexeme zusammen, und es erscheint nicht gerechtfertigt, in der Regression diesen Wert nur genauso stark zu werten wie z.B. den *LG*-Wert für $F = 3040$, der nur die Länge eines einzigen Lexems beschreibt.

Hinzu kommt, dass eine theoretische Funktion, die im hochfrequenten Bereich alle Datenpunkte unter sich läßt (also die Funktion, die durch Gewichtung der Fälle zustande kommt), eher mit den theoretischen Modellannahmen in Einklang zu bringen ist: Wörter, die für ihre Häufigkeit zu kurz sind, belasten das Sprachsystem bzw. die Sprecher an sich nicht. Sie tun es erst dann, wenn alle Kombinationsmöglichkeiten für kurze Lexeme ausgeschöpft sind. Die mögliche Verletzung des Bedürfnisses nach Übertragungssicherheit durch für ihre Frequenz zu kurze Lexeme fällt bei den Lexemen mit den höchsten Frequenzen, die allesamt Funktionswörter sind, nicht stark ins Gewicht, weil aus den meisten Äußerungen der Sinn auch dann erschlossen werden kann, wenn ein Teil der Funktionswörter nicht deutlich genug übertragen wurde. Ist hingegen ein Wort für seine Häufigkeit zu lang, stört es das Ökonomiebedürfnis der Sprecher und gerät so unter einen Anpassungsdruck, der über kurz oder lang zu einer Kürzung des Wortes führt. Es ist also unwahrscheinlich, dass ein Wort mit einer extrem hohen Frequenz für diese Frequenz zu lang ist. Das aber würde eine Regressionskurve implizieren, die auf nicht-gewichteten Fällen basiert.

Die Tabellen 5 bis 7 geben die Ergebnisse der Regression mit gewichteten Fällen wieder.

Tabelle 5
Anpassung der Funktion $LG = A \cdot F^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	8.2351	-0.0995	0.78	0.86	189	12861
WSJ1	8.6625	-0.0635	0.75	0.83	663	44349
WSJ2	8.6564	-0.0641	0.75	0.83	648	43157
TM	7.8068	-0.1467	0.86	0.92	112	4655

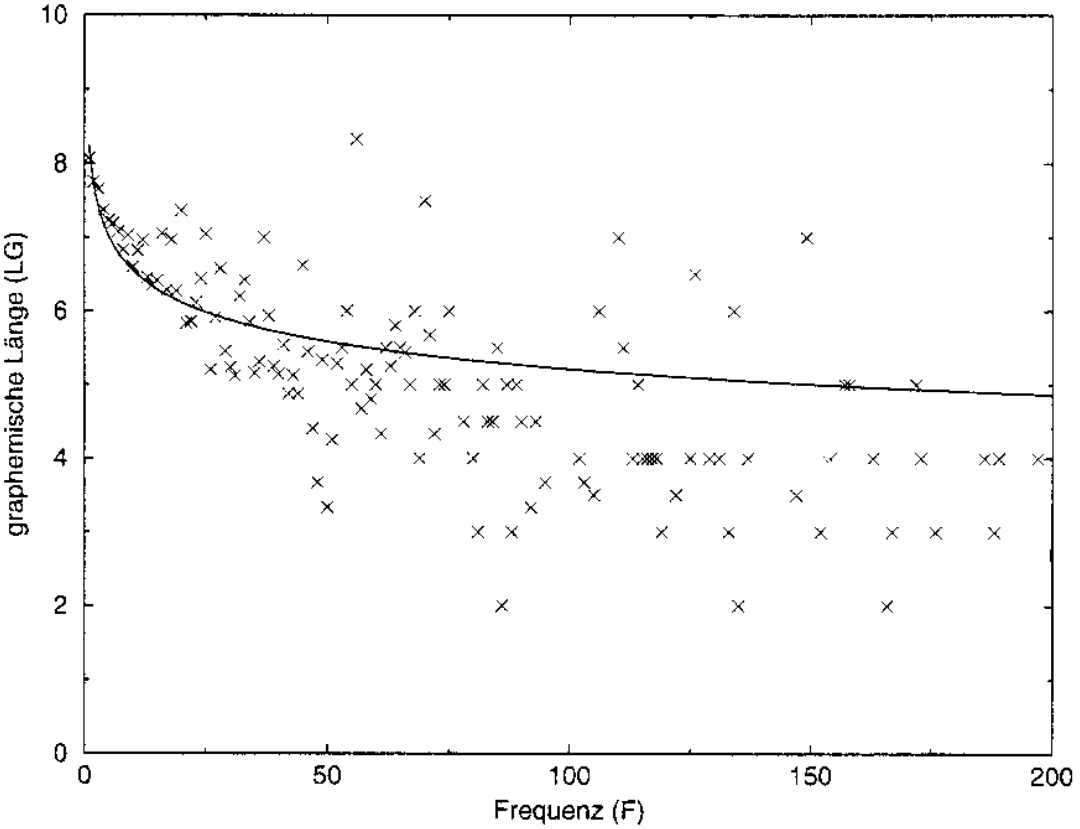

Abb. 6: Die Abhängigkeit der graph. Länge von der Frequenz (Ausschnitt)

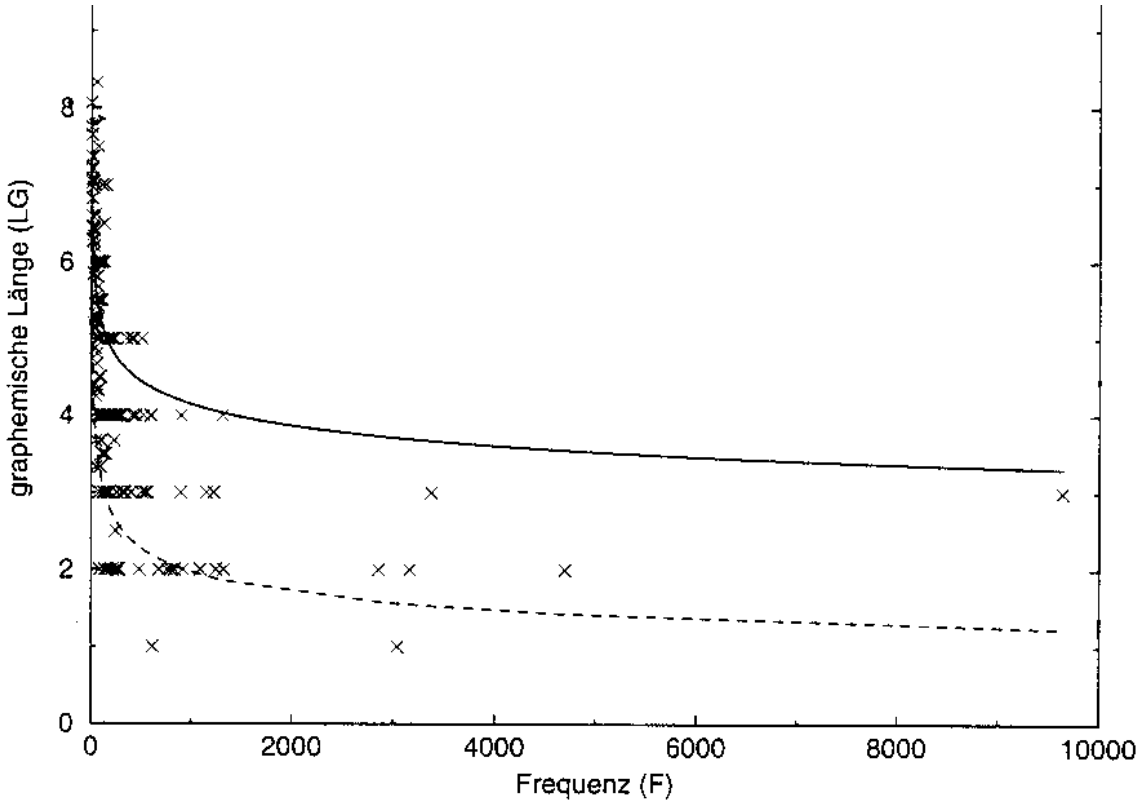

Abb. 7: Die Abhängigkeit der graph. Länge von der Frequenz (alle Daten)

Tabelle 6
Anpassung der Funktion $LP = A \cdot F^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	6.2997	-0.0799	0.65	0.72	189	10969
WSJ1	6.3522	-0.0370	0.38	0.41	654	23908
WSJ2	6.3205	-0.0371	0.39	0.42	638	23651
TM	5.9447	-0.1311	0.82	0.86	112	4103

Tabelle 7
Anpassung der Funktion $LS = A \cdot F^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	2.7602	-0.1250	0.83	0.87	189	7653
WSJ1	2.7665	-0.0655	0.57	0.61	619	16056
WSJ2	2.7635	-0.0676	0.57	0.62	607	15833
TM	2.4271	-0.1654	0.89	0.91	111	3165

Auffällig ist, dass die WSJ-Korpora schlechtere Ergebnisse liefern als das SUSANNE-Korpus und *The Time Machine*. Insbesondere ist dieses Phänomen bei der Silbenzahl und der phonemischen Länge zu beobachten. Da diese beiden Ausprägungen der Systemgröße Länge durch Nachschlagen im Lexikon gewonnen werden, also nur für diejenigen Lexeme aus den Korpora vorliegen, die auch im Lexikon enthalten sind, werden – bei gleich großem Wertebereich – für *LS* und *LP* weniger Wertepaare gemittelt als für *LG*; es gibt mehr Datenpunkte, die nur ein einziges Lexem repräsentieren. (Besonders bei den WSJ-Korpora, die einen extrem großen Wertebereich bei der Frequenz haben, ist das ein Problem.)

Zusätzlich problematisch ist z.B. bei der Silbenlänge, dass alle Lexeme mit einer Frequenz > ca. 250 (beim SUSANNE-Korpus) einsilbig sind, und so natürlich keine weitere Differenzierung mehr möglich ist. In einer solchen Situation verschlechtert jeder hinzukommende Datenpunkt den Determinationskoeffizienten.

Die Abhängigkeit der Frequenz von der Polytextie

Die vierte und letzte direkte Abhängigkeit ist nach dem Köhler-Modell die Abhängigkeit der durchschnittlichen Frequenz einer lexikalischen Einheit von ihrer Polytextie. In der entsprechenden Funktionsgleichung $F = A \cdot PT^B$ repräsentiert die Konstante *A* den Einfluß des Anwendungsbedürfnisses, die Konstante *B* stellt den Einfluß der Polytextie dar.

Abbildung 8 und Tabelle 8 geben die Ergebnisse der Regression mit gewichteten Fällen auf den Daten der drei Korpora wieder.

Tabelle 8
Anpassung der Funktion $F = A \cdot PT^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	1.3718	1.1567	0.99	0.82	64	12861
WSJ1	1.3198	1.0137	0.999	0.86	509	44349
WSJ2	1.3218	1.0115	0.999	0.86	510	43157

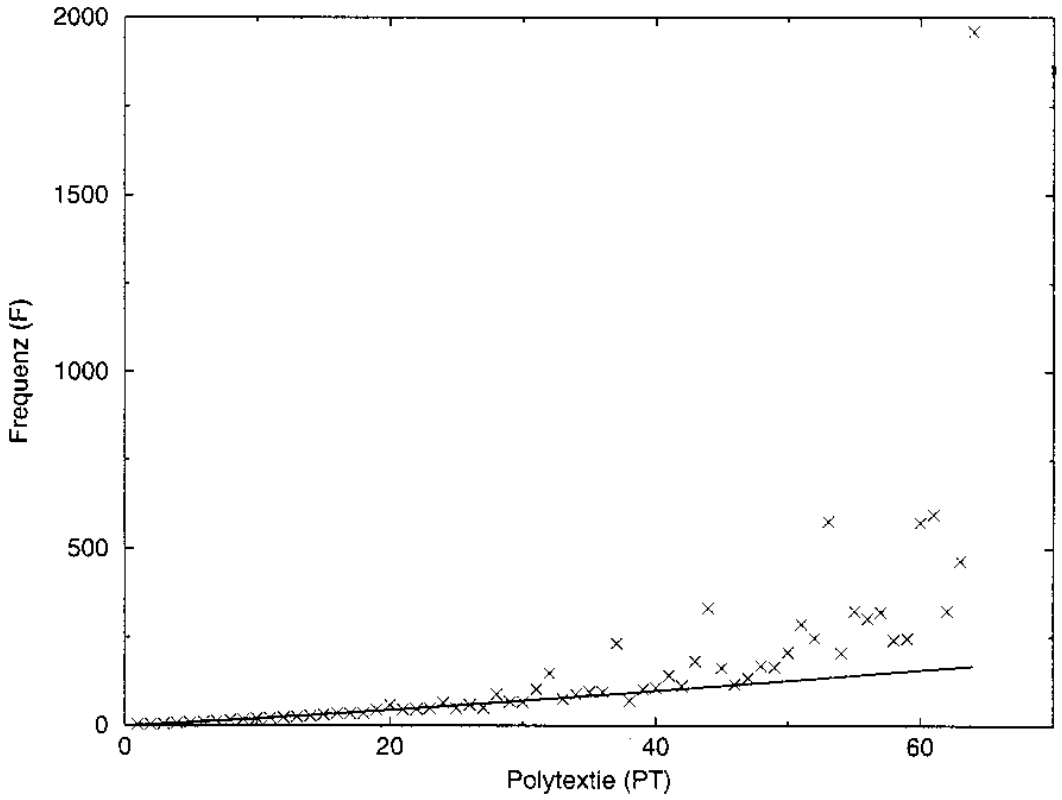

Abb. 8: Die Abhängigkeit der Frequenz von der Polylexie

Es fällt auf, dass der Determinationskoeffizient aus der nicht-linearen Regression für das SUSANNE-Korpus etwas schlechter ist als die entsprechenden Werte für die beiden WSJ-Korpora. Als Grund dafür wurde vermutet, dass die Polytextiewerte für das SUSANNE-Korpus bei der relativ kleinen Größe von 64 (also der Zahl der Einzeltexte im Korpus) enden. Das bedeutet, dass alle Lexeme, die in nahezu jedem existierenden englischen Text vorkommen, den Polytextie-

wert von 64 erhalten. Dieses Problem ergibt sich bei den WSJ-Korpora in wesentlich geringerem Maße. Durch die hohe Anzahl von je über 3000 (teils sehr kurzen) Einzeltexten wird die Menge der in nahezu allen Texten vorkommenden Lexeme stärker differenziert.

Um diese Vermutung zu überprüfen, wurde noch einmal eine Regression auf den Daten des SUSANNE-Korpus unter Auslassung des Wertepaares mit $PT = 64$ durchgeführt. Die Auslassung dieses Wertes bewirkte eine Verbesserung des Determinationskoeffizienten für das SUSANNE-Korpus (Tab. 9), so dass er danach den Determinationskoeffizienten für die WSJ-Korpora entsprach.

Tabelle 9
Anpassung der Funktion $F = A \cdot PT^B$ (modifizierte Daten)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	1.3817	1.1390	0.99	0.86	63	12842

An diesem Variablenzusammenhang wird besonders die Tendenz des linearen Modells sichtbar, große Schätzfehler (wie sie im hinteren Teil der Kurve auftreten) zu reduzieren: Die Linearisierung bewirkt, dass trotz der deutlichen Streuung ein Determinationskoeffizient von fast genau 1 erreicht wird.

2.5 Indirekte Abhängigkeiten

Aus der Struktur des Basismodells sind nicht nur die beschriebenen vier direkten Abhängigkeiten abzuleiten, sondern sie sagt auch indirekte Variablenzusammenhänge voraus. Um eine Funktionsgleichung für eine indirekte Abhängigkeit zu erhalten, werden jeweils zwei der vier bekannten Funktionsgleichungen herangezogen, wobei die unabhängige Variable der ersten Gleichung der abhängigen Variablen der zweiten Gleichung entspricht:

$$y = A_1 \cdot x^{B1}$$
$$x = A_2 \cdot Z^{B2}$$

Dann wird die rechte Seite der zweiten Gleichung in der ersten Gleichung für die unabhängige Variable eingesetzt:

$$y = A_1 \cdot (A_2 \cdot Z^{B2})^{B1}$$

$$A_1 \cdot A_2^{B1} \cdot Z^{B2 \cdot B1}$$

$$A_3 \cdot Z^{B3}$$

Insbesondere von Bedeutung ist der auf diese Weise erhaltene theoretische Parameter B_3. Er stellt eine numerische Vorhersage dar, die sich durch die empirischen Untersuchungen des Zusammenhangs zwischen *Y* und *Z* überprüfen läßt. Dabei bildet der Grad der Übereinstimmung zwischen dem aus den Funktionsgleichungen der direkten Abhängigkeiten theoretisch berechneten und dem empirischen Parameter ein Kriterium zur Beurteilung der Adäquatheit des lexikalischen Basismodells. Abschnitt 2.7 behandelt die statistische Überprüfung dieses Übereinstimmungsgrades.

Die beschriebene Vorgehensweise bei der empirischen Überprüfung der indirekten Abhängigkeiten soll hier anhand des Beispiels der Abhängigkeit der Polylexie von der Frequenz illustriert werden. Aus Platzgründen müssen die Ergebniswerte und -graphiken der anderen indirekten Abhängigkeiten unkommentiert bleiben.

Die Abhängigkeit der Polylexie von der Frequenz

Die indirekte Abhängigkeit der mittleren Polylexie von der Frequenz eines Lexems ergibt sich durch die Verknüpfung von

$$PL = A_1 \cdot L^{B1}$$

und

$$L = A_2 \cdot F^{B2}$$

zu

$$PL = A_3 \cdot F^{B3}.$$

Die Regression mit gewichteten Fällen führte zu den Ergebnissen in Tabelle 10 und Abbildung 9.

Tabelle 10
Anpassung der Funktion $PL = A \cdot F^B$ (Abb. 9 und 10)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	3.6616	0.2152	0.64	0.44	189	6953
WSJ1	2.4774	0.1778	0.65	0.41	619	14518
WSJ2	2.4898	0.1806	0.66	0.43	604	14310
TM	5.1356	0.1427	0.33	0.31	111	2937

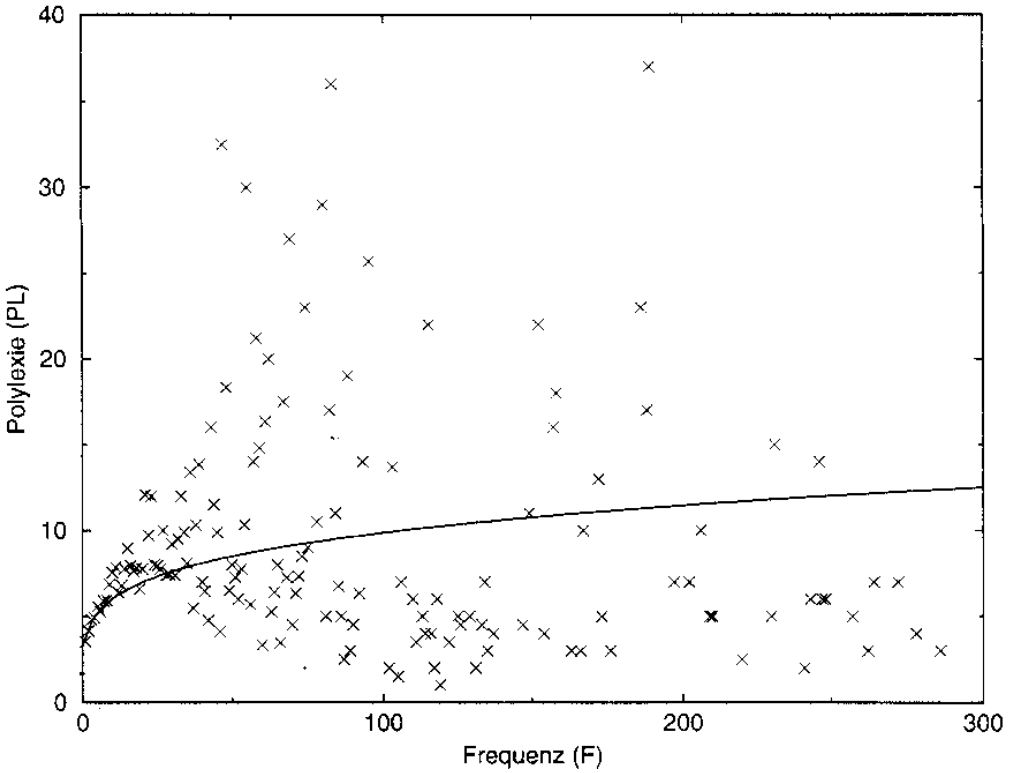

Abb. 9: Die Abhängigkeit der Polylexie von der Frequenz (Ausschnitt)

Hier tritt eine sehr große Streuung auf (sie ist noch größer als bei der umgekehrten Abhängigkeitsrichtung, s.u. Abb. 13), so dass die Graphik (unter Vernachlässigung der Gewichtung) eine fast waagerechte Regressionsgerade nahelegt.

Durch die Gewichtung jedoch erhält man eine Regressionskurve, die – wie die inverse Abhängigkeit – den größeren Teil der Streuung unter sich läßt und deren Anpassung im Bereich der niedrigen Frequenzen, die ja die Mehrzahl der Lexeme repräsentieren, ziemlich gut ist (Abbildung 10 zeigt die Regressionskurve für die Polylexiewerte von $F = 1$ bis $F = 15$; sie repräsentieren ca. 88% der Datenpunkte).

Die Lexeme mit den höchsten Frequenzen, die fast alle unterhalb der Kurve streuen, sind zum größten Teil Funktionswörter, auf sie könnte also das schon bekannte lexikographische Problem mit Funktionswörtern zutreffen. Dazwischen gibt es allerdings eine große Zahl von Datenpunkten, deren große Streuung unerklärt bleibt.

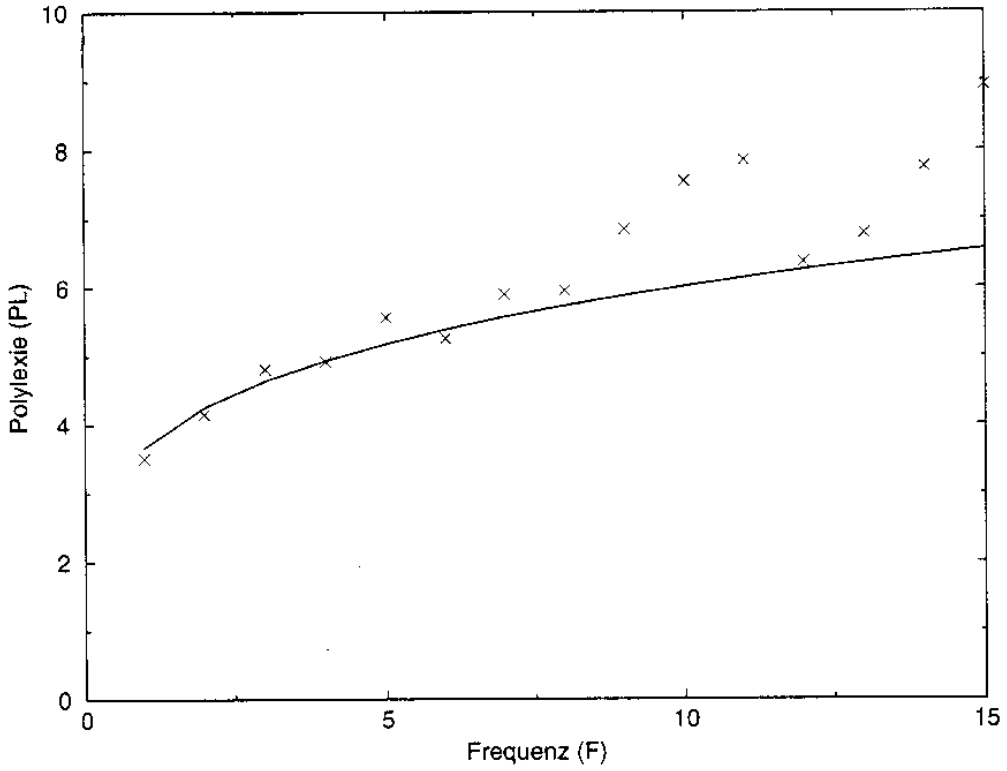

Abb. 10: Die Abhängigkeit der Polylexie von der Frequenz (Ausschnitt)

Die Abhängigkeit der Länge von der Polytextie

Tabelle 11
Anpassung der Funktion $LG = A \cdot PT^B$ (Abb. 11)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	8.1271	-0.1219	0.82	0.89	64	12861
WSJ1	8.5318	-0.0590	0.72	0.82	509	44349
WSJ2	8.5284	-0.0600	0.72	0.83	510	43157

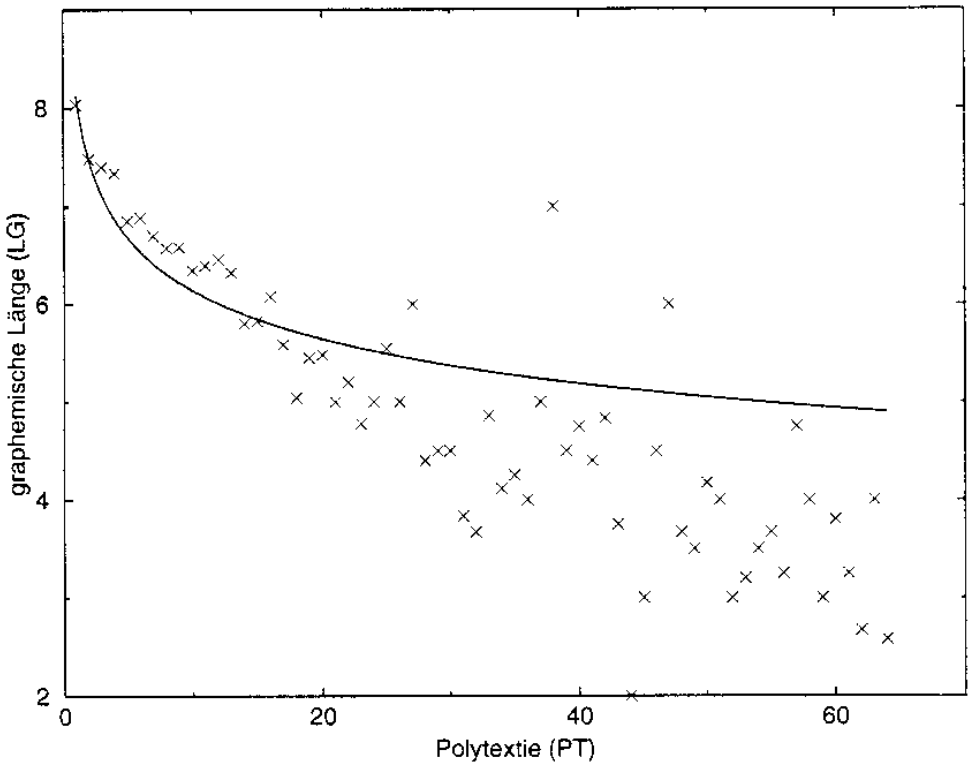

Abb. 11: Die Abhängigkeit der graphemischen Länge von der Polytextie

Tabelle 12
Anpassung der Funktion LS = $A \cdot PT^B$ (o. Abb.)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	2.7177	-0.1565	0.87	0.91	64	7653
WSJ1	2.7396	-0.0658	0.58	0.64	491	16058
WSJ2	2.7360	-0.0679	0.58	0.65	489	15834

Die Abhängigkeit der Polytextie von der Länge

Tabelle 13
Anpassung der Funktion $PT = A \cdot LG^B$ (Abb. 12)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	33.8622	-1.2582	0.95	0.94	21	12856
WSJ1	341.5862	-1.7371	0.96	0.84	28	44349
WSJ2	346.9567	-1.7457	0.96	0.82	30	43157

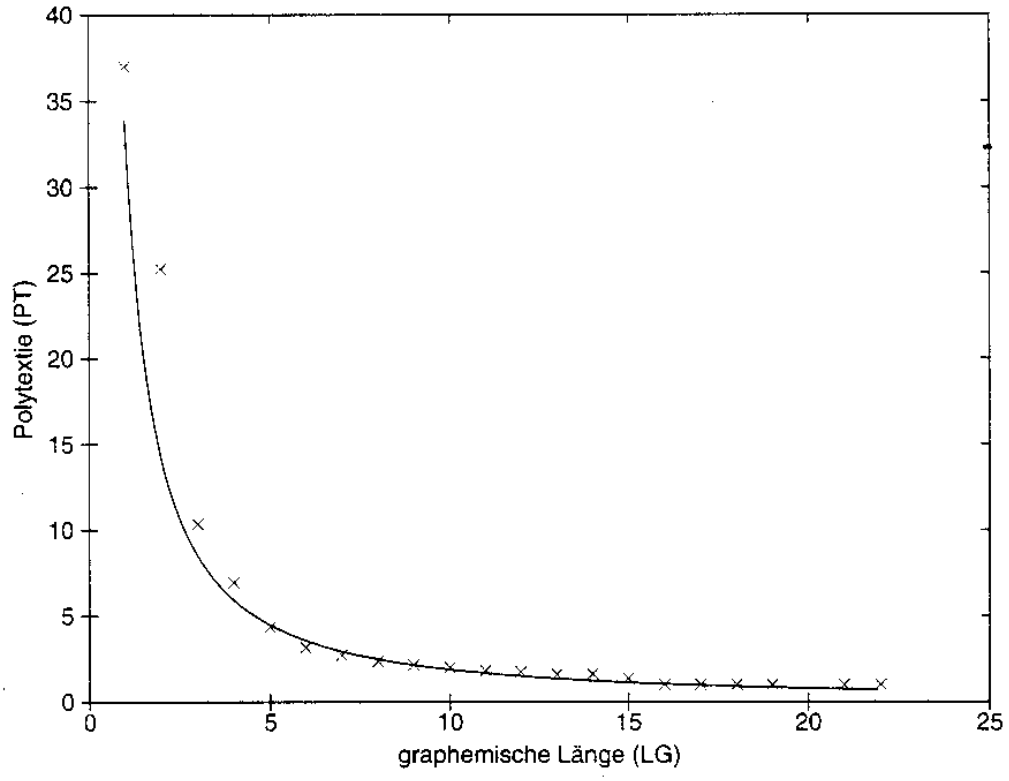

Abb. 12: Die Abhängigkeit der Polytextie von der graphemischen Länge

Tabelle 14
Anpassung der Funktion $PT = A \cdot LS^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	7.3368	-0.9034	0.98	0.99	8	7653
WSJ1	59.7578	-1.2905	0.95	0.98	8	16058
WSJ2	58.8093	-1.2878	0.95	0.98	8	15834

Die Abhängigkeit der Frequenz von der Polylexie

Tabelle 15
Anpassung der Funktion $F = A \cdot PL^B$ (Abb. 13)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	6.9755	0.4695	0.53	0.33	55	6953
WSJ1	19.7174	0.7367	0.83	0.37	55	14518
WSJ2	19.6544	0.7356	0.80	0.38	55	14310
TM	5.6333	0.2761	0.23	0.18	54	2937

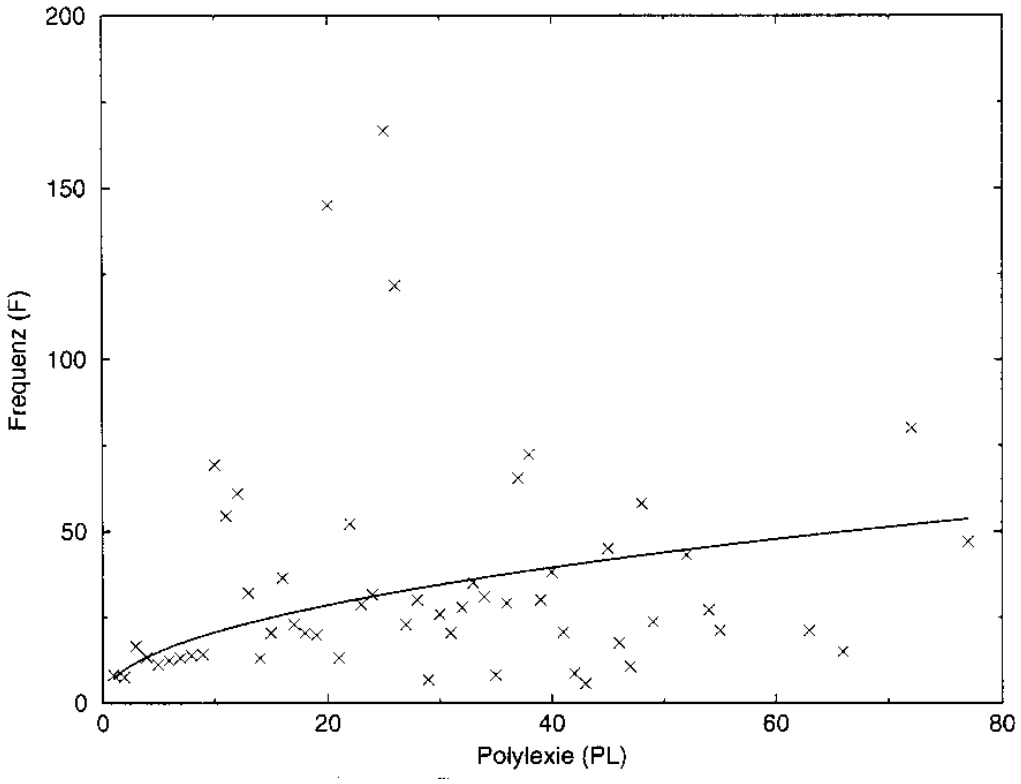

Abb. 13: Die Abhängigkeit der Frequenz von der Polylexie

2.6 Doppelt indirekte Abhängigkeiten

Analog zum beschriebenen Vorgehen bei den indirekten Variablenzusammenhängen lassen sich auch doppelt indirekte Abhängigkeiten aus dem Basismodell deduzieren, und die numerische Parameterausprägung läßt sich für sie vorhersagen. Dazu werden jeweils drei Funktionsgleichungen für direkte Abhängigkeiten ineinander eingesetzt:

$$y = A_1 \cdot x^{B1}$$

$$x = A_2 \cdot W^{B2}$$

$$W = A_3 \cdot Z^{B3}$$

$$= y = A_1 \cdot A_2^{B1} \cdot A_3^{B2 \cdot B1} \cdot Z^{B3 \cdot B2 \cdot B1}$$

$$A_4 \cdot Z^{B4}$$

Der Grad der Übereinstimmung der theoretisch abgeleiteten Parametervorhersagen mit den empirischen Werten wird wieder mit dem gleichen statistischen Test überprüft, der schon für die einfach indirekten Abhängigkeiten verwendet wurde (siehe 2.7). Dabei ist zu erwarten, dass die über zwei vermittelnde Systemgrößen gewonnenen Parametervorhersagen deutlicher von den empirischen Werten abweichen, als das bei den einfach indirekten Abhängigkeiten der Fall ist.

Wie schon bei den einfach indirekten Abhängigkeiten soll hier die Vorgehensweise an einem Variablenzusammenhang illustriert werden. Von den anderen Abhängigkeiten werden wieder lediglich die Ergebniswerte und -graphiken unkommentiert angegeben.

Die Abhängigkeit der Polytextie von der Frequenz

Die Verknüpfung der drei direkten Abhängigkeiten $PT = A \cdot PL^B$, $PL = A \cdot L^B$ und $L = A \cdot F^B$ führt zu $PT = A \cdot F^B$. Die Ergebnisse der empirischen Untersuchungen sind in Tabelle 16 und Abbildung 14 wiedergegeben.

Tabelle 16
Anpassung der Funktion $PT = A \cdot F^B$ (Abb. 14 (SUS) und 15 (WSJ1))

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	0.9945	0.7434	0.99	0.77	189	12861
WSJ1	0.9546	0.9149	0.997	0.86	663	44349
WSJ2	0.9564	0.9151	0.997	0.87	648	43157

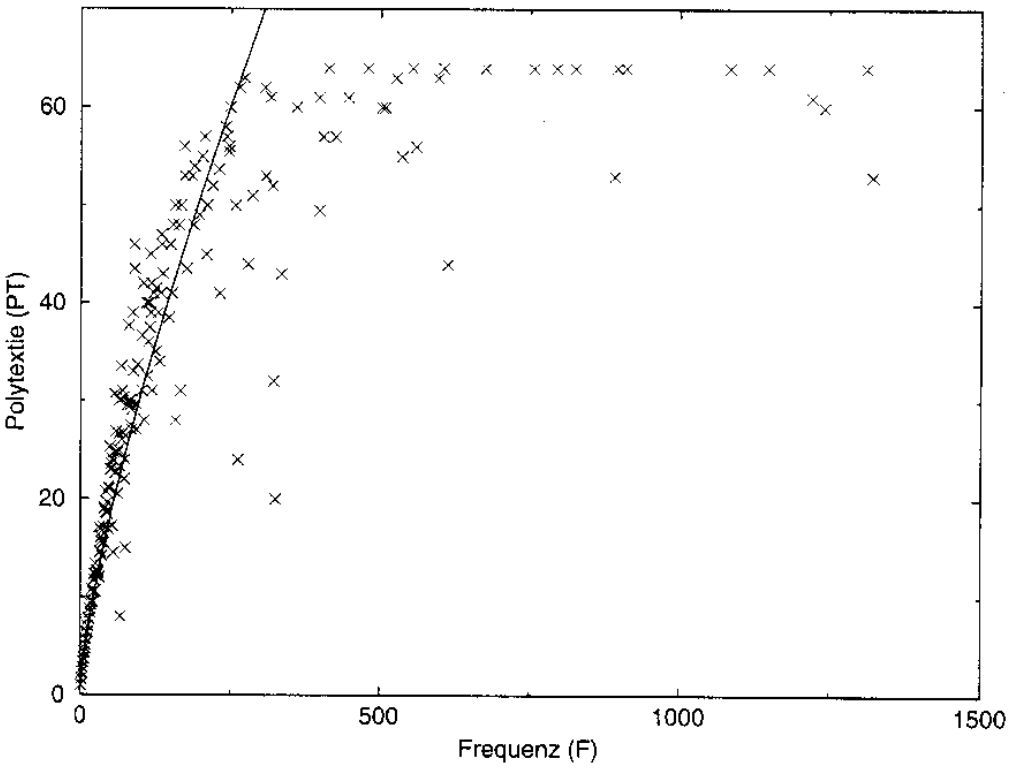

Abb. 14: Die Abhängigkeit der Polytextie von der Frequenz (Ausschnitt)

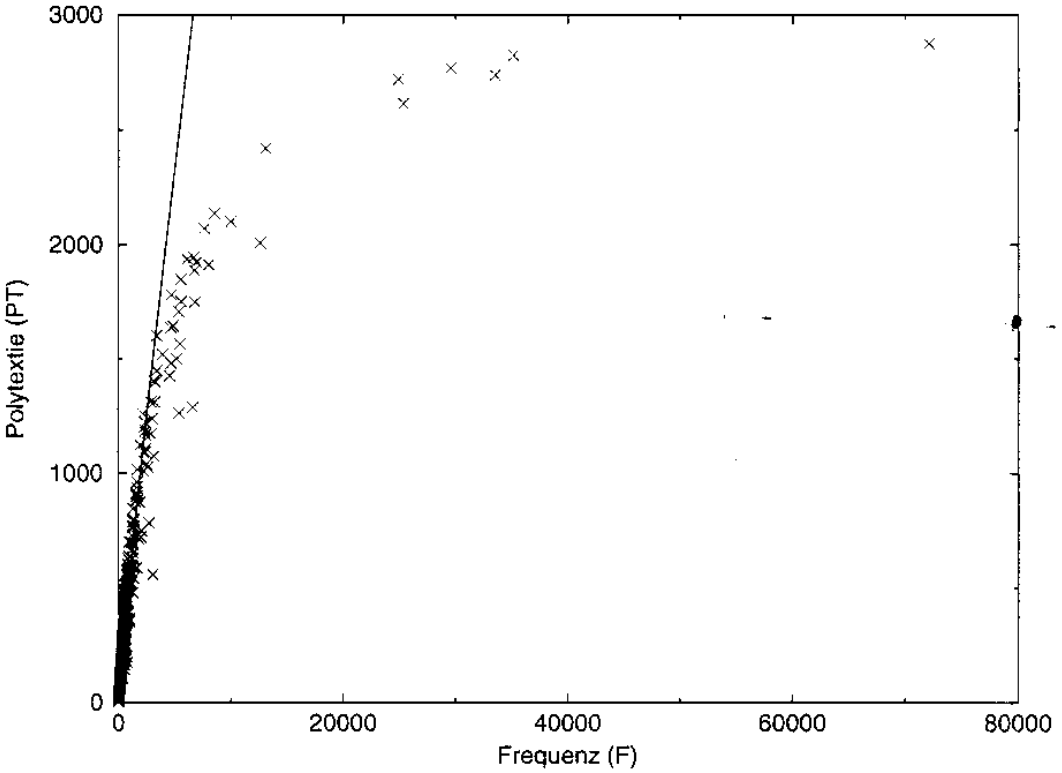

Abb. 15: Die Abhängigkeit der Polytextie von der Frequenz (WSJ, Teil 1)

Die Werte für den Determinationskoeffizienten sind ziemlich gut. Wie schon bei der inversen Abhängigkeit (vgl. Tab. 9 und Abb. 8) ist auch hier das Ergebnis im nicht-linearen Modell für das SUSANNE-Korpus etwas schlechter als für die beiden WSJ-Korpora.

Deshalb wurden wieder testhalber die Daten für das SUSANNE-Korpus unter Auslassung aller Wertepaare, deren Polytextie 64 beträgt, noch einmal untersucht. Dieses Verfahren erbrachte bessere (und wieder den Werten für die WSJ-Korpora entsprechenden) Determinationskoeffizienten (Tab.17).

Tabelle 17
Anpassung der Funktion $PT = A \cdot F^B$ (modifizierte Daten)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	1.0125	0.7532	0.998	0.89	170	12842

Auch an dieser Abhängigkeit wird deutlich, wie sehr im linearisierten Modell die Streuung im oberen Bereich der Kurve, die zu großen Schätzfehlern führt, vernachlässigt wird: trotz dieser unübersehbaren Streuung werden Determinationskoeffizienten von nahe 1 erreicht.

Zum Vergleich zu Abbildung 14 zeigt Abbildung 15 die gleiche Abhängigkeit berechnet auf den Daten des Wall Street Journals (Teil 1). Da bei diesem Korpus die Anzahl der Einzeltexte sehr viel höher ist als beim SUSANNE-Korpus kommt es im Bereich der hochfrequenten Wörter zu einer stärkeren Differenzierung. Die Kurve endet nicht abrupt bei ‚*PT* = Anzahl der Einzeltexte im Korpus', sondern läuft 'natürlich' aus.

Die Abhängigkeit der Polylexie von der Polytextie

Tabelle 18
Anpassung der Funktion $PL = A \cdot PT^B$ (Abb. 16)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	3.6958	0.3068	0.85	0.75	64	6953
WSJ1	2.4532	0.2053	0.74	0.50	490	14520
WSJ2	2.4729	0.2058	0.73	0.47	489	14311

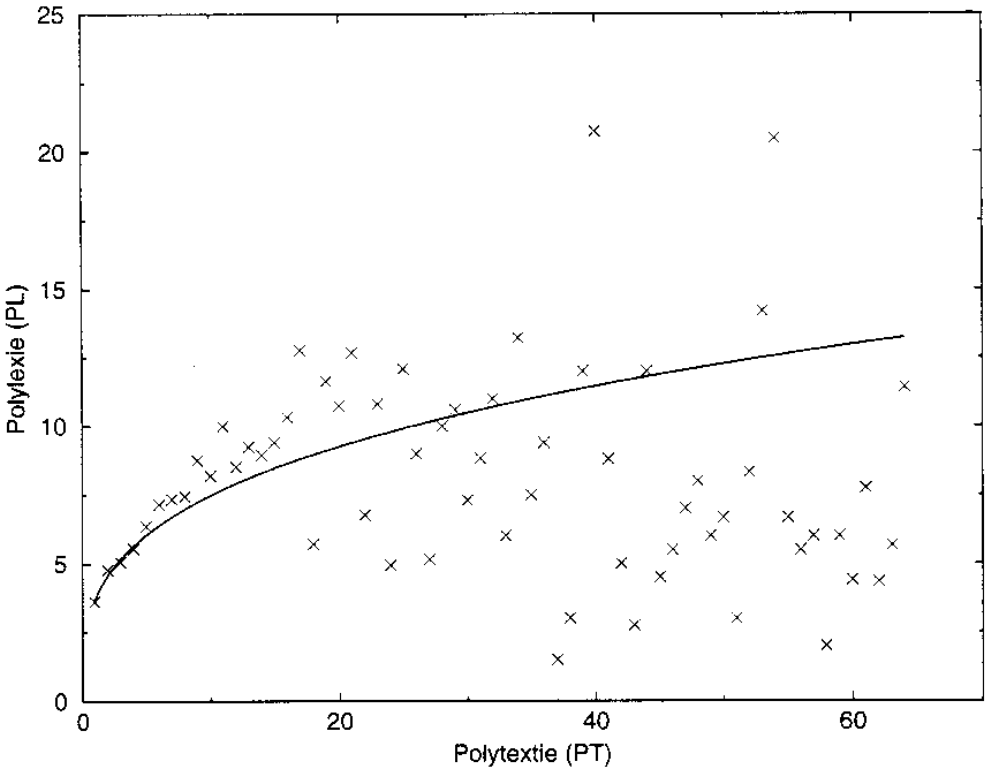

Abb. 16: Die Abhängigkeit der Polylexie von der Polytextie

Die Abhängigkeit der Länge von Polylexie

Tabelle 19
Anpassung der Funktion $LG = A \cdot PL^{B}$ (Abb. 17)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	8.0575	-0.1194	0.77	0.77	55	6953
WSJ1	7.7423	-0.0935	0.71	0.73	55	14520
WSJ2	7.7284	-0.0941	0.71	0.73	55	14311
TM	7.6461	-0.1177	0.76	0.76	54	2937

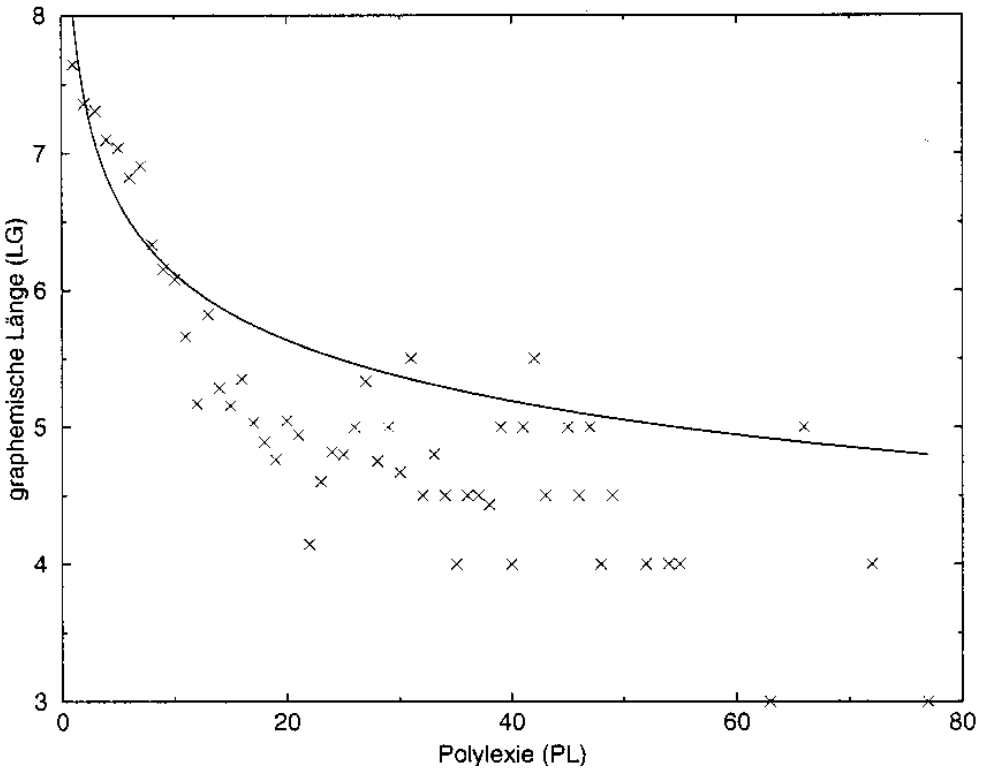

Abb. 17: Die Abhängigkeit der graphemischen Länge von der Polylexie

Tabelle 20
Anpassung der Funktion $LS = A \cdot PL^{B}$ (o. Abb.)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	2.7871	-0.1640	0.65	0.64	55	6953
WSJ1	2.6591	-0.1330	0.64	0.65	55	14520
WSJ2	2.6525	-0.1336	0.64	0.66	55	14311
TM	2.5439	-0.1596	0.64	0.61	54	2937

Die Abhängigkeit der Frequenz von der Länge

Tabelle 21
Anpassung der Funktion $F = A \cdot LG^B$

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	293.096	-2.0557	0.86	0.85	21	12856
WSJ1	1203.874	-2.1728	0.94	0.73	28	44349
WSJ2	1224.025	-2.1817	0.95	0.71	30	43157
TM	221.097	-2.1655	0.89	0.97	19	4655

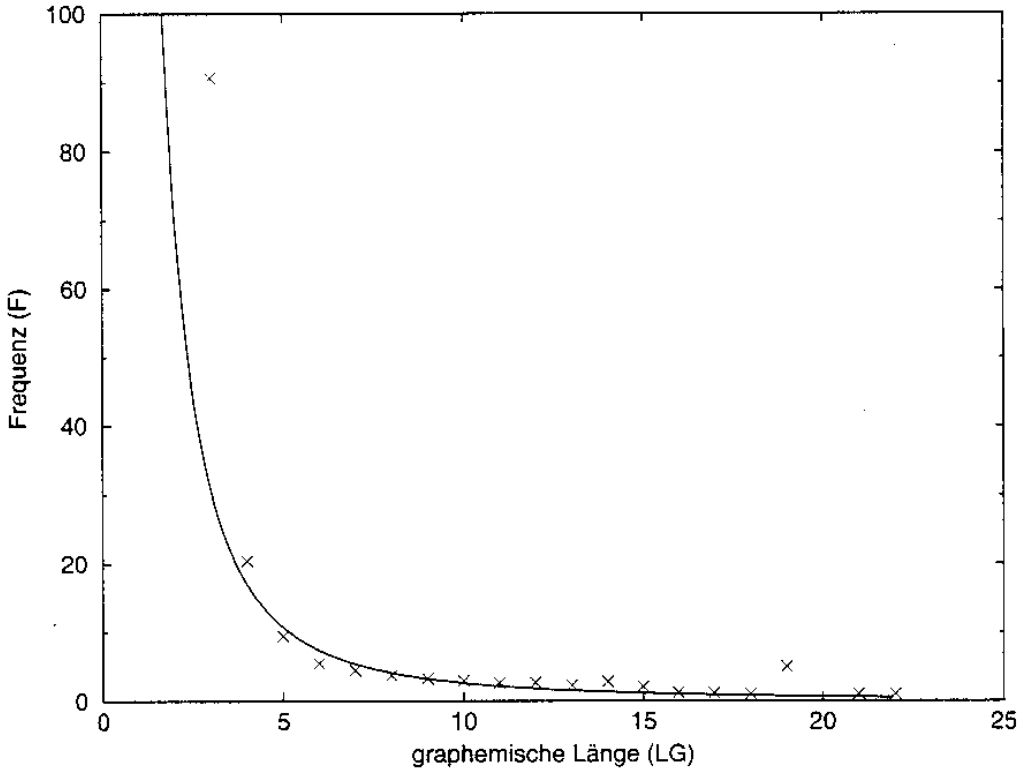

Abb. 18: Die Abhängigkeit der Frequenz von der graphemischen Länge (Ausschnitt)

Tabelle 22
Anpassung der Funktion $F = A \cdot LS^B$ (o. Abb.)

Korpus	*A*	*B*	r^2 (*LR*)	r^2 (*NLR*)	*# MW*	*# Types*
SUS	32.3787	-1.7753	0.91	0.99	8	7653
WSJ1	125.5370	-1.6575	0.92	0.99	8	16056
WSJ2	129.2308	-1.7001	0.92	0.99	8	15833
TM	18.7127	-1.8729	0.94	0.996	7	3165

2.7 Vergleich der theoretischen und empirischen Parameter bei indirekten Abhängigkeiten

Die Ähnlichkeit des theoretischen Parameters b mit dem empirischen Parameter B der indirekten und doppelt indirekten Abhängigkeiten des Basismodells wird mit der Nullhypothese $B - b = 0$ überprüft; es wird also angenommen, dass die Parameter identisch sind. Zum Test dieser Nullhypothese wird definiert:

$$t = \frac{B - b}{s_B}.$$

Diese Definition ist identisch mit der aus Abschnitt 2.3. Der einzige Unterschied in der Durchführung des Tests besteht darin, dass hier für b nicht 0, sondern der aus den direkten Abhängigkeiten berechnete theoretische Wert für den Regressionskoeffizienten eingesetzt wird (vgl. 2.5 und 2.6).

Ist das Ergebnis ein t-Wert, der bei einem Signifikanzniveau von 0.01 über dem Tabellenwert von t mit n-2 Freiheitsgraden liegt (n ist wieder die Summe aller Gewichtungen), bedeutet das, dass es einen signifikanten Unterschied zwischen den zwei Parametern gibt. Liegt der Wert für t unter dem Tabellenwert, wird angenommen, dass sich die beiden Parameter nicht signifikant unterscheiden.

Die Anwendung des t-Tests auf Daten des SUSANNE-Korpus (Tab. 23) ergab, dass bei allen indirekten und doppelt indirekten Abhängigkeiten der Unterschied zwischen dem theoretisch vorhergesagten und dem empirisch ermittelten Regressionskoeffizienten auf dem 0.01-Niveau signifikant war. Dabei unterschieden sich auch die Parameter signifikant, die numerisch sehr ähnlich waren. Deshalb wurden zunächst – unter der Annahme, dass das Modell korrekt ist – zwei mögliche Gründe dafür vermutet:

1. Je nach Variablenzusammenhang und Korpusgröße werden unterschiedliche Datenpunkte unterschiedlich stark gewichtet. In die indirekten Beziehungen

über eine oder zwei vermittelnde Variablen gehen also in die theoretische Funktion andere Gewichtungen ein, als das bei der empirischen Funktion der Fall ist. Dies erschwert die Vergleichbarkeit des theoretischen Parameters mit dem empirischen Parameter.

2. Bei Gewichtung der Datenpunkte wird eine sehr hohe Schätzgenauigkeit errechnet, weil die Summe der Gewichtungen als Anzahl der Datenpunkte interpretiert wird. Tatsächlich werden aber wesentlich weniger Datenpunkte verwendet. Die hohe Schätzgenauigkeit hat zur Folge, dass der Standardfehler des Regressionskoeffizienten s_B äußerst klein wird. Wenn aber s_B (also der Nenner in der Berechnung des *t*-Werts) sehr klein ist, wird selbst ein ganz geringer Unterschied zwischen den Parametern statistisch signifikant.

Tabelle 23
Theoretische und empirische Parameter der indirekten und doppelt indirekten Abhängigkeiten (SUSANNE-Korpus).

Abhängigkeit	*b* (*theoret.*)	*B* (*empir.*)
$F = A \cdot PL^B$	0.4341	0.4695
$PL = A \cdot F^B$	0.0791	0.2152
$LG = A \cdot PT^B$	-0.1151	-0.1219
$LP = A \cdot PT^B$	-0.0924	-0.0998
$LS = A \cdot PT^B$	-0.1446	-0.1562
$PT = A \cdot LG^B$	-0.7951	-1.2582
$PT = A \cdot LP^B$	-0.2696	-0.9212
$PT = A \cdot LS^B$	-0.2327	-0.9034
$F = A \cdot LG^B$	-0.3452	-2.0557
$F = A \cdot LP^B$	-0.3118	-1.6841
$F = A \cdot LS^B$	-0.2691	-1.7753
$PT = A \cdot F^B$	0.0235	0.7434
$PL = A \cdot PT^B$	0.0723	0.3068
$LG = A \cdot PL^B$	-0.0432	-0.1194
$LP = A \cdot PL^B$	-0.0347	-0.1328
$LS = A \cdot PL^B$	-0.0543	-0.1640

Köhler (1986) führte den Test der Nullhypothese auf der Grundlage nicht-gewichteter Mittelwerte aus. Bei nicht-gewichteten Daten treten die beiden oben genannten Probleme nicht auf. Deshalb wurden fürs Englische zusätzlich die theoretischen und empirischen Parameter aus nicht-gewichteten Mittelwerten berechnet und der *t*-Test noch einmal auf diesen Daten ausgeführt. Doch wie schon

bei den gewichteten Mittelwerten waren im Englischen – anders als im Deutschen – auch bei Verwendung nicht-gewichteter Daten auf dem 0.01-Niveau alle beobachteten Unterschiede zwischen theoretischen und empirischen Parametern signifikant.

Diese Ergebnisse werfen Probleme auf. Einerseits unterscheiden sich die empirischen und theoretischen Parameter signifikant. Sie können also das hier untersuchte Modell im Hinblick auf die Vorhersagbarkeit bestimmter Parameterwerte nicht bestätigen, die Nullhypothese $B - b = 0$ muß zurückgewiesen werden.

Auf der anderen Seite stimmen alle empirischen und theoretischen Parameter in ihrer Ausrichtung überein und haben teilweise numerisch sehr ähnliche Werte, so dass das Zurückweisen der Nullhypothese nicht gleichzeitig die Ablehnung des Modells bedeuten sollte.

Trotzdem darf man annehmen, dass das Modell einer Verfeinerung bzw. Modifikation bedarf, um für das Englische nicht nur die Ausrichtung indirekter Parameter korrekt vorherzusagen, sondern auch ihre numerische Ausprägung. Immerhin sind die empirischen Regressionskoeffizienten teilweise ein Vielfaches der theoretischen Regressionskoeffizienten. Zu überdenken wäre z.B. die Art und Weise, auf die die theoretischen Regressionskoeffizienten für die indirekten Abhängigkeiten berechnet werden. Man erhält sie durch Multiplikation der empirischen Regressionskoeffizienten der beteiligten direkten Abhängigkeiten. Da die Regressionskoeffizienten der direkten Abhängigkeiten bis auf einen kleiner als 1 sind, werden dadurch die theoretischen Regressionskoeffizienten der indirekten und doppelt indirekten Abhängigkeiten immer kleiner. Das entspricht nicht den empirischen Befunden.

Eine andere Möglichkeit der Modifikation besteht in der Einbeziehung von direkten Rückwirkungen der Systemgrößen aufeinander, die von Köhler (1986) zwar schon beschrieben, aber noch nicht explizit modelliert worden sind. Beispielsweise wirkt ja nicht nur die Frequenz eines Lexems dahingehend auf seine Länge, dass das Lexem bei häufigem Gebrauch gekürzt wird (z.B. *Universität* → *Uni*), sondern andersherum kann auch die Länge eines Lexems Einfluß auf seine Frequenz nehmen (z.B. wenn sich viele Sprecher bei gleichzeitiger Verfügbarkeit der Lexeme *Universität* und *Uni* aus Ökonomiegründen für das kürzere entscheiden, was dessen Frequenz erhöht).

2.8 Überprüfung der Stichproben auf Homogenität

Die Regressionsanalyse mit den Daten der drei untersuchten Korpora und des einen untersuchten längeren Textes erbrachte für die gleiche Abhängigkeit jeweils verschiedene numerische Parameterwerte. Diese Unterschiede können zufällig sein, sie können aber auch darauf schließen lassen, dass die verschiedenen Stichproben keiner homogenen Grundgesamtheit entstammen.

Diese Frage kann mit Hilfe eines statistischen Tests beantwortet werden. Dabei wird der Parameter *A* außer acht gelassen, weil er nur die Parallelverschiebung der Regressionsgeraden auf der *Y*-Achse beschreibt. Untersucht wird der Regressionskoeffizient *B*, der den Steigungswinkel der Geraden bezeichnet. Dabei lautet die Nullhypothese H_0: $b_1 = b_2 = b_3 = b_4$, es wird also angenommen, dass sich die Regressionskoeffizienten für die vier Korpora/Texte nicht signifikant unterscheiden.

Der Test, der hier zur Überprüfung der Nullhypothese verwendet wird, orientiert sich an Edwards (1976; 109-12).[7] Er kann an dieser Stelle aus Platzgründen nicht näher beschrieben werden, deshalb sei nur gesagt, dass im Verlauf des Tests aus verschiedenen Größen der statistischen Analyse der einzelnen Korpora ein *F*-Wert berechnet wird, der anschließend mit dem kritischen Tabellenwert der *F*-Verteilung verglichen werden kann. Bleibt der für den jeweiligen Variablenzusammenhang errechnete *F*-Wert unter dem Tabellenwert, darf man darauf schließen, dass es keine signifikante Evidenz gegen die Annahme der Nullhypothese $b_1 = b_2 = b_3 = b_4$ gibt.

Bei der Untersuchung sämtlicher Abhängigkeiten im lexikalischen Basismodell ergab sich, dass bis auf drei Abhängigkeiten alle auf dem 0.05-Niveau in bezug auf ihre Regressionskoeffizienten keine signifikanten Unterschiede aufwiesen. Von diesen drei Abhängigkeiten zeigten zwei ($F = A \cdot PT^B$ und $PT = A \cdot LG^B$) auf dem 0.01-Niveau keinen signifikanten Unterschied, und nur bei einer Abhängigkeit ($PT = A \cdot F^B$) stellte sich heraus, dass sich die verschiedenen Regressionskoeffizienten auf dem 0.01-Niveau signifikant unterschieden. Bezeichnenderweise enthält diese Abhängigkeit Variablen, deren Ausprägung zum großen Teil von der Korpusgröße und -struktur abhängt.

Insgesamt kann aufgrund dieser Ergebnisse behauptet werden, dass die vier verwendeten Stichproben einer gemeinsamen Population entstammen und dass die Unterschiede in den Parametern zufällig oder von der Art des Textes beeinflußt sind, wobei die Textsorte aber keinen *signifikanten* Einfluß auf die Parameter hat.

---

[7] Der hier verwendete Test eignet sich nur für nicht-gewichtete Daten. Deshalb wurden als Grundlage die nicht-gewichteten Mittelwerte verwendet. Zwar unterscheiden sich die Regressionskoeffizienten der nicht-gewichteten Mittelwerte von denen der gewichteten Mittelwerte, doch kommt es an dieser Stelle nicht auf eine möglichst gute Vorhersage an, sondern nur auf den Vergleich der Vorhersagen, die mit den Regressionskoeffizienten der verschiedenen Korpora erzielt werden. Da sich die Regressionskoeffizienten der nicht-gewich-teten Mittelwerte in der Bandbreite nicht wesentlich von denen der gewichteten Mittelwerte unterscheiden, erscheint es auch plausibel, bei Homogenität der einen Gruppe von Regressionskoeffizienten auf Homogenität der anderen Gruppe zu schließen.

2.9 Zusammenfassung der Ergebnisse der empirischen Untersuchungen

Die empirischen Untersuchungen haben alle vom lexikalischen Basismodell vorhergesagten direkten und indirekten Variablenzusammenhänge bestätigt (2.4 bis 2.6). Dabei unterschieden sich die Ergebnisse für die verschiedenen Stichproben der englischen Sprache nicht signifikant (2.8).

Die empirischen Untersuchungen konnten jedoch nicht zeigen, dass sich die numerischen Vorhersagen für die Parameterwerte der indirekten Variablenzusammenhänge den entsprechenden empirischen Parameterwerten ausreichend annäherten (2.7). Unter der Annahme der prinzipiellen Korrektheit des Modells wurde auf Möglichkeiten hingewiesen, das lexikalische Basismodell zu verfeinern bzw. zu modifizieren, um zu einer besseren Vorhersage der Parameter zu kommen.

Grundsätzlich darf aber auch nicht die Möglichkeit außer acht gelassen werden, dass die mangelnde numerische Übereinstimmung zwischen den theoretischen und den empirischen Parametern ein Indiz dafür ist, dass in der Sprache allgemein oder speziell im Englischen andere Zusammenhänge zwischen den Systemgrößen vorliegen, als sie aus dem lexikalischen Basis-modell deduziert werden können.

Bei der empirischen Überprüfung des Modells am Englischen ergaben sich für die doppelt indirekten Abhängigkeiten oft höhere Determinationskoeffizienten als für die (einfach) indirekten oder gar direkten Abhängigkeiten. Das widerspricht der Erwartung, dass die Determinationskoeffizienten für die direkten Abhängigkeiten die besten sind und sich zu den doppelt indirekten Abhängigkeiten hin verschlechtern. Auch dies könnte möglicherweise ein Hinweis darauf sein, dass die Modellierung der Beziehungen zwischen den Systemgrößen im lexikalischen Basismodell einer Modifikation bedarf.

Bei allen Änderungen, die am Basismodell vorgenommen werden, ist es entscheidend, dass sie durch theoretische Überlegungen motiviert sind. Ein Alternativmodell soll nicht nur die Anpassung der theoretischen Kurven an die Daten verbessern, was sich zum Beispiel durch das Hinzufügen weiterer Parameter trivialerweise erreichen läßt, sondern es muß für Veränderungen in den Funktionsgleichungen auch linguistische Erklärungen bereitstellen.

3. Mögliche methodische Erweiterungen

In allen hier überprüften Abhängigkeiten wurde – wie in der quantitativen Linguistik nicht unüblich – mit Mittelwerten der abhängigen Variablen gearbeitet. Dabei wurden die entstandenen Datenpunkte mit der Anzahl der Einzelpunkte, aus denen sie sich zusammensetzten, gewichtet.

Neben den erwähnten Problemen mit Gewichtung oder Nicht-Gewichtung der Fälle birgt die Methode der Mittelwertbildung das Defizit, dass bei ihrer An-

wendung die Varianz in den Daten nicht berücksichtigt wird. Das folgende Beispiel soll die Problematik veranschaulichen:

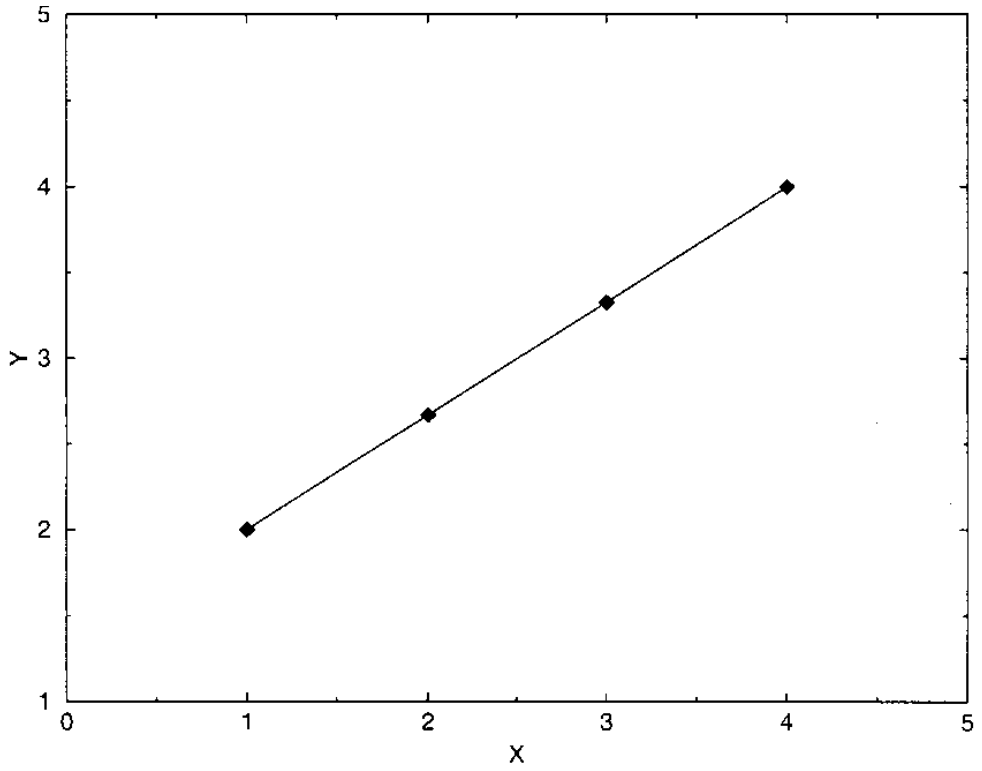

Abb. 19

Abbildung 19 zeigt eine durch lineare Regression ermittelte Regressionsgerade mit einem Determinationskoeffizienten von 1.00. Die Datenpunkte, die hier Mittelwerte repräsentieren sollen, liegen exakt auf der Geraden, die Anpassung ist also perfekt.

Diese Mittelwerte können aus unendlich vielen Mengen von Rohdaten errechnet werden. Abbildungen 20a und 20b zeigen zwei dieser möglichen Menge.

Die Regressionsanalyse auf den linearisierten Daten aus Abbildung 20a ergibt einen Determinationskoeffizienten von 0.97, auf den Daten aus Abbildung 20b führt sie zu einem Determinationskoeffizienten von 0.01. Trotz des großen Unterschieds in den Rohdaten erhält man bei Mittelwertbildung für beide Mengen den selben und sogar einen perfekten Determinationskoeffizienten, der ja höchstens im Fall von Abbildung 20a gerechtfertigt erscheint. Nach dem gleichen Muster sind sicherlich noch extremere Verteilungen denkbar bzw. konstruierbar.

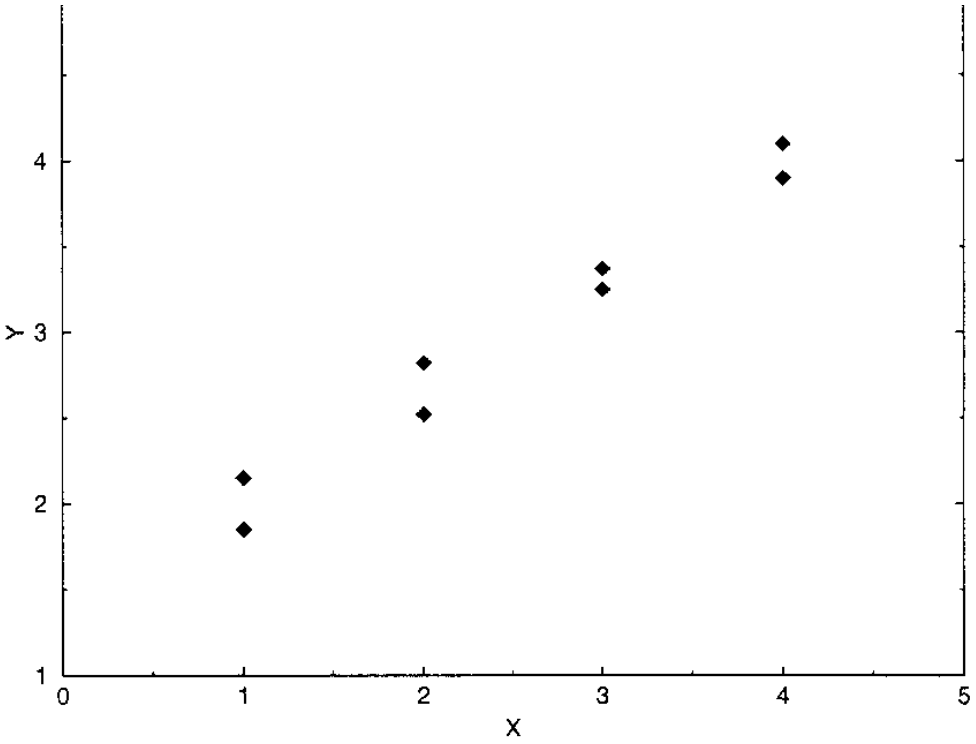

Abb. 20a

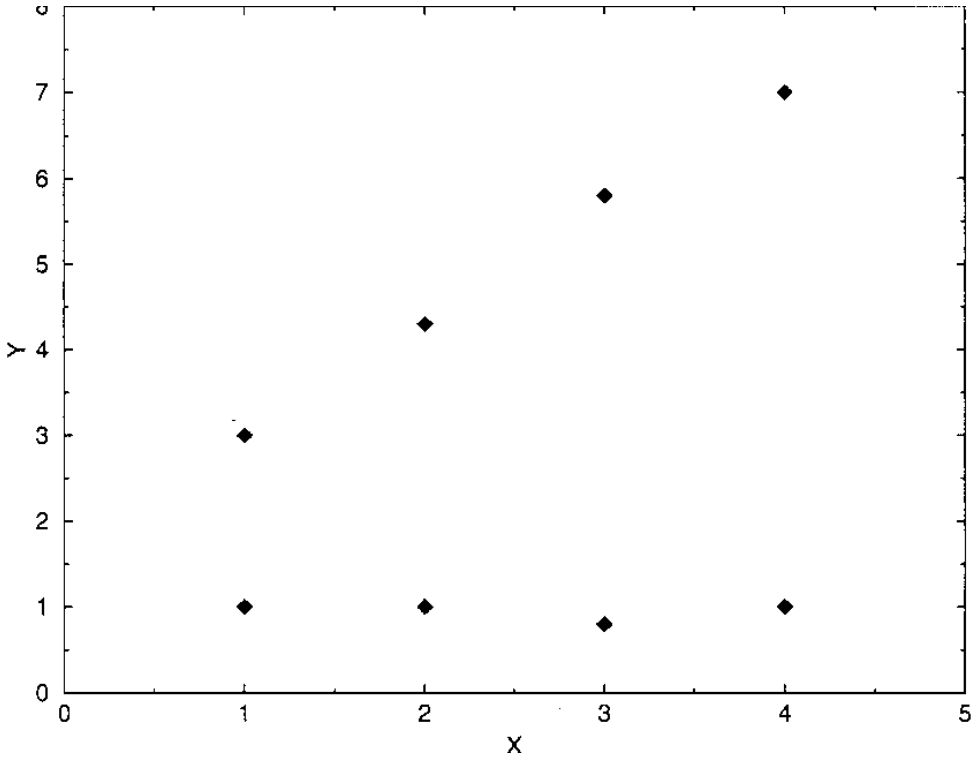

Abb. 20b

Das in der Regressionsanalyse allgemein eher wenig verwendete Verfahren der Mittelwertbildung ist trotz des geschilderten Defizits unter dem Gesichtspunkt der Überprüfung von stochastischen Gesetzen berechtigt. Schließlich macht das lexikalische Basismodell Vorhersagen über funktionale Zusammenhänge zwischen unabhängigen Variablen und den *Mittelwerten* der abhängigen Variablen, *nicht* den Einzelwerten. Für das Modell ist es nicht von Bedeutung, ob ein Mittelwert *M* durch zwei stark voneinander abweichende Werte zustande kommt, die gemittelt den Wert *M* ergeben, oder durch zwei Werte, die beide gleich *M* sind.

Bei der Mittelwertbildung werden jedoch Informationen außer acht gelassen, die in den Rohdaten vorhanden sind, nämlich die Varianz innerhalb einer Klasse der unabhängigen Variablen sowie (bei Nicht-Gewichtung der Datenpaare) die Anzahl der Rohdatenpunkte, die sich in einer Klasse aufhalten. Möglicherweise könnten sich durch Ausnutzung dieser Informationen Ansatzpunkte für eine weitere Erhöhung der Vorhersagekraft und Adäquatheit des lexikalischen Basismodells ergeben. Zur Veranschaulichung soll an dieser Stelle ein kleines Beispiel gebracht werden:

Rohdaten liefern optisch ein ganz anderes Bild als die zugehörigen Mittelwerte. Typisch für Darstellungen von Rohdaten sind nämlich nicht um eine Regressionskurve oder -gerade streuende Datenpunkte, sondern vielmehr sich unter einer fiktiven, oft nicht-monotonen 'Begrenzungslinie' ansammelnde Punkte.

Der Informationsgewinn, den die Rohdaten gegenüber Mittelwerten bieten, soll anhand der Abhängigkeit der Polylexie von der graphemischen Länge demonstriert werden.

Abbildung 21 zeigt die verrauschten Rohdaten[8] dieser Abhängigkeit. Aufgrund der Verrauschung ist neben der Streuung der Datenpunkte auch die Dichte der Besetzung der diskreten Datenpunkte zu erkennen, die im unteren Bereich sehr viel höher ist als im oberen Bereich.

Abbildung 22 zeigt die gleichen Daten, die aber nicht verrauscht sind. Durch die diskrete Darstellung geht (optisch) schon die Information über die Dichte der Besetzung verloren.

Abbildung 23 schließlich zeigt noch einmal die aus den Rohdaten gebildeten Mittelwerte, aus denen auch nicht mehr die Streuung zu ersehen ist, die den Mittelwerten zugrundeliegt.

---

8 Die Verrauschung der Daten wird dadurch erreicht, dass für jeden Datenpunkt (*X*; *Y*) jeweils ein Zufallswert aus den Intervallen [X+1] und [Y+1] generiert und – aufgefaßt als neue Koordinaten – graphisch dargestellt wird.

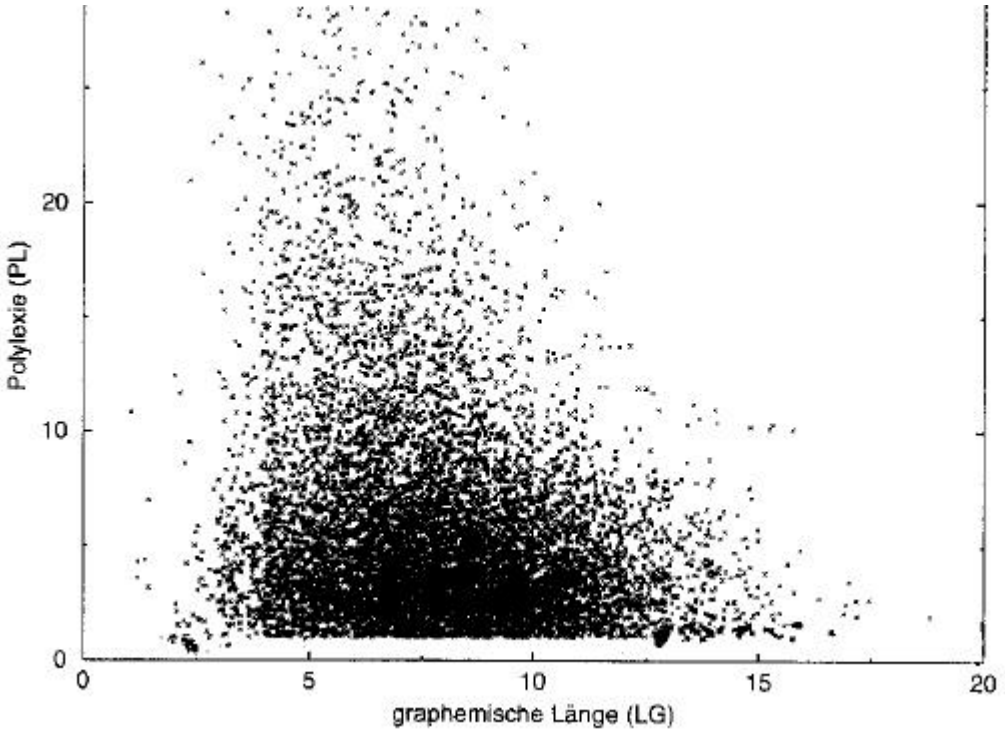

Abb. 21: Die Abhängigkeit der Polylexie von der graph. Länge (verrauschte Rohdaten; Ausschnitt)

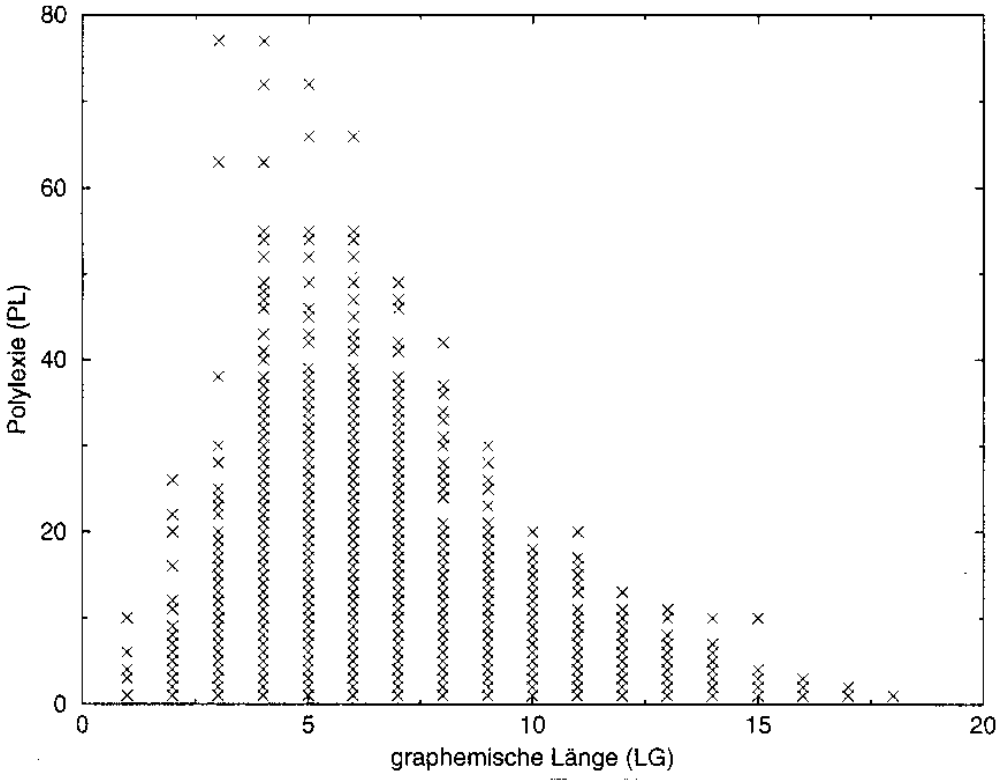

Abb. 22: Die Abhängigkeit der Polylexie von der graph. Länge (diskr. Rohdaten)

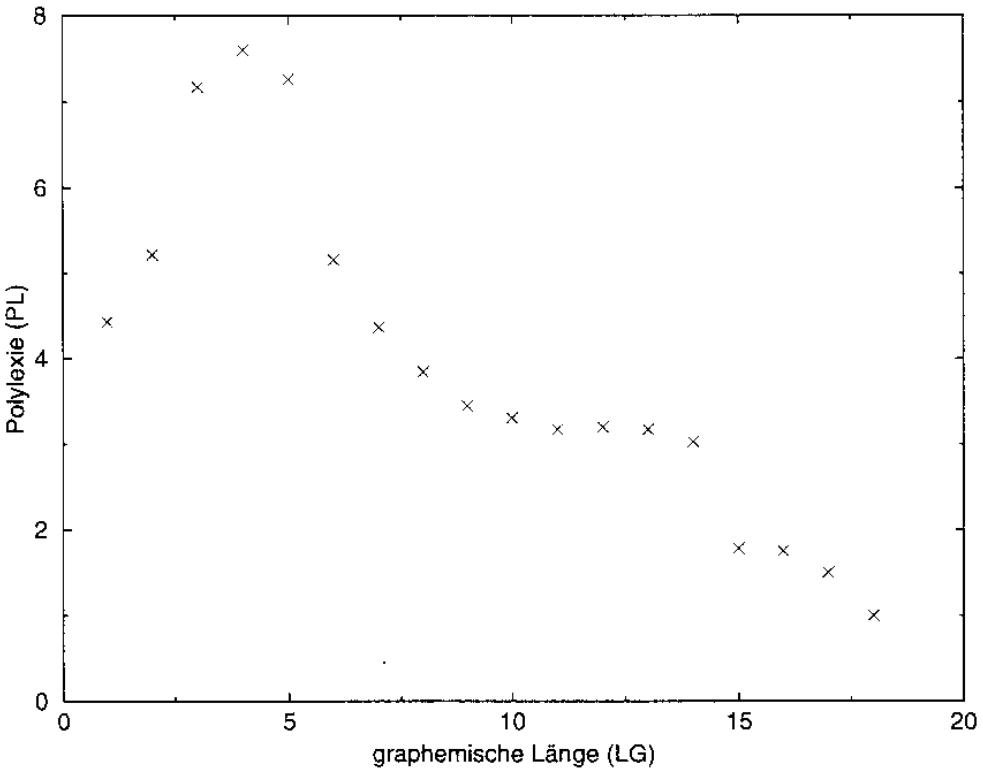

Abb. 23: Die Abhängigkeit der Polylexie von der graph. Länge (Mittelwerte)

4. Zusammenfassung

Neben den Ergebnissen der empirischen Untersuchungen (2.9) wurden in dieser Arbeit auch einige Fragestellungen aufgeworfen, die im Rahmen dieser Arbeit nicht abschließend geklärt werden konnten und die vor weiteren empirischen Untersuchungen behandelt werden bzw. diese anregen sollten. So wäre zum Beispiel zu überdenken,

- ob und wann die Verwendung gewichteter Mittelwerte gerechtfertigt ist,
- wie die in den Rohdaten enthaltenen Informationen in die linguistische Modellbildung eingebracht werden könnten,
- welche statistischen Testverfahren für die Überprüfung der Hypothesen aus linearen, aber auch aus nicht-linearen Modellen geeignet bzw. die aussagekräftigsten sind.

Nur wenn solche Fragen geklärt sind, können wirklich zuverlässige Aussagen über die Abweichungen der empirischen Werte von den Modellvoraussagen gemacht werden.

Die mangelnde Übereinstimmung zwischen den theoretisch vorhergesagten und den empirischen Parametern der indirekten Abhängigkeiten im lexikalischen

Basismodell zeigt die Notwendigkeit einer Verfeinerung bzw. Modifikation des Modells auf. Es wurde darauf hingewiesen, dass die theoretische Motivation solcher Modifikationen unumgänglich ist. Solange für Alternativmodelle keine linguistischen Begründungen vorliegen sollte weiterhin das existierende Basismodell als Untersuchungsrahmen verwendet werden.

Literatur

Altmann, G. (1992). Das Problem der Datenhomogenität. In B. Rieger (Hg.), *Glottometrika 13* (S. 287-298), Bochum: Brockmeyer.

Edwards, A.L. (1976). *An Introduction to Linear Regression and Correlation*. San Francisco: Freeman.

Gieseking, K. (1993). *Synergetische Aspekte von Struktur und Dynamik der englischen Lexik.* Unveröffentlichte Magisterarbeit, Universität Trier.

Köhler, R. (1986) *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. (= Quantitative Linguistics; 31). Bochum: Brockmeyer.

Köhler, R. (1990) Elemente der synergetischen Linguistik. In R. Hammerl (Hg.), *Glottometrika 12* (S. 179-187), Bochum: Brockmeyer.