Quantitative Untersuchungen im Französischen: Häufigkeitsverteilungen und funktionale Zusammenhänge

Inaugural-Dissertation

zur Erlangung der Doktorwürde der Philosophie

der Universität Trier

Dem Fachbereich II: Sprach-, Literatur- und Medienwissenschaften

vorgelegt von:

Jacob Kamta

Erstgutachter: Prof. Dr. Reinhard Köhler

Zweitgutachter: Univ. Prof. Mag. Dr. Peter Grzybek

Tag der mündlichen Prüfung: 01.Juli 2009

Ngouanfouo Yvette

Tamossoc Kamta Fred Loic

Tsondjou Lélys Babelle

Vorwort

Diese Arbeit wurde im Sommersemester 2009 als Dissertation dem Fachbereich II der Universität Trier vorgelegt und für die Publikation nicht mehr verändert.

Es ist kaum möglich, all jene zu nennen, die mir beim Verfassen dieser Studie geholfen haben. Stellvertretend und ohne Anspruch auf Vollständigkeit möchte ich mich jedoch bei einigen Personen besonders bedanken: Mein Dank gilt zuallererst Herrn Prof. Dr. Reinhard Köhler nicht nur für die unermüdliche Betreuung der Arbeit und seine zuständige Gesprächsbereitschaft, sondern auch für die Literatur, die er mir zur Verfügung stellte. Mein herzlicher Dank gilt ebenfalls Herrn Univ. Prof. Mag. Dr. Peter Grzybek für die wertvollen Verbesserungsvorschläge und die zugige Erstellung des Zweitgutachtens. Das Korrekturlesen übernahmen Lisa Gutbroad und Anna Matschke. Meine Familie leistete moralischen Beistand.

Trier, im August 2009 Jacob Kamta

Inhaltsverzeichnis

1. Einleitung

1.1. Zum Gegenstand der Arbeit

Auf verschiedenen Ebenen der Sprache lassen sich Einheiten unterscheiden, bei denen man unterschiedliche Eigenschaften beobachten und quantifizieren kann. Quantitative Eigenschaften sind „zur Beschreibung und zum tieferen Verständnis der Entwicklung und der Funktionsweise von Sprachsystemen und ihren Bestandteilen nötig“ (KÖHLER 2005a, 1). Die vorliegende Arbeit befasst sich daher mit quantitativen Eigenschaften sprachlicher Einheiten. Untersucht werden dabei ihre (Häufigkeits-)Verteilungen sowie die Wechselbeziehungen zwischen ihnen anhand von Daten aus der französischen Sprache. Die Arbeit beschränkt sich auf einige Probleme aus unterschiedlichen Bereichen, die sich mit Softwareprogrammen lösen, und widmet sich nur der Überprüfung bestimmter zu Gesetzen gewordener Hypothesen.

Im der Arbeit wird mit den sprachlichen Entitäten *Wortbildungsaffix*, *Wort*, *Satz* bzw. *Teilsatz*, *Buchstabe* und *Phonem* und den quantitativen Eigenschaften *Länge* bzw. *Komplexität*, *Frequenz* und *Polylexie* operiert.

Die genannten Einheiten spielen eine herausragende Rolle im Sprachsystem. *Wort* und *Satz* bzw. *Teilsatz* sind neben *Buchstaben* und *Lauten* für jeden Sprachbenutzer die geläufigsten sprachlichen Entitäten. *Wort* und *Satz* sind darüber hinaus zentrale Einheiten der Sprachstruktur, d.h. jeglicher Textkonstitution. *Wortbildungsaffixe* ermöglichen es, aus bereits bestehenden Wörtern neue abzuleiten. Durch die Ableitung wird die Lexik der Sprache erweitert. Diese Erweiterung ist notwendig, um die Lexik der sich ständig ändernden Systemumwelt und den sich ständig vermehrenden Kommunikationssituationen anzupassen.

Länge bzw. *Komplexität*, *Polylexie* und *Frequenz* gehören zu den fundamentalen Eigenschaften sprachlicher Einheiten. Über die Größe *Länge* schreibt (TULDAVA 1998, 98), dass sie „ein quantitativ-typologisches Kriterium ist, das nicht nur strukturelle Merkmale der Sprache, sondern auch individuelle und funktionale Besonderheiten von Texten und Wörterbüchern prognostiziert“. Die Eigenschaft *Häufigkeit* spielt in der quantitativen Linguistik eine zentrale und herausragende Rolle. Im Gegensatz zu der Länge eines Ausdrucks ist eine Zurückführung der Häufigkeiten auf Wahrscheinlichkeiten möglich: Je wahrscheinlicher eine Einheit, desto häufiger kommt sie vor.

Die in dieser Arbeit dargestellten Untersuchungen zu den Verteilungen der weiter oben genannten Spracheigenschaften und -einheiten behandeln die semantische Diversifikation von Wortbildungsaffixen, die Ranghäufigkeitsverteilungen von Wörtern, Buchstaben sowie Pho-

nemen, die Satzlängenverteilung, das Textblock-Gesetz und die Verteilung der Komplexität und Frequenz von Teilsätzen.

Die synergetische Linguistik betrachtet die Sprache als selbstorganisierendes und selbstregulierendes System. Dabei werden nicht nur Einheiten und Beziehungen zwischen diesen Einheiten erfasst, sondern auch die Eigenschaften des Systems aufgrund ihrer Funktionen innerhalb des Systems erklärt.

Zahlreiche Zusammenhänge auf verschiedenen Ebenen der Sprache wurden bereits nachgewiesen. KÖHLER (1986) hat ein synergetisches Modell der Lexik für das Deutsche entwickelt, das Anspruch erhebt, für alle natürlichen Sprachen gültig zu sein. Die Struktur dieses Modells wird durch Systembedürfnisse, quantitative Eigenschaften und ihre Relation zueinander bestimmt. Die vorliegende Arbeit beschreibt die Untersuchung einiger Zusammenhänge zwischen sprachlichen Eigenschaften, die von diesem Modell vorhergesagt werden. Es handelt sich dabei um die Abhängigkeit der Polylexie von der Länge, der Länge von der Frequenz und der Polylexie von der Frequenz. Neben diesen Zusammenhängen werden in dieser Arbeit auch das Menzerath-Gesetz sowie eine vermutete Beziehung zwischen Frequenz und Komplexität von Teilsätzen, die aus der Koppelung von Frequenz und Länge von Lexemen abgeleitet wird, behandelt. Das Menzerath-Gesetz ist eine der wichtigsten Errungenschaften der quantitativen Linguistik. Es strukturiert die Sprache von den größten Einheiten bis hinunter zu den Lauten bzw. Phonemen. Untersucht werden soll der Zusammenhang von Satzlänge zu Teilsatzlänge.

Bei der Behandlung der einzelnen Häufigkeitsverteilungen und Zusammenhänge gilt es, die Hypothese zu überprüfen, dass sprachliche Erscheinungen immer theoretisch begründbaren Gesetzen unterliegen (ALTMANN 1985,7). Zu allen genannten Fragestellungen soll entsprechend jeweils ein Gesetzesvorschlag, der bereits in der Literatur für solche Fälle geprüft wurde, aufgegriffen und erneut überprüft werden[1].

1.2. Zur Zielsetzung der Arbeit

Eine wesentliche Aufgabe der quantitativen Linguistik besteht in der Erforschung von Gesetzmäßigkeiten, die sprachliche Phänomene steuern. Die Notwendigkeit, Sprache mathematisch zu erforschen, betont u.a. KÖHLER (2005a). Mathematische Mittel ermöglichen es, das höchste Ziel jeder Wissenschaft zu realisieren, nämlich „die Erklärung der Phänomene (und damit auch die Möglichkeit zu ihrer Vorhersage)" (KÖHLER 2005a, 12). Erklärungen gelten

[1] Vgl. Kap. 3.

dann als gelungen, wenn nachgewiesen werden kann, dass die beobachteten Zustände bestimmten Gesetzmäßigkeiten folgen.

Die quantitative Linguistik hat eine Vielzahl von Gesetzeshypothesen entdeckt. Eine der Voraussetzungen dafür, dass eine Gesetzeshypothese vorläufig als gültig angesehen werden kann, ist deren Überprüfung durch möglichst vielfältige Daten verschiedener Sprachen. Die vorliegende Arbeit ist in diesem Zusammenhang zu sehen; sie soll einen Beitrag dazu leisten, dass sprachstatistische Erhebungen der Differenzierung linguistischer Befunde oder Hypothesen dienen.

Mit dieser Arbeit soll auch dazu beigetragen werden, eine Lücke zu schließen: Beim Studium der einschlägigen Literatur zu den quantitativen Untersuchungen zum Französischen fällt auf, dass die Zahl der Überprüfungen der bisher von der quantitativen Linguistik entdeckten Sprachgesetze im Vergleich zu Sprachen wie Deutsch, Englisch, Russisch, Slowakisch, etc. sehr gering ist[2]. Soweit Daten vorliegen, sind sie entweder (zum Teil) sehr alt oder im Grunde genommen unbrauchbar, denn es fehlen die Kriterien, die Auskunft darüber geben, wie gut die Übereinstimmung zwischen Beobachtung und Theorie ist und wie die Werte der Parameter sich darstellen. Die vorliegende Arbeit soll nun dazu dienen, die Beobachtungen und Tests zu den quantitativen Sprachgesetzen in dieser Sprache zu ergänzen und damit die genannten Defizite ein wenig zu verringern.

1. 3. Begriffliche Erklärungen

Zur Durchführung der Untersuchung ist es nötig festzulegen, wie die untersuchten Einheiten identifiziert und deren Eigenschaften gemessen werden sollen.

1.3.1. Sprachliche Einheiten

a) Wort, Phonem und Buchstabe

Die Einheit *Wort* wird hier auf orthographischer Ebene definiert und als jede Buchstabenfolge verstanden, die durch Leerräume und/oder Interpunktionszeichen begrenzt ist:

> „Là où le linguiste part d'un texte écrit, il peut accepter si cela lui convient, pour définir une unité 'mot', les critères de l'orthographie: un mot dans l'écriture est un segment séparé des autres segments par des espaces blancs." (MARTINET, A. 1969, 30)

Alternative Wortdefinitionen und deren Auswirkungen auf quantitative Untersuchungen finden sich u.a. in ANTIĆ et al. (2006).

[2] Vgl. Kap. 2

Mit Bindestrich verbundene Wörter stellen im Französischen ein Problem dar: Sie lassen sich nicht immer als jeweils ein Wort auffassen. Eine Form wie "as-tu" oder "allez-vous" kann umgestellt und daher ebenso als Folge von zwei Wörtern verstanden werden. Dagegen ist eine Form wie "amour-propre" oder "va-et-vient" lexikalisiert; sie gilt als Kompositum und wird daher jeweils als nur ein Wort behandelt.

Ein Buchstabe bezeichnet ein Kommunikations-Aufschreibzeichen eines Sprachsystems, ein Phonem die kleinste bedeutungsunterscheidende Einheit.

b) Satz und Teilsatz

In Anlehnung an WINTER (1974) und PIEPER (1979) wird in dieser Arbeit unter einem Satz eine endliche Sequenz von Graphemen und Zwischenräumen verstanden, die zwischen zwei Satzende-Zeichen bzw. zwischen dem Textanfang und dem ersten Satzzeichen steht, und durch Großschreibung des ersten Graphems am Satzanfang gekennzeichnet ist. Als Satzende-Zeichen gelten: der Punkt (**.**), das Fragezeichen (**?**) und das Ausrufungszeichen (**!**). Steht ein Doppelpunkt (**:**) und ist das erste Graphem des nächsten Wortes groß geschrieben, so gilt der Satz als abgeschlossen.

Ein Teilsatz wird als der Teil des Satzes definiert, der ein finites Verb enthält, wobei finite Verben durch ihre Konjugationsmorpheme bestimmt sind.

c) Wortbildungsaffix

Wortbildungsaffixe bzw. -morpheme dienen dazu, neue Wörter zu bilden. Sie ergeben in Verbindung mit einem Wortstamm bzw. Grundmorphem neue Wörter.

1.3.2. Sprachliche Eigenschaften

a) Polylexie

Für die Mehrbedeutung von Lexemen werden in der Sprachwissenschaft die Begriffe Polysemie und Homonymie verwendet. Die Zuordnung von Mehrdeutigkeit zu einer der beiden Kategorien ist aber umstritten; es herrscht keine Einigkeit über eine klare Differenzierung der beiden Termini. Dazu SCHIERHOLZ (1991, 65):

„Wenn die Termini ‚Polysemie' und ‚Homonymie' tatsächlich etwas Verschiedenes bezeichnen, so sollte es möglich sein, diese Unterschiede definitorisch zu verankern. Unbefriedigend sind Differenzierungsversuche, in denen Äußerungen wie ‚subjektive Faktoren sind dabei wohl nicht ganz auszuschließen, (…)' als Lösungshilfe akzeptiert werden."

In der Quantitativen Linguistik wird statt Polysemie der Begriff Polylexie als Maß für die Mehrdeutigkeit von Lexemen verwendet[3]. KÖHLER (1986, 57) definiert Polylexie als „die Anzahl der verschiedenen Bedeutungen, die eine lexikalische Einheit zu einem gegebenen Zeitpunkt trägt.“ Als Schätzung für diesen Wert schlägt er „die Zahl der Wörterbucheinträge zu dieser Einheit“ (S. 92) vor.

Abhängig von dem benutzten Wörterbuch können die Untersuchungsergebnisse jedoch unterschiedlich ausfallen. SCHIERHOLZ (1991, 178ff.) schlägt deshalb die Verwendung verschiedener Wörterbücher für die Zählung vor. Problematisch sei aber, so STEINER (2002, 213), „bei einem solchen Verfahren (…) die Frage, wie die Polylexiewerte aus den verschiedenen Wörterbüchern miteinander verrechnet werden sollen.“

Analog zu Köhlers Definition wird in der vorliegenden Arbeit die Polylexie eines Wortes als die Anzahl der Eintragungen unter dessen Stichwort im Wörterbuch bestimmt. Wie bei HAMMERL (1991, 48) werden - bei der Messung - Unterbedeutungen, d.h. Bedeutungen mit Ordnungsnummern wie 1.1, 1.2, etc. außer Acht gelassen.

Die *Polylexie* spricht für die Wahl des Lemmas als Beobachtungsgröße von Wörtern, da bei der Ermittlung der Bedeutung eines Wortes unwichtig ist, in welcher Form dieses Wort in der Rede beobachtbar ist. Bedeutungen sind, so KÖHLER (1986, 89), „Entitäten, die durch Abstraktion von der Aktualisierung der Wörter in einem grammatischen Gefüge zustande kommen.“ Zu einem Lemma werden hier alle Formen eines Nomens, Verbs, Adjektivs und Artikels gerechnet.

Die *Morph-Polylexie*, d.h. die Bedeutungen von Wortbildungsaffixen, wird anhand der Wortbildungsliteratur bestimmt: Zu jedem Wortbildungsaffix werden aus der Literatur zur Wortbildung im Französischen eine oder mehrere Funktionen bzw. Bedeutungen entnommen. Die Anzahl dieser Funktionen wird als Maß für die Morph-Polylexie genommen.

b) Länge

Die Eigenschaft *Länge* bezeichnet die Anzahl der konstituierenden Elemente eines sprachlichen Ausdrucks. Sie wird für die Einheit *Wort* auf zweierlei Weise gemessen, d.h. sowohl in der Buchstabenanzahl als auch in der Anzahl der Silben in einem Wort. Für beide Messungen wird die Lemmalänge, nicht die Wortformlänge berücksichtigt.

Die Buchstabenanzahl wird dadurch ermittelt, dass lediglich die Buchstaben pro Wort gezählt werden. Die Silbenanzahl pro Wort wird nach der Anzahl der im Wort vorkommenden Vokale und Diphthonge bestimmt. D.h.: Vokale und Diphthonge werden als Hauptträger der Sil-

[3] Vgl. z.B. KÖHLER, R. (1986)

ben, d.h. als Silbengipfel aufgefasst. Geschriebene Vokale, die nicht gesprochen werden, werden nicht als Silbenträger betrachtet.

Für die Bezugseinheit *Satz* wird die Länge durch die Anzahl der Teilsätze pro Satz gemessen. Es wäre selbstverständlich auch möglich, die Satzlänge entsprechend der Anzahl anderer Einheiten wie Silben, Morpheme, Phoneme oder sogar Buchstaben zu messen.
Die (durchschnittliche) Teilsatzlänge wird als Quotient aus der Anzahl der Wörter pro Satz und finiten Verben ermittelt.

c) Frequenz

Die Frequenz gibt die Vorkommenshäufigkeit einer linguistischen Einheit an. Sie kann als absolute oder relative Anzahl angegeben werden.
In der vorliegenden Untersuchung werden nur absolute Häufigkeiten verwendet. Die Häufigkeit der Einheit *Wort* wird im Text sowie in einem Häufigkeitswörterbuch ermittelt. Gezählt werden dabei die Wortformen und nicht die Lemmata. Eine Wortform ist hier die konkret realisierte Form eines Wortes, d.h. das, was man an der Oberfläche beobachten kann. Dies ist nicht nur im Text möglich; auch das von uns verwendete Häufigkeitswörterbuch bietet diese Möglichkeit an.

Für die Einheit *Teilsatz* wird die Frequenz definiert als die Anzahl der Vorkommen eines Teilsatzes mit einer bestimmten Komplexität.

d) Komplexität

Die Komplexität eines Teilsatzes wird in der Anzahl der direkt vom Verbknoten abhängigen Knoten gemessen.

1.4. Untersuchungsmaterial

In dieser Arbeit werden Ergebnisse vorgestellt, die durch die Auswertung von Wörterbüchern und Texten erzielt wurden. Die Daten wurden entsprechend den festgelegten Operationalisierungs- und Messvorschriften erhoben.

1.4.1. Wörterbücher

Für die Bestimmung der Polylexie eines Stichwortes wurde das Wörterbuch *„Le Petit Larousse illustré 2007“* verwendet. Neben der Anzahl der Bedeutungen pro Wort kann in diesem Wörterbuch auch die Wortlänge − zumindest für einige Wörter − in Silben gemessen

werden. Die Wörter, deren Polylexie zu ermitteln waren, stammen aus einer Textbearbeitung, einem Textkorpus und aus dem Wörterbuch selbst[4].

Zur Bestimmung der Frequenz im Lexikon wurde das Frequenzwörterbuch von A. JUILLAND et al. (1970) verwendet. Die Wörter, deren Frequenz abgelesen wurde, sind die gleichen Wörter wie die aus *„Le Petit Larousse illustré 2007“*.

Die Struktur der 5.082 Einträge des Frequenzwörterbuches zeigt den nachstehenden Auszug:

accident n		**20.00**		**32**		**62.50**	
	32	7	3	6	15	1	
accident	23	6	3	4	10	0	
accidents	9	1	0	2	5	1	

Wichtig sind hier die Elemente am Anfang jeder Zeile. Die erste Zeile gibt das Lemma, *accident*, und dessen Wortartzugehörigkeit, hier Nomen, an. In der zweiten Zeile findet sich die Summe der Häufigkeit aller Wortformen. Die dritte und vierte Zeile enthalten die Häufigkeit der einzelnen Wortformen.

1.4.2. Texte

Bei der empirischen Überprüfung der verschiedenen Häufigkeitsverteilungen und funktionalen Zusammenhänge wurden Texte mit verschiedenen Eigenschaften verwendet:

Zum einen wurden vollständige Texte ausgewertet, wobei unter vollständigen Texten sowohl in sich abgeschlossene Kapitel eines Textes als auch abgeschlossene Texte verstanden werden. Untersucht wurden literarische Texte sowie Zeitungsartikel. Letztere entstammen den Online-Ausgaben der französischen Tageszeitung „Le Monde“ und sind im gleichen Jahr, 2006, entstanden. Bei den literarischen Texten handelt es sich in erster Linie um Romane. Sie sind unter *http://www.lexique.org/public/corpatext.php* abrufbar[5].

Zum anderen wurden Textausschnitte und Textmischungen verwendet. Textausschnitte stellen Textpassagen dar, die zufällig und willkürlich ausgewählt wurden, wie z.B. Gedichte einer Gedichtsammlung; bei Textmischungen geht es um die Zusammenstellung von willkürlich ausgewählten Texten[6].

Bei der Auswertung der Texte wurden Titel, Kapitelüberschriften und alles, was nicht zum eigentlichen Text gehört, wie z.B. Fußnoten und Autorennamen, nicht berücksichtigt. Eine Übersicht über die verwendeten Texte stellen Tabellen 1 und 2 im Anhang dar. Auf die

[4] Vgl. Abschnitt 6.1.
[5] Stand 2007
[6] Siehe dazu Texte 19 und 20 in Tabelle 2 im Anhang

Nummer des Textes, auf die in den einzelnen Analysen Bezug genommen wird, folgen dabei die Autorennamen bzw. die Textquelle sowie die Bezeichnung des Textes.
Die Textwahl für die einzelnen Analysen erfolgte relativ willkürlich.

Dargestellt werden in der vorliegenden Arbeit zunächst Vorarbeiten zu den zu behandelnden Häufigkeitsverteilungen und Zusammenhängen zwischen sprachlichen Eigenschaften sowie quantitative Untersuchungen, die in der französischen Sprachwissenschaft durchgeführt wurden. Damit befasst sich Kapitel 2. In dem darauf folgenden Kapitel werden die theoretischen Modelle dargestellt, die bei der Modellierung der einzelnen Fragestellungen verwendet werden. Die Anpassungsgüte dieser Modelle an die empirischen Daten wird mit Hilfe statistischer Methoden beurteilt. Mit den verwendeten Testmethoden befasst sich Kapitel 4.
Im Kapitel 5 werden die einzelnen Häufigkeitsverteilungen empirisch überprüft. Behandelt wird zunächst die semantische Diversifikation der Bedeutungen von Wortbildungsaffixen. Dies ist Thema des Abschnitts 5.1. Die Ranghäufigkeitsverteilung von Wörtern ist Gegenstand des nächsten Abschnitts. Darauf folgen die Rangverteilung von Buchstaben und Phonemen. Die Gesetzmäßigkeiten, welchen die Verteilung der Satzlänge und die Häufigkeitsverteilung in Textblöcken folgen, werden jeweils in Abschnitten 5.4 und 5.5 überprüft. Das Kapitel schließt mit der Überprüfung der Verteilung der Komplexität und Frequenz von Teilsätzen.

Im Anschluss an die Häufigkeitsverteilungen werden die einzelnen Zusammenhänge zwischen sprachlichen Eigenschaften überprüft. Behandelt werden zunächst die KÖHLER (1986) entnommenen Zusammenhänge von sprachlichen Eigenschaften. Damit befasst sich Abschnitt 6.1. Im Anschluss daran wird auf den Zusammenhang zwischen Satz- und Teilsatzlänge eingegangen. Der Zusammenhang zwischen Frequenz und Komplexität von Teilsätzen ist Gegenstand des Abschnitts 6.3.

2. Quantitative Untersuchungen in der französischen Sprachwissenschaft und Vorarbeiten zum Gegenstand der Arbeit

2.1. Quantitative Untersuchungen in der französischen Sprachwissenschaft

In Form eines kurzen Überblicks soll im Folgenden auf quantitative Untersuchungen eingegangen werden, die in der französischen Sprachwissenschaft durchgeführt wurden.

In der französischen Forschungsliteratur finden sich Arbeiten, die statistische Erhebungen durchführen. Diese Untersuchungen behandeln jedoch im Gegensatz zu quantitativen Arbeiten zum Deutschen, Englischen oder Russischen nur einige Themengebieten. Die meisten davon sind lexikostatistischer Art. Von wissenschaftshistorischer Bedeutung in diesem Bereich sind die Arbeiten von GUIRAUD (1960, 1963, 1968 …) und MULLER (1968, 1973, 1977a, 1977b…), die im Hinblick auf die Anwendung quantitativer Verfahren auf sprachliches Material einen wichtigen Beitrag geleistet haben. Nach MULLER (1968, 6) stellt GUIRAUD (1960) in dessen „problèmes et méthodes de la statistique linguistique“ vor:

> „ses recherches sur le rapport qui pourrait exister entre les emprunts aux langues étrangères et le système phonologique de la langue emprunteuse; proposition hardie; puisque sa démonstration tend à prouver une recherche d’équilibre phonologique par des apports lexicaux, à introduire la notion de fréquence dans la description de ce système et à déceler une finalité.”

Mit seiner Publikation "Initiation à la statistique linguistique" vermittelt MULLER (1968) Sprachwissenschaftlern ohne mathematische Ausbildung nützliche und praktische Kenntnisse aus dem Bereich der quantitativen Linguistik. Es geht dabei,

> „de présenter les principes et non les résultats de la statistique linguistique, de décrire ses méthodes d’exploration et non ses découvertes ou ses conquêtes” et „de familiariser le linguistique avec le raisonnement statistique, de l’habituer ou de le réhabituer au langage algébrique qui sert de support à ce raisonnement.” (MULLER 1968, S. 5-6)

In MULLER (1979) findet sich eine Sammlung von Aufsätzen zur Lexikostatistik in der französischen Tradition. Weitere wichtige Arbeiten auf diesem Gebiet sind GUIRAUD (1954), BRUNET (1978) und MENARD (1983).

Die meisten lexikostatistischen Arbeiten sind dem Wortschatz gewidmet. Man findet u.a Untersuchungen zur Erarbeitung des Grundwortschatzes, z. B. in DOTTRENS & MASSARENTI (1948) und HAYGOOD (1937); zum Wachstum des Wortschatzes in Texten, z.B. in BERNET (1988), THOIRON (1986), DUGAST (1978) und BRUNET (1988); zum Mutter- und Fremdsprachenerwerb sowie zur Entwicklung von Wortschatzbeherrschung und Redegeschwindigkeit, z.B. in DUNN-LARDEAU (1986), KUCHNER (1932) und MICHÉA (1949).

Die bekanntesten lexikostatistischen Untersuchungen sind wohl Häufigkeitswörterbücher, in denen der Wortschatz nach seiner Frequenz geordnet aufgeführt wird. Hier lassen sich sowohl „einfache“ Frequenzwörterbücher, wie zum Beispiel JUILLAND et al. (1970), EATON (1961) und IMBS (1971) als auch Fach-Häufigkeitswörterbücher finden, wie z.B. HOFFMANN (1976). Ein statistisches Handbuch des Französischen scheint es nicht zu geben.

Die quantitative Stilanalyse sowie die statistische Erforschung poetischer Sprache im Allgemeinen behandeln u.a. DUGAST (1979a, 1979b), FIALA (1986), GRAMMONT (1937) und MONSONÉGO (1966). Die Untersuchungen zu den Gebieten der Phonetik, Phonologie und Morphologie gelten der Laut- und Phonemhäufigkeit sowie der Phonem- und Morphemdistribution.

Im Bereich der Buchstaben lassen sich ebenfalls Untersuchungen finden: Mit der Rang-Frequenz-Verteilung von Buchstaben hat sich NASVYTIS (1953) beschäftigt. Die Daten sind allerdings nicht nur veraltet, sondern auch gering. Im Internet finden sich verschiedene Analyzer für französische Buchstabenhäufigkeiten, die alle unterschiedliche Inventare verwenden. Es lassen sich auch Untersuchungen zur Wortlängenhäufigkeit und deren Verteilung finden. Diese beziehen sich auf Pressetexte und Briefe, und werden in DIECKMANN & JUDT (1996), FELDT, JANSSEN & KULEISA (1997) und HEINICKE (2008) behandelt.

2.2. Quantitative Vorarbeiten zum Gegenstand der Arbeit

Im Folgenden soll auf quantitative Vorarbeiten zu den in der vorliegenden Untersuchung zu behandelnden Häufigkeitsverteilungen und funktionalen Zusammenhängen eingegangen werden. Berücksichtigt werden dabei Untersuchungen, die sich nicht nur auf das Französische beziehen, sondern auch auf andere Sprachen.

2.2.1. Häufigkeitsverteilungen

2.2.1.1. Häufigkeitsverteilung der Bedeutungen von Wortbildungsaffixen

Wie bereits erwähnt, besteht ein wesentliches Forschungsziel der Quantitativen Linguistik in der Untersuchung von Gesetzmäßigkeiten, die sprachliche Phänomene steuern. Zu diesen Sprachgesetzen zählt das Diversifikationsgesetz, das erstmals in ALTMANN (1985) und in ALTMANN (1991) erneut vorgestellt wurde. Dieses Gesetz besagt, dass sich Diversifikationsprozesse, d.h. formale bzw. funktional-semantische Differenzierungen sprachlicher Einheiten als gesetzmäßig erweisen. Ein Beispiel für eine funktional-semantische Diversifikation ist das Auftreten der Bedeutungen von Wortbildungsaffixen. Bei diesen Entitäten verläuft die semantische Diversifikation dadurch, dass die Verwendung eines Präfixes oder Suffixes z.B. die

Bedeutung des Grundwortes verändert bzw. modifiziert. Dies zeigt beispielsweise das Wortpaar *Enfant – Enfantin.*

Eine Erklärung der Diversifikation sprachlicher Einheiten findet sich in dem Zipf'schen Prinzip der geringsten Anstrengung, nach dem Sprecher- und Hörerbedürfnisse gegeneinander konkurrieren: Der Sprecher spart am effektivsten an Anstrengung, wenn er möglichst wenig Entitäten mit jeweils möglichst vielen Bedeutungen benutzen kann. Der Hörer rezipiert auf der anderen Seite am ökonomischsten, wenn jeder Entität nur eine einzige Bedeutung zugeordnet wird.

Auf Grund der Minimierung des Kodierungs- und Dekodierungsaufwands kommt es beispielsweise dazu,

> „daß die häufigsten Wörter (durchschnittlich) die kürzesten sind, daß diese wiederum die meisten Bedeutungen haben und häufiger in unterschiedlichen Sprachschichten und -stils vertreten sind bzw. in Texten unterschiedlicher Gattung (Polytexie) als längere mit wenig Bedeutungen." (ROTHE 1989, 122)

Untersuchungen zur semantischen Diversifikation von Affixen wurden bereits in etlichen Sprachen gemacht. BEÖTHY & ALTMANN (1991)[7] konnten zeigen, dass die Bedeutungen von ungarischen Präfixen im Text dem Prinzip der Diversifikation unterliegen. Auch die Rang-Häufigkeitsverteilungen der Übersetzungen ins Niederländische folgen diesem Gesetz. Die Autoren verwenden als Anpassungsmodell die (verschobene) negative Binomialverteilung, die von ALTMANN (1985, 179) aus den Geburts- und Todesprozessen der Bedeutungen einer Einheit abgeleitet wurde.

In ALTMANN, BEST & KIND (1987) wird die Verteilung der Bedeutungen von deutschen Affixen untersucht. Die Anpassung der (verschobenen) negativen Binomialverteilungan an die Daten gelingt aber nicht in allen Fällen (vgl. S. 130). Unter der Annahme, dass hier mehrere Diversifikationsprozesse gleichzeitig ablaufen, schließen die Autoren, dass es sich hier um eine gemischte Verteilung handeln muss. Sie zeigen, dass sich die gemischte negative Binomialverteilung gut an die Daten anpassen lässt.

Mit der gemischten negativen Binomialverteilung modelliert BEST (1990) erfolgreich einen frühneuhochdeutschen Wortbildungstyp bei Substantiven. Es handelt sich hierbei um Substantive des Typs "Geschöpf", "Gesetz" und "Gesicht", deren Auftreten nach Funktionsständen geordnet wird. NEMCOVÁ (1991) untersucht die semantische Diversifikation der Bedeutungen slowakischer Verbalpräfixe. An die Daten passt sie zwei Modelle an, die 0-gestutzte negative

[7] Siehe auch BEÖTHY & ALTMANN (1984).

Binomialverteilung und die Zipf-Alekseev-Verteilung[8]. Mit dem letztgenannten Modell erzielt NEMCOVÁ in allen Fällen hervorragende Ergebnisse:

> „As can easily be seen the Zipf-Dolinsky distribution yields in all cases an excellent result even if the fit could not always be tested by means of the chi-square test.„ (NEMCOVÁ 1991, 71).

Die 0-gestutzte negative Binomialverteilung ließ sich auch mit guten Ergebnissen anpassen.

Untersuchungen zur semantischen Diversifikation der Bedeutungen französischer Affixe konnten nicht gefunden werden. Es liegen aber etliche Arbeiten zur funktional-semantischen Diversifikation anderer Entitäten vor. ROTHE (1985) beschäftigt sich mit der Semantik der französischen Konjunktion "et". An die verschiedenen Bedeutungen bzw. Verwendungsweisen dieser Konjunktion konnte sie die 1-verschobene negative Binomialverteilung mit Erfolg anpassen. Für das Französische zitiert BEST (2001a, 87) eine weitere Untersuchung, nämlich die von HUG (2000), der auf der Grundlage eines Korpus das Auftreten des Französischen "que" nach vorhergehenden Wörtern und dessen verschiedene Verwendungsweisen untersucht. An die im ersten und zweiten Fall ermittelten Daten kann BEST (2001a, 88-89) jeweils die 1-verschobene Hyperpoisson-Verteilung und die modifizierte negative Binomialverteilung mit sehr gutem Ergebnis anpassen.

2.2.1.2. Ranghäufigkeitsverteilung von Wörtern

Zur Untersuchung der Häufigkeitsverteilung von Wörtern in einer Sprache werden in der Regel alle Wörter − eines vorliegenden Textkorpus − in einer Rangfolge nach ihrer Häufigkeit aufgelistet, wobei die absolute Häufigkeit des Wortes notiert wird. Betrachtet man die Worthäufigkeit in Abhängigkeit vom Rang, so ergibt sich eine Ranghäufigkeitsverteilung der Wörter. Rang und Häufigkeit sind dabei indirekt proportional zueinander.

Erste systematische Untersuchungen zu der Ranghäufigkeitsverteilung von Wörtern wurden von ZIPF (1949) durchgeführt. Stellt man die Häufigkeiten der einzelnen Einheiten eines Textes fest und gibt man dem häufigsten Wort den Rang 1, dem zweithäufigsten den Rang 2, usw. ergibt sich nach ZIPF aus dem Produkt Rang und Frequenz eine etwa gleich bleibende Zahl. Die entsprechende Gleichung hat die Form:

$$(1) \qquad r \times f = K$$

ZIPF begründet sein Gesetz mit der Neigung des Menschen zur geringsten Anstrengung, d.h. mit dem Ausgleich zwischen dem Versuch des Produzenten nach wenig und dem des Rezipienten nach viel Vokabular. „The entire behavior of an individual is at all times motivated by the urge to minimize effort“, so ZIPF (1949, 3).

[8] In der Literatur findet man gelegentlich auch die Bezeichnung "Zipf-Dolinsky"-Verteilung.

ZIPFs Gesetz, welches auf den Beobachtungen von ESTOUP (1916) zur Verbesserung des Sprachunterrichts bzw. zur Konstruktion eines optimalen stenographischen Codes beruht, modelliert die mittleren Ränge sehr gut, die sehr häufigen und seltenen Ränge aber nicht; hier kommt es zu großen Abweichungen.

Dieses Gesetz wurde von MANDELBROT (1953, 1954, 1957) auf Grund von Ökonomie-Argumenten erweitert zu:

$$(2) \qquad P_x = \frac{(b+x)^{-a}}{F(n)}, \quad x = 1, 2, 3, \ldots n,$$

wobei n der Inventarumfang, a und b Parameter sind. Das modifizierte Gesetz ist heute als Zipf-Mandelbrot-Gesetz bekannt. Laut ALTMANN (1985b, 10) zählt es zu den wohl größten linguistischen Leistungen. Damit lassen sich Ranghäufigkeiten im Bereich der sehr häufig und sehr selten auftretenden Wörter besser beschreiben.

Längere Wörter sind eindeutiger und werden in der Regel seltener verwendet als kürzere Wörter. Letztere haben grundsätzlich viele Bedeutungen. In den Ranghäufigkeitsverteilungen wird dieser Tatsache dadurch Rechnung getragen, dass die Vorkommenshäufigkeit von Wörtern zu ihren Rängen umgekehrt proportional ist.

Neben der Mandelbrot'schen Modifikation des Zipf'schen Gesetzes wurde eine große Zahl alternativer Modelle und Modifikationen entwickelt. Die Literatur zählt hunderte von Arbeiten (vgl. ALTMANN 1988a, 70). Einen kleinen Überblick findet man in GUITER & ARAPOV (1982). Die umfangreichsten Informationen liefern CHITASHVILI & BAAYEN (1993).

Das Zipf-Mandelbrot-Gesetz wird immer herangezogen, um die Ranghäufigkeitsverteilung von Wörtern zu erfassen. Bereits BILLMEIER (1969) hatte versucht, diese Verteilung einer Überprüfung am deutschen Textmaterial zu unterziehen. Die Ergebnisse seiner Untersuchung fasst er mit folgenden Worten zusammen: „Die Untersuchung zeigt, dass die Zipf'schen Gesetze – einschließlich ihrer Modifikationen – für deutsche Texte nur grob die empirischen Fakten wiedergeben.” (BILLMEIER 1969, 45)

Neben der Untersuchung, ob die nach absteigender Frequenz geordneten Häufigkeiten der Wörter eines vollständigen Textes gemäß dem Zipf-Mandelbrot-Gesetz verteilt sind, geht KNÜPPEL (2001) auch der Frage nach, ob die Häufigkeiten der Wörter einzelner Wortlängenklassen (Einsilber, Zweisilber, etc) ebenfalls der Zipf-Mandelbrot-Verteilung folgen. Als Datengrundlage benutzt sie 20 deutsche Texte bzw. Briefe aus einem Zeitraum von 1811 bis 1983. Die in den einzelnen Texten ermittelten empirischen Ranghäufigkeitsverteilungen lassen sich hervorragend mit der Zipf-Mandelbrot-Verteilung erfassen; die Autorin erzielt in allen Fällen Wahrscheinlichkeiten von P ≈ 1. Das gleiche Ergebnis erzielt sie bei 73 der 75

bearbeiteten Ranghäufigkeitsverteilungen in den einzelnen Wortlängenklassen. Zwei Dateien belegen mit P = 0.99 bzw. 0.89 ebenfalls ausgezeichnete Resultate.
Dass das Zipf-Mandelbrot-Gesetz auch dann gilt, wenn die Ranghäufigkeitsverteilungen der Wörter einer Längenklasse betrachtet werden, konnte bereits UHLÍŘOVÁ (1995) an 10 Texten der tschechischen Sprache zeigen.
Anhand der erzielten hervorragenden Ergebnisse will KNÜPPEL (2001, 265) zeigen, dass „Billmeiers Schluss, die Gültigkeit des Gesetzes [d.h. des Zipf-Mandelbrot-Gesetzes] für die Modellierung von deutschen Texten pauschal in Frage zu stellen (...), voreilig war."

Die Angemessenheit des Zipf-Mandelbrot-Gesetzes für Ranghäufigkeitsverteilungen konnten auch ALTMANN (1988a, 72ff.) und BEST (2001a, 78ff.) bestätigen. ALTMANN passte dieses Modell an die Rangverteilung der Phoneme, Wortformen sowie Lexeme in Goethes *"Erlkönig"* an. Für die Phoneme ergab das Modell ein Resultat von P = 0.71, während es bei Wortformen und Lexemen jeweils zu einem Ergebnis von P ≈ 1 führte. Das gleiche Ergebnis erzielte BEST für die Rangverteilung der Wörter in LICHTENBERGs *"Sudelbücher"*. BEST (2001a, 81) weist auch darauf hin, dass man an die Ranghäufigkeitsverteilungen der Wörter in PESTALOZZIs *"Hühner, Adler und Mäuse"* die Zipf-Mandelbrot-Verteilung mit gleich gutem Ergebnis anpassen könne. Er erklärt die negative Einschätzung dieser Verteilung durch BILLMEIER mittels der von diesem Autor verwendeten Texte:

> „Es handelt sich bei ihm z.T. um Textmischungen, auch um sehr umfangreiche Werke, wie Kants *Kritik der reinen Vernunft*, bei denen man Probleme mangels hinreichender Homogenität geradezu erwarten muss." (BEST 2001a, 81)

In ORLOV (1982a, 1982b) werden weitere Untersuchungen und Weiterentwicklungen des Zipf-Mandelbrot-Gesetzes in der Theorie und in Anwendung auf Textebene unternommen. ORLOV sieht in der Abgeschlossenheit des untersuchten Textes eine wesentliche Bedingung für die Gültigkeit der Zipf-Mandelbrot-Verteilung. Dies begründet er mit dem „Zipfschen Umfang", der Länge, mit der ein Autor im Voraus seinen Text plant.

GRZYBEK (2000) untersucht die Rangverteilung der 1000 häufigsten Wörter aus dem Cortes-Korpus. Die Modellierung durch das Zipf-Mandelbrot-Gesetz zeigt eine gute Übereinstimmung zwischen empirischen und theoretischen Werten.

Eine Datei zur Ranghäufigkeitsverteilung der Wörter im Französischen liegt in MULLER (1968, 167-168) vor. Es handelt sich hierbei um die Rangverteilung der Wörter in P. CORNEILLEs *"L'illusion comique"*, Rolle von Alacandre. MULLER wendet hier das Zipf-Mandelbrot-Gesetz in seiner ursprünglichen Form an. Er berechnet das Produkt von Rang und Frequenzrang. Über die dabei erzielten Ergebnisse schreibt er:

„Le résultat est assez frappant: à part les 5 fréquences les plus élevées et la fréquence 1, tous les produits se situent entre 240-320, avec une montée lente jusqu'à la fréquence 20, puis une descente.” (MULLER 1968, 167)

Die Zipf'schen Überlegungen haben eine ganze Reihe von wissenschaftlichen Forschungs-Richtungen ins Leben gerufen. So wurden beispielsweise in Russland von ARAPOV (1988) und ARAPOV & CHERC (1983) auf der Grundlage von Rangordnungsanalysen Modelle für die Textdynamik und Textentwicklung entwickelt. TULDAVA (1995, 1998) verdankt man mathematische Analysemethoden für zahlreiche Texterscheinungen.

2.2.1.3. Ranghäufigkeitsverteilung von Buchstaben und Phonemen

Die Beschäftigung mit Buchstabenhäufigkeiten stellt eine der traditionellsten Fragestellungen in der Geschichte der quantitativen Linguistik dar. In der Vergangenheit ging es in den Untersuchungen um die bloße Erhebung von Buchstabenhäufigkeiten; das Interesse war auf die Vorkommenshäufigkeit der einzelnen Grapheme ausgerichtet. So beschäftigte sich beispielsweise FÖRSTEMANN (1846, 1852) mit den relativen und wechselnden Häufigkeiten u.a. von Buchstaben im Alt- und Mittelhochdeutschen sowie im Griechischen und Lateinischen. Entsprechende Untersuchungen zum Altkirchenslawischen findet man in SCHLEICHER (1852).
In der jüngeren Zeit rückte immer wieder die theoretische Frage nach der Regularität bzw. Modellierbarkeit der Buchstabenhäufigkeiten in den Vordergrund der Untersuchungen. Neben der Häufigkeit der einzelnen Buchstaben in verschiedenen Texten und Korpora ging es vor allem um das systemische Häufigkeitsverhalten von Buchstaben. Man überführte zu diesem Zweck zuerst die Häufigkeitsverteilung in eine Rangverteilung, bei der die Frequenz des häufigsten Buchstabens den ersten Rang, die des zweithäufigsten den zweiten, usw. zugewiesen bekam. Anschließend untersuchte man, ob zwischen den Häufigkeiten der einzelnen Ränge bestimmte Beziehungen bestehen, und ob bzw. wie diese Relationen am besten beschrieben werden können. D.h.: Untersucht wurde vorrangig die Ranghäufigkeitsverteilung von Buchstaben, wobei das Ziel der theoretischen Modellierung darin bestand, den Abstand zwischen den jeweiligen Häufigkeiten mathematisch zu modellieren.

Es gibt heute eine Reihe von Untersuchungen, die sich mit dem Ranghäufigkeitsverhalten von Buchstaben befassen. Die meisten dieser Untersuchungen sind den slawischen Sprachen gewidmet: Für das Russische in GRZYBEK & KELIH (2003), GRZYBEK, KELIH & ALTMANN (2004) und GRZYBEK, KELIH & ALTMANN (2005); für das Slowakische in GRZYBEK, KELIH & ALTMANN (2006) und für das Ukrainische in GRZYBEK & KELIH (2005a). In all diesen Arbeiten erstreckt sich der Inventarumfang von Buchstaben im Bereich von I = 25 (Slowenisch) bis I = 46 (Slowakisch mit Diakritien). Die Untersuchungen werden an Daten mit verschiedenen Ei-

genschaften durchgeführt: Es werden vollständige Texte, Textausschnitte, -kumulationen und -kombinationen sowie ein sich aus diesen Texten zusammensetzendes Gesamtkorpus ausgewertet.

Für die Modellierung der Ranghäufigkeitsverteilung der Buchstaben in diesen Sprachen haben die Autoren die Eignung bzw. Adäquatheit aller relevanten Ansätze erprobt, die in der quantitativen Linguistik für solche Probleme in Betracht gezogen wurden bzw. werden. Es stellte sich heraus, dass die meisten dieser Modelle ungeeignet sind; lediglich ein einziges Modell, die negative hypergeometrische Verteilung, hat sich in allen untersuchten Sprachen durchgehend als passendes Modell sowohl für die Einzeltexte als auch für das Gesamtkorpus erwiesen. Diese Verteilung konnte auch an die Phoneme des Russischen mit Erfolg angepasst werden (vgl. z.B. KELIH 2007).

Das Modell der negativen hypergeometrischen Verteilung konnte BEST (2005a) auch bei seinen umfangreichen Untersuchungen zu Ranghäufigkeitsverteilungen von Buchstaben und anderen Schriftzeichen, Lauten und Phonemen im Deutschen vielfach bestätigen; diese Verteilung hat sich immer wieder als sehr gutes Modell für derartige Phänomene erwiesen:

> „In allen hier vorgelegten Fällen wurde nachgewiesen, dass die Buchstaben, in eine Ranghäufigkeitstabelle gebracht, entsprechend der negativen hypergeometrischen Verteilung vertreten sind. Das gilt sowohl für die Einzeltexte als auch für die hier betrachteten Textkorpora (…)." (BEST 2005a, 27)

GRZYBEK (2007a, b) hat die Daten von BEST (2005a) nochmals analysiert. In dieser erneuten Analyse wurde das Datenmaterial ausschließlich auf die eigentlichen Buchstaben reduziert. An dieses modifizierte Datenmaterial konnte er wieder die negative hypergeometrische Verteilung mit gutem Erfolg anpassen:

> „Without exception, the negative hypergeometric distribution turns out to be a very good model: With the exception of the very first text (extremely short at 675 letters), the values of the discrepancy coefficients are in the interval of $0.0046 \leq C \leq 0.0140$ – thus proving the negative hypergeometric distribution indeed to be a good model. " (GRZYBEK 2007a, 86)

Die negative hypergeometrische Verteilung ist kein immer geeignetes Modell für die Buchstabenverteilung im Deutschen. So wird in ALTMANN (1988a, 69ff.) und BEST (2001a, 78-85) gezeigt, dass Buchstaben und Phoneme in etlichen Fällen gemäß der Zipf-Mandelbrot-Verteilung auftreten; in anderen Fällen folgen sie der geometrischen Verteilung (vgl. ALTMANN & LEHFELDT 1980, 144). BEST (2005a, 28) weist darauf hin, dass bei der Anpassung der Häufigkeitsränge der Anfangs- und der Endbuchstaben in einem Lexikon die negative

hypergeometrische Verteilung nicht als das am besten geeigneten Modell betrachtet werden könne.

Arbeiten zur theoretischen Modellierung der Buchstabenverteilungen im Französischen konnten nicht gefunden werden. Es gibt zwar seit langem Untersuchungen zur Häufigkeit der Buchstaben des Französischen, die in Arbeiten wie z.B. von GALLAND (1941) und NASVYTIS (1953) fortgesetzt werden und heutzutage in verschiedenen Internetquellen Anwendung finden[9]. Doch all diese Quellen weisen Probleme auf, die eine systematische Analyse erschweren: Abgesehen davon, dass die meisten Quellen auf unterschiedlichen Inventarumfängen beruhen, findet man keinen Hinweis darauf, auf welches Material sich die gemachten Angaben beziehen. Zudem werden nur relative Häufigkeiten angegeben, die für eine Reihe weiterführender Fragen unzureichend sind.

Es kann festgehalten werden, dass sich die bisherigen Arbeiten nur mit der Frage nach der Häufigkeit jeweils individueller Buchstaben beschäftigen. Solche Studien können laut GRZYBEK (2007b, 108-109)

> „im Rahmen von systematischen und system-bezogenen Untersuchungen allerdings nicht mehr und nicht weniger als Rohmaterial und Ausgangspunkt für weiterführende Fragestellungen sein: Denn unter system-bezogener bzw. systemtheoretischer Perspektive geht es nicht um das spezifische Verhalten individueller Elemente in verschiedenen Stichproben, sondern um deren systemisches Verhalten im relationalen und funktionalen Kontext des gegebenen Systems (…)."

Die Verteilung der Buchstabenhäufigkeit in französischen Texten ist also ein bisher kaum erforschtes Feld.

2.2.1.4. Häufigkeitsverteilung der Satzlängen

Der Frage, inwiefern Satzlängen in Texten gesetzmäßig verteilt sind, ist von vielen Autoren nachgegangen worden. Arbeiten, die sich mit diesem Sachverhalt befassen, reichen bis in die 40er Jahre des zwanzigsten Jahrhunderts zurück. So wurde bereits von YULE (1939), der einige englische Daten gesammelt hat, die Vermutung geäußert, dass die Vorkommenshäufigkeit der Satzlänge in Texten nicht chaotisch ist, sondern bestimmten Gesetzmäßigkeiten unterliegt. YULE legt sich zwar auf kein Verteilungsmodell fest, weist aber auf Folgendes hin: „They [Sätze] are not of the poisson type but of the type in which the square of the standard derivation largely exceeds the mean." (S. 371)[10]

---

[9] Siehe z.B. die Internetseiten: http://www.cryptogram.org/cdb/words/frequency.txt#calcs, http://www.central.edu/homepages/LintonT/classes/spring01/cryptography/letterfreq.html

[10] Vgl. dazu auch NIEHAUS (1997, 216)

In Anlehnung an die Yule'sche Untersuchung schlug WILLIAMS (1940) vor, bei der Satzlängenverteilung statt der absoluten Anzahl der Wörter pro Satz die jeweilige logarithmische Wortanzahl pro Satz als Zufallsvariable zu betrachten (vgl. ALTMANN 1988b). Er beschränkte sich bei der Re-Analyse der Daten von YULE auf die graphische Darstellung. Diese zeigt seiner Ansicht nach, dass die Häufigkeitsverteilung einer Normalverteilung unterliegt. Auf Grund der empirischen Untersuchung von drei Texten postulierte er die Lognormal-Verteilung als Modell für die Satzlängenverteilung (vgl. WILLIAMS 1940, 360). Die Begründung für diese Verteilung sieht ALTMANN (1988b, 150) darin, dass

> „man für einen Textautor einen festen Satzlängenmittelwert als sein Charakteristikum annimmt und zufällig ‚normale', jedoch logarithmische transformierte Abweichungen von diesen Mittelwert vermutet."

Gegen den Williams'schen Ansatz brachte SICHEL Argumente statistischer Natur. Er kritisierte, dass die Lognormal-Verteilung nur auf der Basis einer graphischen Darstellung der Satzlängenverteilung postuliert wurde und dieses Modell keinem statistischen Test unterzogen wurde, der einen entsprechenden Nachweis der Übereinstimmung zwischen Theorie und Empirie erbringe. SICHEL schlug seinerseits die zusammengesetzte Poisson-Verteilung als theoretisches Modell der Satzlängenverteilung vor. Er testete das Modell an Sprachmaterial des Lateinischen, Griechischen und Englischen.
ALTMANN (1988b, 151) beurteilt diese Verteilung wie folgt:

> „Sichels Verteilung bringt viele Vorteile mit sich, weil sie zahlreiche spezial- oder Grenzfälle hat, die in der Linguistik bereits angewendet wurden, sodaß man sie nicht außer Acht lassen sollte, sondern ihre Tragweite in der Linguistik weiter prüfen und ihre Interpretation versuchen sollte."

Obwohl die Lognormal-Verteilung und die zusammengesetzte Poisson-Verteilung gute Übereinstimmung mit den empirischen Befunden ermöglichen bzw. ermöglichten, wurden sie einer allgemeinen Kritik unterzogen (vgl. z.B. ALTMANN 1988a). Demgegenüber stellte ALTMANN (1988b) einen grundlegenden Neuansatz vor. Dieser besteht darin, die Verteilung der Satzlängen, gemessen sowohl in der Anzahl der Teilsätze als auch in der Wortanzahl, auf der Grundlage der Theorie der linguistischen Synergetik zu erforschen. ALTMANN zieht bestimmte systeminterne und -externe Faktoren in Betracht, die seiner Ansicht nach auf die Satzlängenverteilung einwirken, nämlich (a) den Einfluss des Produzenten, (b) den Einfluss des Rezipienten, (c) Textfaktoren und (d) Faktoren der Ebene.

Davon ausgehend, dass jegliche Längenverteilung im Text einer Gesetzmäßigkeit folgt und der relative Unterschied zweier jeweils benachbarter Wahrscheinlichkeiten P_x und P_{x-1} der Satzlängen x und x-1 eine Funktion von x ist, stellt ALTMANN (1988b) die folgende Formel auf:

$$(3) \qquad D = \frac{P_x - P_{x-1}}{P_{x-1}}$$

Die Differenz D ist laut ALTMANN nicht konstant, sondern ändert sich mit x, der Satzlänge. Welchen der weiter oben genannten Einflussfaktoren diese Differenz unterliegt, hängt von der Einheit ab, in deren Anzahl die Satzlänge gemessen wird. Im Hinblick auf die Operationalisierung der Satzlänge nach der Anzahl der direkten Konstituenten (Clauses) kommt der Faktor (d) nicht zum Tragen; damit ergibt sich nach ALTMANN (1988b) die folgende Gleichung:

$$(4) \qquad D = \frac{b - ax}{cx}$$

Durch Einsetzen von (4) in (3) und nach Umformungen erhält ALTMANN (1988b, 152ff.) die negative Binomialverteilung, deren Formel:

$$(5) \qquad P_x = \binom{k + x - 1}{x} p^k q^x, \quad x = 0, 1, 2, \cdots n$$

lautet, wobei 0 < p < 1, q = 1-p und k > 0.

Misst man aber die Satzlänge in Wortanzahl, kommt der Faktor (d) hinzu, was nach ALTMANN zu der Gleichung (6) führt:

$$(6) \qquad D = \frac{b - ax}{cx + d}$$

Durch Einsetzen von (6) in (3) und nach Reparametrisierung erhält ALTMANN (1988b, 15ff.) die Hyperpascal-Verteilung mit der Formel:

$$(7) \quad P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0, x = 0,1,2,\ldots n$$

D.h.: Für die Satzlänge, gemessen in der Anzahl der Teilsätze pro Satz, wird die negative Binomialverteilung als adäquates Modell hergeleitet. Falls die Satzlänge in der Wortanzahl gemessen wird, wird die Hyperpascal-Verteilung als geeignetes Modell postuliert.

Beide Modelle hat ALTMANN (1988b) mit gutem Erfolg erprobt: Die Wort/Satz-Variante hat er an 245 altgriechischen, englischen und slowakischen Texten überprüft, wobei das postulierte Modell nur in 9 Fällen versagte (vgl. ebd. S.158). Für die Clause/Satz-Variante wurden 10

Texte unterschiedlicher Sprachen ausgewertet, wobei die Anpassung der negativen Binomialverteilung in allen Fällen gelang (vgl. ebd., S.159).

Es sind inzwischen weitere Untersuchungen zur Satzlängenverteilung durchgeführt worden. Für die Wort/Satz-Verteilung liegen mehrere Arbeiten zu verschiedenen Sprachen vor. BEST (2001b) untersuchte die Satzlängenverteilung in 25 deutschen Texten aus drei Funktionalstilen, wobei Texte auf zwei Arten ausgewertet wurden: (a) Zunächst wurden in jedem Text alle Satzlängen einzeln aufgelistet; (b) die Häufigkeitsklassen wurden in einem zweiten Schritt in Intervallen von 1-5, 6-10, etc. zusammengefasst. Die Anpassung der Hyperpascal-Verteilung an die aus dieser zweifachen Auswertung der Texte entstandenen Dateien lieferte insgesamt keine guten Ergebnisse. An die Textdateien konnte hingegen die negative Binomialverteilung in ihrer 1-verschobenen Form mit guten Resultaten angepasst werden. Bei der Anpassung der 1-verschobenen negativen Binomialverteilung an die Dateien ohne Zusammenfassung der Längenklassen wurden nur in zwei Texten keine guten Ergebnisse erzielt. Die Anpassung derselben Verteilung an die Dateien mit zusammengefassten Längenklassen lieferte ein ähnlich gutes Ergebnis. Auf Grund seiner Anpassungsergebnisse schlug BEST (2001b) die 1-verschobene negative Binomialverteilung als theoretisches Modell für die Wort/Satz-Verteilung im Deutschen vor.

Als geeignet erweist sich dieses Modell auch in der Arbeit von NIEHAUS (2001), in der die Verteilung der Satzlänge in 20 literarischen Texten des Deutschen untersucht wird. Wie bei BEST (2001b) erfolgt auch bei NIEHAUS die Auswertung der Texte mit und ohne Zusammenfassung der Satzlängen. Für die in beiden Fällen ermittelten Dateien erweist sich die 1-verschobene negative Binomialverteilung als ein recht geeignetes Modell; mit dieser Verteilung erzielt NIEHAUS bessere Ergebnisse als mit der 1-verschobenen Hyperpascal-Verteilung (NIEHAUS 2001, 210).

Das sich bei BEST (2001b) und NIEHAUS (2001) bewährte Modell der 1-verschobenen negativen Binomialverteilung hat BEST (2002) einer zusätzlichen Überprüfung anhand von weiteren 20 Pressetexten unterzogen. Auch hier konnten die bereits bei anderen Untersuchungen erzielten guten Anpassungsergebnisse untermauert werden:

> „Es konnte am Beispiel eines weiteren Datensatzes zur Pressesprache gezeigt werden, dass Satzlängen in abgeschlossenen Texten bestimmten Gesetzen gehorchen. Bei deutschen Texten scheint es meist die [1-verschobene] negative Binomialverteilung zu sein, wenn man die Satzlänge nach der Zahl der Wörter pro Satz bestimmt.” (BEST 2002, 27)

Die Wort/Satz-Verteilung in 20 englischsprachigen Texten untersuchen KAßEL & LIVESEY (2001). Ähnlich wie bei BEST (2001b) und NIEHAUS (2001) werden die Texte mit und ohne Zu-

sammenfassung der Satzlängen ausgewertet. Für Dateien mit Aufführung der einzelnen Sätze erweist sich die 1-verschobene negative Binomialverteilung als gutes Modell. An die Dateien mit Zusammenfassung von Längenklassen lässt sich die 1-verschobene gemischte Poissonverteilung erfolgreich anpassen (vgl. KAßEL & LIVESEY 2001, S. 31ff.).

Laut der Untersuchung von ROUKK (2001a) stellt sich für die Wort/Satz-Verteilung im Russischen die 1-verschobene Hyperpoisson-Verteilung als geeignetes Modell dar. Dieselbe Verteilung erweist sich laut BEST (2005c, 300) in der Untersuchung von JING (2001) als geeignet für die Modellierung der Satzlängen im Chinesischen. Die Angemessenheit dieses Modells konnte auch GRZYBEK (1999) auf der Grundlage einer slowenischen Sprichwortsammlung nachweisen.

Weitere Untersuchungen zur Wort/Satz-Verteilung in slowenischen Texten und Texten anderer slawischer Sprachen findet man in GRZYBEK & KELIH (2004).

Zur Clause/Satz-Verteilung wurden auch etliche Untersuchungen anhand von Daten verschiedener Sprachen durchgeführt. Für das Deutsche findet sich eine solche Untersuchung beispielsweise in NIEHAUS (1997). Die Autorin untersucht die Satzlängenverteilung in 85 Texten unterschiedlicher Funktionalstile. Für diese Dateien erweist sich die 0-gestutzten negativen Binomialverteilung als recht adäquates Modell: Nur in 7 von insgesamt 85 Fällen versagt dieses Modell. Aus ihren Untersuchungsergebnissen schließt NIEHAUS (1997, 257), dass

> „die negative Binomialverteilung in ihrer 0-gestutzten Form als Modell der Satzlängen-Häufigkeitsverteilung, gemessen in der Anzahl der Teilsätze, bis auf weiteres beibehalten werden kann, da nur wenige Texte nicht der Verteilung folgen und sich in der Regel Gründe für die Abweichungen nennen lassen."

Spätere Berechnungen der Daten von NIEHAUS haben nach BEST (2005c, 301) ergeben, dass mit der 1-verschobenen Hyperpoisson-Verteilung sogar noch bessere Ergebnisse erzielt werden können. Dieses Modell konnte sich laut BEST (2005c, 301) in weiteren Untersuchungen zum Deutschen, nämlich STREHLOW (1997) und WITTEK (2001) zum größten Teil bewähren. An den insgesamt 245 untersuchten Textdateien konnte dieses Modell nur in einem Fall nicht angepasst werden (vgl. BEST 2005c, 301).

Die 0-gestuzte negative Binomialverteilung und die 1-verschobene Hyperpoisson-Verteilung haben sich auch als geeignet für die Modellierung der Satzlängen im Chinesischen (vgl. BOHN 1998) und Russischen (vgl. ROUKK 2001b) erwiesen. BOHN (1998) benutzte als Datengrundlage ein chinesisches Korpus und einen Einzeltext. An die Daten aus dem Korpus ließ sich die 0-gestutzte negative Binomialverteilung besser anpassen, während sich für den Einzeltext die

1-verschobene Hyperpoisson-Verteilung als geeignetes Modell erwies. Die Untersuchung von ROUKK ergab, dass

> „die positive [oder 0-gestutzte] negative Binomialverteilung insgesamt ein etwas besseres Ergebnis für die 22 Erzählungen von Tschechow ergab, da sie sich in allen Fällen mit gutem Ergebnis anpassen ließ; die Hyperpoisson-Verteilung konnte an einen der Texte nur mit einem nicht zufrieden stellenden Ergebnis angepasst werden. ” (ROUKK 2001b, 114)

Die Frage der Beschreibung der Satzlängenverteilungen durch theoretische Modelle ist im Französischen wenig erforscht. Es konnte keine Untersuchung in diesem Bereich gefunden werden.

2.2.1.5. Häufigkeitsverteilung in Textblöcken

Die Idee, die Häufigkeitsverteilung von sprachlichen Entitäten in Textabschnitten bestimmter Länge zu untersuchen, geht auf die 30er Jahre des zwanzigsten Jahrhunderts zurück (vgl. BEST 2005b, 1). Erste Untersuchungen dieser Art fanden sich bereits in ZWIRNER & ZWIRNER (1935, 38). Beide Autoren untersuchten die Häufigkeitsverteilung einiger Laute des Deutschen in 200 Textblöcken mit je 100 Lauten und stellten dabei fest, dass das Auftreten dieser Einheiten „dem Gesetz der kleinen Zahlen” unterliegt (ZWIRNER/ZWIRNER 1935, 44). Dies ist eine andere Bezeichnung für die Poissonverteilung. ZWIRNER & ZWIRNER führten an die Vorkommenshäufigkeit der untersuchten Laute eine Anpassung der Poisson-Verteilung durch, gaben aber in keinem Fall das Testkriterium P an, sodass nur ein optischer Vergleich zwischen den empirischen und theoretischen Werten möglich ist.

In der Nachfolge wurden weitere Untersuchungen zur Häufigkeitsverteilung in Textblöcken durchgeführt: FRUMKINA (1962) untersuchte die blockmäßige Wiederholung von mehreren russischen Wörtern[11] – aus einem Werk von PUSCHKIN – in 1000-Wort-Textblöcken. An die ermittelten Werte passte sie ebenfalls die Poisson-Verteilung an, wobei sie allerdings nicht in allen Fällen gute Ergebnisse erzielte; das Modell konnte sich an 5 von insgesamt 12 Fällen nicht bewähren (vgl. z.B. ALTMANN & BURDINSKI 1982, 148). BRAINERD (1972) untersuchte das Vorkommen englischer Artikel in 39 verschiedenen Textblöcken von je 50 Wörtern. Im Vergleich zu FRUMKINA war er „more successful in applying the poisson distribution to the frequencies (…) He demonstrated that with one exception all data fit the poisson distribution (...).” (ALTMANN & BURDINSKI 1982, 148)

[11] Es ging dabei nach ALTMANN & BURDINSKI (1982,148) u.a. um die Wörter "den'", "drugoj", "novyj", etc.

In PIOTROWSKI (1984) finden sich Untersuchungsergebnisse mehrerer russischer Autoren, die an die Häufigkeitsverteilung in Textblöcken auch eine Anpassung der Poisson-Verteilung durchgeführt haben.

Angeregt durch die vorhergehenden Arbeiten entwickelten ALTMANN & BURDINSKI (1982) aus der theoretischen Verteilung von Einheiten in Textpassagen ein Gesetz für die blockmäßige Wiederholung; sie schlugen die negative hypergeometrische Verteilung als Grundmodell vor, welches als Grenzfälle die Poisson-Verteilung, die Binomialverteilung und die negative Binomialverteilung umfasst (vgl. ALTMANN & BURDINSKI 1982,153; ALTMANN 1988, 177). Das nach FRUMKINA benannte Gesetz hat sich in dieser neuen Form in einer ganzen Reihe von Fällen bewährt. ALTMANN & BURDINSKI (1982, 155ff.) konnten eine Anpassung der negativen hypergeometrischen Verteilung an die Daten aus FRUMKINA (1962) und BRAINERD (1972) erfolgreich durchführen. Auch mit dieser Verteilung konnten die Autoren die blockmäßige Wiederholung des deutschen Artikels "das" (im Nominativ) in *"Deutschstunde"* von S. Lenz sowie die Häufigkeitsverteilung einiger bulgarischer und indonesischer Wörter erfolgreich modellieren.

Die Angemessenheit der negativen hypergeometrischen Verteilung für die blockmäßige Wiederholung wird auch durch die Untersuchungen in BEST (2001a, 2005b) unterstützt. Anhand der Erzählung *"Dazugehören"* von JÄGERSBERG untersucht BEST (2001a, 97-100) die Vorkommenshäufigkeit von 5 Wörtern in 50- und 100-Wort-Textblöcken. In BEST (2005b) geht es um die blockmäßige Wiederholung unterschiedlicher Einheiten, von Buchstaben über Laute bis hin zu Wörtern und Wortarten. In den beiden Untersuchungen stellt die negative hypergeometrische Verteilung ein gutes Modell dar.

Das Frumkina-Gesetz hat sich nicht nur im Bereich des grammatischen Wortschatzes, sondern auch in anderen Bereichen bewährt. PIOTROWSKI (1984, 143) und PIOTROWSKI et al . (1985, 252) konnten zeigen können, dass auch die Häufigkeitsverteilung von Drei-Wort-Segmenten in Textabschnitten dem Frumkina-Gesetz folgt. KÖHLER (2001) passt das Gesetz erfolgreich an die Häufigkeitsverteilung von Klauseln in deutschen Texten an. Als Verteilungsmodell benutzt er die negative Binomialverteilung, einen der Grenzfälle der negativen hypergeometrischen Verteilung. KÖHLER (2001, 147) vermutet:

> „ If future investigations (…) corroborate these results, we can conclude that Frumkina's law which was first found and tested for words, can be generalised (as already supposed by Altmann) to possibly all types of linguistic units (…)."

Erhebungen zur Häufigkeitsverteilung in Textabschnitten der französischen Sprache konnten nur in MULLER (1968, 81ff.) gefunden werden. Der Autor untersucht das Auftreten der Wör-

ter „et“ und „vous“ in 55 bzw. 17 Textabschnitten mit je 50 bzw. 100 Versen in *„Phèdre“* von RACINE. MULLER untersucht in zwei Fällen auch das Vorkommen von Bedeutungsgruppen grammatischer Wörter, nämlich die Negationswörter „ ne, n’, ni, non“ in 66 Textblöcken mit je 25 Versen und die Pronominalformen „tu, te, t’, ton, ta, tes“ in 55 Textabschnitten mit je 30 Versen. Als Textbasis der Untersuchung dient hier wiederum RACINEs ***„Phèdre“***. An die Daten führt MULLER aber keine Anpassung durch; auf Grund der Verfügbarkeit der Rohdaten ist aber eine Modellierung der Verteilung möglich. BEST (2005b, 7ff.) zeigt, dass diese Daten gemäß der negativen hypergeometrischen Verteilung auftreten.

2.2.1.6. Häufigkeitsverteilung der Komplexität und Frequenz von Teilsätzen

Sieht man von den Arbeiten zur Satzlängenverteilung ab, sind quantitative Analysen zur Distribution der Eigenschaften von Einheiten auf der höheren sprachlichen Ebene (Syntagma, Phrase, Clauses …) selten. Die Ursachen für die Zurückhaltung auf diesem Gebiet liegen wohl in erster Linie in der Uneinheitlichkeit der syntaktischen Lehre, die die Quantifizierung erschwert sowie in einem Mangel an Vorstellungen darüber, wie sich solche Untersuchungen in der Praxis verwerten lassen.

Arbeiten zur Verteilung der Eigenschaften von Teilsätzen in französischen Texten oder Texten anderer Sprachen konnten nicht gefunden werden. Es liegen aber einige Untersuchungen zur Distribution der Eigenschaften anderer syntaktischen Einheiten vor: KÖHLER (1999, 300ff.) setzt sich mit der Verteilung der Konstituentenhäufigkeiten und -komplexität im Susanne-Korpus und Negra-Korpus auseinander. Eine weitere Untersuchung in diesem Bereich findet man in KÖHLER & ALTMANN (2000). Beide Arbeiten verwenden einen phrasenstrukturgrammatischen Ansatz. Die quantitative Analyse zur Syntax in KÖHLER (2005b) beruht auf der Dependenzgrammatik und behandelt Valenz und distributionelle Eigenschaften deutscher Verben.

2.2.2. Zusammenhänge zwischen sprachlichen Eigenschaften

2.2.2.1. Zusammenhang zwischen Polylexie und Länge

Zu der Wortlänge gibt es eine Vielzahl von Arbeiten, die in verschiedenen Sprachen durchgeführt wurden[12]. Der Wechselbeziehung zwischen Bedeutungszahl und Wortlänge sind bisher aber nur einige Untersuchungen gewidmet worden.

Die erste Überprüfung der Abhängigkeit der Polylexie von der Wortlänge wurde vermutlich von ALTMANN, BEÖTHY & BEST (1982) durchgeführt. Die Autoren begründeten diese Be-

[12] Vgl. z.B. Glottometrika 16.

ziehung über das Menzerath-Gesetz. Sie stellten eine Hypothese über die Verringerung der Menge der Bedeutungen des Wortes bei wachsender Wortlänge auf:

> „Die Möglichkeiten sind aber damit nicht erschöpft, denn die Größe der Konstrukte (x) braucht nicht unbedingt in der Zahl der als *y* zu ihnen in bezug gesetzten Konstituenten gemessen zu werden. (…) Eine Beziehung, die nach unseren Annahmen dem Menzerathschen Gesetz folgen sollte, ist die Verringerung der Menge der lexikalischen Bedeutungen des Wortes bei wachsender Wortlänge.“ (ALTMANN, BEÖTHY & BEST 1982, 537)

D.h.: Die Verlängerung eines Wortes hat eine Bedeutungsspezifikation zur Folge. Letztere führt wiederum zu einer Verringerung der potentiellen Anzahl der Bedeutungen des Wortes. ALTMANN, BEÖTHY & BEST testeten die Hypothese an Daten des Deutschen, Ungarischen und Slowakischen mit Erfolg und ziehen den Schluss, dass das Menzerath-Gesetz sowohl auf der Ausdrucksseite als auch auf der Inhaltsseite Gültigkeit besitzt.

Ähnlich wie ALTMANN, BEÖTHY & BEST (1982) leitet ROTHE (1983) den Zusammenhang zwischen Polylexie und Wortlänge auch aus dem Menzerath-Gesetz ab:

> „(…) Diese Beziehung [Konstruktlänge und Konstituentenlänge] gilt für verschiedene sprachliche Ebenen und Bereiche: Die Länge eines Elements ist eine Funktion der Länge seiner hierarchisch nächsthöheren Einheit; aber auch die Menge der semantischen Repräsentanten kann als eine Funktion der Länge des Ausdrucks angesehen werden. Letztere der Beziehungen (…) zu überprüfen, ist Ziel der vorliegenden Arbeit.“ (ROTHE 1983, 101)

Die Überprüfung erfolgt hier an sprachlichem Material (Lexika) aus den romanischen Sprachen Französisch, Portugiesisch und Spanisch. Für jede Sprache wird eine Stichprobe von 1000 Wörtern aus dem jeweils benutzten Wörterbuch erstellt, wobei nur das letzte Wort jeder bzw. jeder zweiten Seite des benutzten Wörterbuches berücksichtigt wird. Die Wörter werden hinsichtlich ihrer Länge in Buchstaben und in Silben ausgewertet. ROTHE erzielt für die untersuchten Sprachen signifikante Ergebnisse.

Auch in FICKERMANN, MARKNER-JÄGER & ROTHE (1984) wird die Beziehung zwischen Polylexie und Länge als Konsequenz des Menzerath-Gesetzes angesehen:

> „Es wird angenommen, dass mit der Vergrößerung der Länge des Wortes seine Bedeutung spezifiziert wird und daher seine Bedeutungskomplexität abnimmt.“ (S. 115)

Der Zusammenhang wird an Daten aus dem Englischen, Schwedischen und Indonesischen überprüft. Die Autoren erzielen dabei signifikante Resultate für alle drei Sprachen.

Für das Polnische und Russische zitieren FICKERMANN, MARKNER-JÄGER & ROTHE (1984, 115) eine Untersuchung von SAMBOR (1983).

In allen bisher dargestellten Arbeiten zur Wechselbeziehung zwischen Polylexie und Länge wurde die Potenzkurve der Form

$$y = ax^b$$

an die Daten angepasst. Der Parameter *a* dieser Kurve wird als durchschnittliche Zahl der Bedeutungen von Wörtern der Länge 1 und die Konstante *b* als Umfang der Strukturinformation interpretiert.

In KÖHLER (1986) wird im Vergleich zu anderen Arbeiten die Beziehung zwischen Polylexie und Wortlänge nicht über das Menzerath-Gesetz, sondern über den Mechanismus der Bedeutungsspezifikation, d.h. über das Spezifikationsbedürfnis begründet. Letzteres repräsentiert „die Notwendigkeit, gegebene Mehrdeutigkeiten und Vagheiten einer lexikalischen Einheit zu verringern. “ (KÖHLER, 1990a, 3)

Zur Bedienung eines Systembedürfnisses gibt es verschiedene Methoden. KÖHLER (1990b, 183) unterscheidet dazu folgende Mittel der Kodierungs- und Spezifikationsbedürfnisse: die lexikalische, die syntaktische, die morphologische und die prosodische Methode. Die Verringerung der Mehrdeutigkeiten eines Ausdrucks kann durch die Verwendung von syntaktischen (zum Beispiel durch einen Relativsatz oder eine Apposition) oder morphologischen (zum Beispiel durch Komposition oder Affigierung) Mitteln erreicht werden. Die Verwendung eines dieser Mittel wirkt sich nach KÖHLER (1986) auf den Zusammenhang zwischen Polylexie und Länge aus. Von den beiden Methoden beeinflusst in erster Linie die morphologische Methode die Länge aus:

> „Die Abhängigkeit der Polylexie von der Länge ist umso stärker, je mehr eine Sprache von morphologischen gegenüber syntaktischen Mitteln zur Bedeutungsspezifikation Gebrauch macht. Diese typologische Eigenschaft einer Sprache soll Synthetizität heißen.“ KÖHLER (1986, 60)

Den Zusammenhang zwischen Polylexie und Länge überprüft KÖHLER (1986) erfolgreich anhand von 1325 Lemmata aus dem LIMAS-Korpus. Er führt dabei eine Anpassung der Funktion

$$y = ax^b$$

an die Daten durch. Der Parameter *a* wird von ihm als Kompromiss aus dem Einfluss der Bedürfnisse nach Minimierung des Kodierungsaufwands und Minimierung des Dekodierungsaufwands, die Konstante *b* als Grad der Synthetizität der untersuchten Sprache, d.h. des Einflusses der Bedeutungsspezifikation auf die Polylexie interpretiert.

Weitere Arbeiten, in denen die Abhängigkeit der Polylexie von der Wortlänge untersucht wird, sind u.a. HAMMERL (1991), in der ein synergetisch-linguistisches Modell für die Lexik

des Polnischen aufgestellt und überprüft wird, und GIESEKING (2003), in der das Köhler'sche Modell der Lexik (1986) einer Überprüfung für das Englische unterzogen wird.

2.2.2.2. Zusammenhang zwischen Länge und Frequenz

Erste Studien zur Beziehung zwischen Worthäufigkeit und Wortlänge gehen auf die 1. Hälfte des zwanzigsten Jahrhunderts zurück. Bereits ZIPF (1932, 1935) konnte diesen Zusammenhang auf Grund von empirischen Untersuchungen zur Wortlänge in Texten des Chinesischen, Lateinischen und Englischen nachweisen:

> „In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency; and in view of the influence of high frequency on the shortenings from truncations and from durable and temporary abbreviatory substitution, it seems a plausible deduction that, as the relative frequency of a word increases, it tends to diminish in magnitude."(ZIPF 1935, 38)

Seit den Arbeiten von ZIPF gilt der Zusammenhang zwischen Länge und Frequenz als allgemein abgesicherte und bekannte linguistische Hypothese. In Anlehnung an diese Hypothese, die besagt, dass die Häufigkeit von Wörtern invers zu Länge korreliert, wurde eine Reihe von empirischen Untersuchungen auf der Grundlage von Texten und Korpora durchgeführt. KÖHLER (1986, 69) benutzt die Länge-Frequenz-Beziehung zur „Erklärung der individuellen Ausprägung der Länge." Er geht von der Hypothese aus, dass die relative Veränderung der Länge beim Übergang von einer Frequenz zu einer anderen proportional zur relativen Frequenzänderung ist:

> „Es ist offensichtlich, daß die Länge lexikalischer Einheiten von ihrer Frequenz etwa in der Weise abhängt, wie Zipf es darstellte. Als z.B. das Wort 'Automobil' in die Lexik mehrerer Sprachen aufgenommen wurde, war es noch sehr selten (…). Mit zunehmender Häufigkeit (…) wurde der Ausdruck in allen Sprachen stark gekürzt: dt. 'Auto' (…)." (KÖHLER 1986, 69)

KÖHLER überprüft diesen Zusammenhang an deutschem Sprachmaterial. Er passt an die empirischen Daten die Funktion

$$y = ax^b$$

an. Den Parameter *a* interpretiert er als die globale Komponente und die Konstante *b* als Kürzungseinheit. Die Untersuchung bestätigt unter Verwendung des F-Testes die Abhängigkeit der Länge von der Frequenz.

Untersuchungsgegenstand der Arbeit von HAMMERL (1990) ist u.a. auch die Überprüfung der Abhängigkeit der Länge von der Frequenz. HAMMERL untersucht zu diesem Zweck das Vokabular aus zwei polnischen Häufigkeitslisten. Die erste Liste beruht auf 50 Gedichten von

SZARZYŃSKI und hat einen Gesamtumfang von 5653 Tokens und 1703 Types. Die zweite Liste betrifft das polnische Häufigkeitswörterbuch von SŁOWNICTWO (1974-1977) und umfasst 30041 Lexeme. An die aus den beiden Häufigkeitslisten gewonnenen empirischen Daten passt HAMMERL das Modell von KÖHLER (siehe weiter oben) an. Die Überprüfung der Güte der Anpassung unter Anwendung des Determinationskoeffizienten ergibt, dass das entsprechende Modell die empirischen Daten signifikant beschreibt. Insgesamt liegt R^2 im Intervall $0.82 \geq R^2 \geq 0.58$.

In HAMMERL (1991) wird die Beziehung zwischen Länge und Frequenz sowohl als Abhängigkeit der Länge von der Frequenz als auch umgekehrt untersucht. Als Datengrundlage für die Überprüfung dieser Zusammenhänge und aller weiteren in seiner Arbeit aufgestellten Hypothesen benutzt HAMMERL (1991, 49) „die von J. Sambor und R. Hammerl für das Projekt 'Sprachliche Synergetik' vorbereiteten polnischen Daten (…), die 30041 Lexeme des polnischen Häufigkeitswörterbuches (SŁOWNICTWO 1974-1977) betreffen (…).” HAMMERL passt andere Funktionen als KÖHLER an.

Auch in der Untersuchung von STRAUSS, GRZYBEK & ALTMANN (2006) geht es um die Überprüfung der Abhängigkeit der Wortlänge von der Worthäufigkeit. Als Datengrundlage verwenden die Autoren 11 Texte unterschiedlicher Stile, die aus dem Russischen, Kroatischen, Slowakischen, Slowenischen, Deutschen, Englischen, Indonesischen, Sudanesischen und Ungarischen stammen. Die Texte werden einzeln und in ihrem Gesamtumfang analysiert. An die Dateien passen die Autoren ein anderes mathematisches Modell an als KÖHLER (1986) und HAMMERL (1990), nämlich

$$y = ax^{-b}+1$$

Für die Überprüfung der Güte der Anpassung der nach diesem Modell errechneten Daten an die entsprechenden empirischen Daten verwenden STRAUSS, GRZYBEK & ALTMANN den Determinationskoeffizienten. Es zeigt sich, dass die entsprechende Funktion die empirischen Daten hervorragend beschreibt. Die untersuchte Abhängigkeit konnte sich in jedem Text bestätigen. Insgesamt liegt R^2 im Intervall $0.96 \geq R^2 \geq 0.84$.

Weitere Untersuchungen zum Zusammenhang zwischen Wortlänge und Wortfrequenz findet man u.a. auch in NEWMAN & FRIEDMAN (1958), ARAPOV (1988) und GRZYBEK & ALTMANN (2002).

2.2.2.3. Zusammenhang zwischen Polylexie und Frequenz

Die Entdeckung der Beziehung zwischen Worthäufigkeit und Bedeutungszahl geht auf ZIPF (1946, 1972) zurück. Seiner Argumentation zufolge ist die Zahl der Bedeutungen des Wortes

mit einem bestimmten Häufigkeitsrang gleich der Wurzel aus seiner Häufigkeit (Vgl. ZIPF 1972, 27-28). Ausführliche Untersuchungen wurden von GUITER (1974) an Daten verschiedener romanischer Sprachen (Französisch, Spanisch, Italienisch, Portugiesisch, etc.) sowie dem Englischen unternommen, der den Trend mit der Formel

$$S = CF^{1/a}$$

erfasste, wobei *S* die Bedeutungszahl, *F* die Häufigkeit und *C* eine Konstante ist. a > 2.
Die Arbeiten von ZIPF und GUITER enthalten nur Graphen, keine Zahlen. Es fehlt auch − eine plausible − linguistische Begründung für die untersuchte Wechselbeziehung. Eine solche Begründung besteht erst seit den Arbeiten von ALTMANN(1980) und ALTMANN & SCHWIBBE (1989). In diesen Studien wird die Beziehung zwischen Bedeutungszahl und Frequenz aus dem Menzerath-Gesetz hergeleitet. In ALTMANN & SCHWIBBE (1989) wird dieser Zusammenhang an Daten des Slowakischen und Russischen überprüft, jedoch nur mit einem schwach signifikanten Ergebnis für das Russische. Angepasst wird dabei die Funktion

$$y = ax^b$$

In KÖHLER (1986) wird jedoch die Beziehung zwischen Bedeutungszahl und Worthäufigkeit indirekt aus dem Zusammenhang zwischen Länge und Frequenz und dem zwischen Polylexie und Länge abgeleitet. Das Verhältnis zwischen diesen beiden Größen wird nach KÖHLER (1986) stark durch das Anwendungsbedürfnis beeinflusst. Polyseme Wörter werden häufiger verwendet als nicht- polylexe Wörter. Und je öfter eine Einheit zum Ausdruck einer beliebigen Bedeutung verwendet wird, desto größer ist deren Frequenz. D.h.: Die Frequenz hängt davon ab, „wie oft eine der Bedeutungen, die die Einheit tragen kann, kommunikationsrelevant wird“ (KÖHLER, 1986, 67). Die Bedürfnisse nach Minimierung des Dekodierungsaufwands (minK) und des Kodierungsaufwands (minD), welche als die Zipf'schen Kräfte der Unifikation und Diversifikation verstanden werden, wirken sich auf die Polylexie aus. Das Bedürfnis minK wirkt sich positiv auf die Polylexie aus, während sein Pendant die Polylexie negativ beeinflusst.
Die Abhängigkeit der Polylexie von der Worthäufigkeit überprüft KÖHLER an einer Stichprobe zur deutschen Sprache und erzielt eine hohe Anpassungsgüte. Er passt dabei die Funktion

$$y = ax^b$$

an. Diese Funktion gewinnt er dadurch, dass er die Abhängigkeit der Polylexie von der Länge und die Abhängigkeit der Länge von der Frequenz miteinander verknüpft.

Im Hinblick auf andere lexikalische Eigenschaften wurden im Bereich der quantitativen Linguistik u.a. auch die Zusammenhänge der Polylexie mit der Polytextie (vgl. z.B. KÖHLER

1986), den Komposita (vgl. z.B. ROTHE 1988) und der Wortbildungsaktivität (vgl. z.B. STEINER 2002) untersucht.

2.2.2.4. Zusammenhang zwischen Satz- und Teilsatzlänge

Die Erkenntnis, dass zwischen der Größe eines sprachlichen Konstrukts und der Größe seiner Konstituenten eine Beziehung besteht, verdankt man in erster Linie den Arbeiten des Psychologen und Phonetikers MENZERATH. Von wesentlicher Bedeutung ist seine 1954 veröffentlichte Arbeit „Architektonik des deutschen Wortschatzes". In dieser Publikation untersucht er den Zusammenhang zwischen der Wortlänge, gemessen in der Anzahl der Silben, und der Silbenlänge, bestimmt durch die Anzahl der Laute. Zu diesem Zweck wertet er 20453 Stichwörter eines deutschen Aussprachewörterbuches aus. MENZERATH (1954, 108) konnte zeigen, dass „die Zentralwerte der Lautzahlen fast gleichmäßig mit steigender Silbenzahl fortschreiten." Er schließt daraus auf eine Beziehung zwischen der Anzahl der Silben und der Anzahl der Laute. Diesen Zusammenhang gibt er wie folgt wieder:

> „Die relative Lautzahl nimmt mit steigender Silbenzahl ab, oder mit anderer Formel gesagt: je mehr Silben ein Wort hat, um so (relativ) kürzer (lautärmer) ist es." MENZERATH (1954, 100)

Die Ergebnisse zu dem beobachteten Phänomen fasst MENZERATH (1954, 101) in der allgemeinen Hypothese zusammen: „*Je größer das Ganze, um so kleiner die Teile*". Diese Hypothese wurde von ALTMANN (1980, 1) auf den Bereich der Linguistik übertragen und wie folgt formuliert:

„*The longer a language construct the shorter its components (constituents)*",

woraus sich als Konsequenz eine Reihe von Hypothesen ableiten lässt (vgl. ALTMANN & SCHWIBBE 1989, 5).

Ausgehend von der Annahme, dass die Konstituentenlänge von der Konstruktlänge abhänge und dass ihre relative Veränderungsrate umgekehrt proportional zu der Konstruktlänge stehe, leitete ALTMANN (1980) ein Gesetz für den von MENZERATH vorhergesagten Zusammenhang ab. Dieses in der einschlägigen Literatur als Menzerath-Gesetz bzw. Menzerath-Altmann-Gesetz bekannte Gesetz lautet in seiner allgemeinsten Form:

$$y = ax^b e^{-cx} \tag{8}$$

a, b und c sind dabei Koeffizienten und lassen sich aus den Daten schätzen.

Die abgeleitete theoretische Funktion weist zwei Spezialfälle auf:

Für den Fall, dass $b = 0$, $c \neq 0$ gilt die Funktion:

(9) $y = a\, e^{-cx}$

Für den Fall, dass b ≠ 0, c = 0 gilt die Funktionsgleichung (10), die als Standardfall angesehen wird.

(10) $y = ax^b$

Seit der Ableitung des Menzerath-Gesetzes ist eine Vielzahl von Arbeiten zur Überprüfung seiner Validität durchgeführt worden; viele aus dem allgemeinen Satz ableitbare Hypothesen für verschiedene sprachliche Ebenen wurden entsprechenden empirischen Daten verschiedener Sprachen gegenübergestellt. Dargestellt werden in diesem Abschnitt ausschließlich Arbeiten zur funktionalen Beziehung zwischen Satz- und Teilsatzlänge, d.h. der Hypothese:
„In long sentences (measured in number of clauses) the clauses are shorter and vice versa otherwise the sentence loses it clearness." (ALTMANN 1980, 9)

Die funktionale Beziehung zwischen Satz- und Teilsatzlänge ist empirisch untersucht worden. Die Validität dieses Zusammenhangs überprüft KÖHLER (1982) an 843 Sätzen, die aus der Bearbeitung von englischen Texten gewonnen wurden. Die Sätze werden hinsichtlich ihrer Länge in der Anzahl der Teilsätze und die Teilsätze hinsichtlich ihrer Länge in der durchschnittlichen Wortanzahl gemessen. Als Kriterium eines Teilsatzes wird zunächst das Vorhandensein eines finiten Verbs festgelegt.
An die gewonnenen empirischen Daten für die Satzlänge und die durchschnittliche Teilsatzlänge passt KÖHLER das Menzerath-Gesetz in seinen drei Ausprägungen an. Die dabei erzielten Ergebnisse unterstützten die Hypothese, dass zwischen der Satzlänge und der Teilsatzlänge der stochastische Zusammenhang „je länger der Satz, desto kürzer die Teilsätze" besteht.
Um zu zeigen, dass die von ihm ermittelten Ergebnisse unabhängig von der Vorgehensweise zur Ermittlung von Konstrukt und Konstituenten zu sehen sind, wertet KÖHLER die gleichen Daten noch einmal mit anderen Untersuchungskriterien aus: Als Teilsatzindikatoren betrachtet er neben den finiten Verben zusätzlich noch Partizipien mit Objekt oder Präpositionalphrase. Eine Anpassung der drei Ausprägungen des Menzerath-Gesetzes an die Daten ergibt ebenfalls hoch signifikante Ergebnisse. KÖHLER schließt daraus Folgendes:

> „Besteht ein gesetzmäßiger Zusammenhang zwischen zwei oder mehreren Größen, so gilt er natürlich unabhängig davon, in welchen Einheiten diese Größen gemessen werden. Auch linguistische Einheiten wie Satz, Teilsatz, Wort, Morphem usw. sind konventionell festgelegt bzw. festlegbar." (KÖHLER 1982, 106)

Mit der Überprüfung des Zusammenhangs zwischen Satzlänge und Teilsatzlänge befasst sich auch HEUPS (1983). Die Datengrundlage ihrer Untersuchung besteht aus 10668 Sätzen, die

aus 5 deutschen Textklassen (Gesetzestexten, wissenschaftlichen Texten, Zeitungstexten, Romanen und Briefen) gewonnen wurden. Somit wird ein breites Spektrum abgedeckt. Die Texte werden nach ähnlichen Methoden ausgewertet, wie sie auch KÖHLER in seiner Arbeit anwendet.

HEUPS' Arbeit enthält neben einer Analyse der Textklassen auch Einzelanalysen der Texte. Damit beabsichtigt sie, Kennwerte zu ermitteln, die Anhaltspunkte für Stil und Textsorte sein könnten:

> „Die Variablen Satzlänge, Clauselänge und Satztypenhäufigkeit dieser Arbeit können für die Kategorisierung von Texten und Stilen als erste Anhaltspunkte nützlich sein, stellen aber lediglich die Behandlung eines Einzelproblems dar." (HEUSPS 1983, 122)

An die empirischen Werte passt HEUPS die Funktionsgleichung (8) an. Die erzielten Ergebnisse bestätigen, dass entsprechend dem Menzerath-Gesetz mit zunehmender Satzlänge die Teilsatzlänge abnimmt; der Vergleich der empirischen und der berechneten Werte zeigt eine signifikante Übereinstimmung.

Bei der Arbeit von TEUPENHAYN & ALTMANN (1984) geht es ähnlich wie bei KÖHLER und HEUPS um die Überprüfung der funktionalen Beziehung zwischen Satzlänge und Clauselänge. Die Autoren selbst formulieren dies folgendermaßen:

> „In the present article we want to test the dependence of clause length on sentence length on a somewhat larger scale." (S.127)

Die Satzlänge wird hier wiederum in Teilsätzen bestimmt und die Teilsatzlänge durch die Anzahl der im Satz vorkommenden finiten Verben festgelegt. Die Datengrundlage der Untersuchung bilden Sätze aus 29 Texten verschiedener Sprachen (Deutsch, Englisch, Französisch[13], Schwedisch, Ungarisch, Slowakisch, Indonesisch und Tschechisch). Pro Satz werden mindestens 200 Sätze bearbeitet. Die Anpassung des Menzerath-Gesetzes in der einfachsten Form führt zu guten Ergebnissen: Nur in einem Text kann keine gute Übereinstimmung der empirischen Werte mit den theoretischen Daten festgestellt werden.

Die Satz-Teilsatz-Beziehung im Chinesischen untersucht BOHN (1998) an Korpus- und Textdaten. Satz- und Teilsatzlänge werden nach ähnlichen Kriterien gemessen, wie sie auch TEUPENHAYN & ALTMANN (1984) in ihrer Arbeit anwenden. Die Anpassung der Funktion

[13] Die Autoren haben nur einen Text ausgewertet, nämlich: PEYREFITTE, A. (1976): Le Mal Français. Plon.

$y = ax^b$ an die empirisch ermittelten Daten führt bei der Korpus-Untersuchung zu besseren Ergebnissen als bei der Einzeltext-Untersuchung[14].

Neben der Frage nach dem Verhältnis zwischen Wortlänge und Bedeutungshaltigkeit geht SCHWIBBE (1984) in seiner Arbeit auch dem Verhältnis von Satzlänge zu Clauselänge nach. Anders als bei den bereits vorgestellten Arbeiten führt er seine Untersuchung nicht auf der Satzebene, sondern auf der Textebene durch. Er bearbeitet hierzu 744 Einzeltexte verschiedener Sorten. Für jeden Text werden die durchschnittliche Satzlänge (in Clauses) und die Clauselänge (in Wörtern) ermittelt. Eine Clause wird dabei definiert als „eine Menge von Wörtern, die (a) größer als 2 und (b) zwischen zwei Satzzeichen steht." (SCHWIBBE 1984, 157)
SCHWIBBE passt das Exponentialmodell $y = a\,e^{-cx}$ und das Potenzmodell $y = ax^b$ an die Werte der abhängigen Variable an. Wie bei den vorhergehenden Arbeiten werden auch hier trotz unterschiedlicher Methoden der Clause-Berechnung gute Ergebnisse erzielt:

> „Sowohl auf der Textebene als auch auf der Wortebene ist der Nachweis der Gültigkeit der MR [Menzerathschen Regel] erbracht worden: je höher die Satzlänge, umso geringer ist die Clauselänge." (SCHWIBBE 1984, 161)

In allen vorgestellten Arbeiten gilt, dass die Untersuchungsergebnisse signifikant sind und daher das Menzerath-Gesetz durch eine Vielzahl von Daten aus den unterschiedlichsten Sprachen bestätigt wird.

2.2.2.5. Zusammenhang zwischen Frequenz und Komplexität von Teilsätzen

Quantitative Untersuchungen zu diesem Zusammenhang konnten nicht gefunden werden. Es liegen aber etliche Arbeiten zur Beziehung zwischen Eigenschaften anderer syntaktischer Einheiten vor, beispielsweise in KÖHLER (1999, 2005).

KÖHLER (1999) befasst sich mit Zusammenhängen zwischen Eigenschaften von syntaktischen Konstruktionen, wobei er die Konstruktionen auf der Basis der Phrasenstrukturgrammatik als Konstituenten operationalisiert. KÖHLER begründet und überprüft u.a. folgende Untersuchungshypothesen: (a) Die mittlere Frequenz von Konstituenten ist eine Funktion deren Komplexität, (b) die mittlere Komplexität von Konstituenten ist eine Funktion deren Frequenz, (c) die Verschachtelungstiefe von Konstituenten ist eine Funktion deren Position. Die empirische Überprüfung dieser und anderer Zusammenhänge erfolgt an deutschen und englischen Texten. An die Dateien wird eine Anpassung der Funktion

$$y = ax^b e^{cx}$$

[14] Der Determinationskoeffizient liegt bei der Korpus- und Einzeltext-Untersuchung jeweils bei D = 0.9750 und D = 0.5628.

mit guten Ergebnissen durchgeführt.

KÖHLER (2005,18ff.) untersucht die funktionale Abhängigkeit der Selektionsbeschränkungen von der Anzahl der Aktanten. Er zeigt, dass die Anzahl der Alternativen linear mit der Anzahl der überhaupt möglichen Aktanten steigt. Für diesen Zusammenhang erweist sich die Funktionsgleichung

$$y = ax^b$$

als geeignet.

3. Theoretische Modelle für die einzelnen Verteilungen und Zusammenhänge

Im Folgenden wird darauf eingegangen, welche theoretischen Modelle verwendet werden, um die einzelnen Verteilungen und funktionalen Zusammenhänge zu erfassen. Es wird hier eine direkte Anwendung der Statistik betrieben. D.h.: Es werden von der Statistik bereitgestellte Modelle ohne jegliche Veränderungen und Innovation verwendet. Von der Ableitung von neuen Modellen ist hier also keine Rede, zumal

> „the modelling of dynamical language phenomena as well as the formulation of language and text laws meet with the difficulty of finding appropriate concepts for describing the dependence of an entity on another one." (ALTMANN & KÖHLER 1995, 62)

Zudem erfordert die Ableitung von neuen Modellen eine einigermaßen aktive Beherrschung der Statistik, wenn man Erfolge erzielen will (vgl. ALTMANN & LEHFELDT1980, 23).

Die zu verwendeten Modelle werden auf der einen Seite deduktiv, auf der anderen Seite aus bisherigen Befunden abgeleitet. Bei der letzteren Herangehensweise wird von allgemeinen, sprachübergreifenden Gültigkeiten ausgegangen.

3.1. Die Häufigkeitsverteilungen

3.1.1. Häufigkeitsverteilung der Bedeutungen von Wortbildungsaffixen

Der im Abschnitt 2.2.1.1. gelieferte Überblick über Untersuchungen zur Diversifikation der Bedeutungen von Affixen zeigt ein wenig homogenes Bild; der Diversifikationsprozess wird hier mit vielen Verteilungen getestet. Es lässt sich aber feststellen, dass zwei Modelle immer wieder mit guten Ergebnissen verwendet werden. Es handelt sich zum einen um die negative Binomialverteilung, die wegen der Rangordnung gewöhnlich in der 1-verschobenen Form verwendet wird, wie sie in (11) gegeben ist:

$$(11) \qquad P_x = \binom{k+x-1}{x} \frac{p^k q^x}{1-p^k}, \quad x = 1, 2, \cdots n$$

Zum anderen handelt es sich um die 0-gestutzte bzw. positive negative Binomialverteilung mit der Formel:

$$(12) \qquad P_x = \frac{\binom{k+x-1}{x}}{\left(1-p^k\right)} p^k q^x \ , x = 1, 2, \cdots n$$

Beide Verteilungsmodelle wurden auch immer wieder verwendet, um andere Arten von Diversifikation zu modellieren. Sie wurden beispielsweise von BEST (1991) für die Erfassung der Diversifikation des deutschen Partikels *"von"* herangezogen. Die Anpassungsergebnisse sind mit Wahrscheinlichkeiten von fast 1 für beide Verteilungen sehr zufrieden stellend. FUCHS (1991) erzielte auch signifikante Resultate bei der Durchführung der Anpassung der beiden Verteilungen an die Diversifikation der deutschen Präposition *"auf"*. HENNERN (1991) überprüfte, ob beide Verteilungsmodelle auf die semantische Diversifikation von *"in"* im Englischen anwendbar sind. Die von ihr erzielten Ergebnisse liegen für beide Modelle in einem Bereich, den auch BEST (1991) für seine Untersuchung erhielt.

Da die negative Binomialverteilung und die 0-gestutzte negative Binomialverteilung sich in den bisherigen Untersuchungen zur semantischen Diversifikation an verschieden Sprachen mit guten Ergebnissen bewährt haben, sollen auch diese Verteilungen für die Modellierung der semantischen Diversifikation der Bedeutungen von Affixen im Französischen herangezogen werden. Daher die folgende Hypothese:

> „*Die Häufigkeit der Bedeutungen von Wortbildungsaffixen im Französischen folgt in ihrer Verteilung dem Modell der 1-verschobenen negativen Binomialverteilung bzw. 0-gestutzten (positiven) negativen Binomialverteilung.*"

3.1.2. Ranghäufigkeitsverteilung von Wörtern

Auf Grund der Arbeiten zu den Worthäufigkeitsverteilungen in Texten kann feststellt werden, dass sich das Zipf-Mandelbrot-Gesetz für die Beschreibung dieser Sprachphänomene gut eignet; dieses Modell hat sich in den meisten Untersuchungen bewährt; der Großteil der Untersuchungsergebnisse hat dessen Adäquatheit bestätigt. Daraus ist nach ORLOV (1982, 143) zu schließen, dass ein Autor

> „im Laufe der Schöpfung der Texte imstande ist, ihre Häufigkeitsstruktur so zu organisieren, dass die volle Länge dem Zipfschen Umfang nahe kommt."

ALTMANN & ZÖRNIG (1983, 205) schreiben dem Modell einen Universalitätscharakter zu:

„In linguistics it is generally accepted that the ranked frequencies of linguistic entities are distributed according to the Zipf-Mandelbrot law."[15]

Da die meisten empirischen Ranghäufigkeitsverteilungen sich (sehr) gut mit der Zipf-Mandelbrot-Verteilung modellieren lassen, soll an die Ranghäufigkeitsverteilung von Wörtern in französischen Texten auch dieses Modell angepasst werden. Daher die Hypothese:

„Die Ranghäufigkeitsverteilung von Wörtern im Französischen folgt dem Modell der Zipf-Mandelbrot-Verteilung."

3.1.3. Ranghäufigkeitsverteilung von Buchstaben und Phonemen

Für die Erfassung der Häufigkeitsverteilung von Buchstaben und anderen "niedrigsten" Spracheinheiten wie Phonemen sind verschiedene Modelle diskutiert worden, wobei die Diskussion keine einheitliche Festlegung auf ein Verteilungsmodell erbracht hat. In der einschlägigen Literatur tauchen immer wieder folgende Modelle auf:

- Zipf-Mandelbrot-Verteilung
- Zeta-Verteilung
- Geometrische Verteilung
- Good-Verteilung
- Negative hypergeometrische Verteilung

Für die mathematische Herleitung dieser Verteilungen sei hier auf die einschlägige Literatur hingewiesen[16].

Aus den bisher durchgeführten Untersuchungen zur Modellierung der Häufigkeitsverteilung von Buchstaben lässt sich als vorläufiges Ergebnis resümieren: Bei den Arbeiten zu den slawischen Sprachen erwies sich die negative hypergeometrische Verteilung als das beste Modell. Für die Verteilung der Buchstaben im Deutschen kommen drei Modelle in Frage, nämlich die Zipf-Mandelbrot-Verteilung, die geometrische Verteilung und die negative hypergeometrische Verteilung.

Damit stellt sich die Frage, welche Verteilung für die Erfassung der Buchstabenverteilungen im Französischen verwendet werden kann. Um dies zu klären, sollen alle weiter oben aufgeführten Modelle auf ihre Adäquatheit hin geprüft werden. Auch für die Erfassung der Phonemhäufigkeit sollen diese Modelle herangezogen werden. Damit verbunden ist die Frage nach einem einheitlichen Modell, dem verschiedene Stichproben bzw. Einzeltexte unterliegen.

[15] Vgl. auch: BEST (2001), ZÖRNIG & ALTMANN (1995), ALTMANN (1988a, 73ff.).
[16] Vgl. z.B. GRZYBEK et al. (2004, 25).

3.1.4. Häufigkeitsverteilung der Satzlängen

Untersucht wird hier nur die Clause/Satz-Verteilung. Die bisherigen empirischen Untersuchungen hierzu zeigen, dass die 0-gestutzte negative Binomialverteilung und die Hyperpoisson-Verteilung geeignete Modelle sind. Beide Verteilungen haben sich bei vielen Sprachen – und unterschiedlichen Spracheinheiten wie z.B. der Wortlängenverteilung – bewährt. Daher sollen sie auch in der vorliegenden Arbeit herangezogen werden. Da keine Sätze der Länge 0 definieret wird, wird die Hyperpoisson-Verteilung in der 1-verschobenen Form verwendet, deren Formel lautet:

$$(14) \qquad P_x = \frac{a^{x-1}}{b^{(x-1)}\, {}_1F_1(1;b;a)},\ x = 1, 2, \cdots n$$

Neben der positiven negativen Binomialverteilung und der 1-verschobenen Hperpoisson-Verteilung soll auch die (1-verschobene) negative Binomialverteilung in Betracht gezogen werden. Es handelt sich hierbei um dasjenige Modell, das von ALTMANN (1988b) für die Clause/Satz-Verteilung postuliert und erfolgreich an 10 Texten unterschiedlicher Sprachen überprüft wurde.

3.1.5. Häufigkeitsverteilung in Textblöcken

Aus der einschlägigen Literatur geht hervor, dass das Textblock-Gesetz eine Vielfalt von Einheiten verschiedener Sprachebenen zwischen Lauten und Syntagmen steuert. Die bisherigen Ergebnisse zeigen, dass die Häufigkeitsverteilungen dieser Einheiten sich am besten mit der negativen hypergeometrischen Verteilung erfassen lassen, deren Formel lautet:

$$(13) \qquad \binom{-M}{x} \frac{\binom{-k+m}{n-x}}{\binom{-k}{n}},\ x = 0, 1, 2. \cdots n$$

$$K > M > 0$$

Dieses Verteilungsmodell hat sich in den meisten Untersuchungen an verschiedenen Sprachen als (sehr) gutes erwiesen. Man darf festhalten:

> „The distribution of the text-blocks containing x words (or other text units) is derived in the form of the negative hypergeometric (Beta-binomial) distribution."(ALTMANN & BURDINSKI 1982, 147)

In Übereinstimmung mit ALTMANN & BURDINSKI (1982) und früheren Untersuchungen soll in der vorliegenden Arbeit auch eine Anpassung der negativen hypergeometrischen Vertei-

lung an die Häufigkeitsverteilung sprachlicher Einheiten in Textblöcken französischer Sprache durchgeführt werden. Daher die Hypothese:

„*Im Französischen folgt die Häufigkeitsverteilung in Textpassagen dem Modell der negativen hypergeometrischen Verteilung*."

3.1.6. Häufigkeitsverteilung der Komplexität und Frequenz von Teilsätzen

Die Bezugseinheit, mit der hier operiert wird, ist der Teilsatz, der auf der Grundlage der Dependenzgrammatik bearbeitet wird. Die Bearbeitung bezieht sich dabei ausschließlich auf die dem Verb direkt untergeordneten Einheiten[17].

Für die Modellierung der Häufigkeitsverteilungen der Eigenschaften *Komplexität* und *Frequenz* von Teilsätzen werden die quantitativen Analysen zur Syntax von KÖHLER (1999) und KÖHLER & ALTMANN (2000) verwendet, welche auf einem phrasenstrukturgrammatischen Ansatz basieren. Die Komplexität einer Konstituente definieren KÖHLER & ALTMANN (2000) als die Anzahl der unmittelbaren Konstituenten der betroffenen Konstituenten. Den Autoren zufolge wird diese Eigenschaft durch folgende Faktoren gesteuert:

(a) Bedürfnis nach Minimierung der Komplexität der syntaktischen Konstruktion. Dadurch wird der mit der Produktion einer Äußerung eines gegebenen Ausdrucks verbundene physische Aufwand reduziert.

(b) Bedürfnis nach Maximierung der Kompaktheit. Dadurch wird die Komplexität reduziert.

(c) Informationsgehalt

(d) Inventar der syntaktischen Konstruktionen. Je mehr verschiedene Konstruktionen es gibt, desto weniger komplex sind sie.

Aus diesen Faktoren leiten sie die Hyperpascal-Verteilung als geeignetes Modell für die Verteilung der Komplexität von Konstituenten ab. Die Formel dieses Verteilungsmodells ist in (15) gegeben:

$$(15) \quad P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0, x = 0,1,2,....n$$

[17] Näheres dazu vgl. Abschnitt 5.6.

In Anlehnung an KÖHLER & ALTMANN (2000) wird hier angenommen, dass auch die weiter oben genannten Faktoren die Häufigkeitsverteilung der Komplexität von Teilsätzen steuern. An die empirischen Daten soll deshalb auch die Hyperpascal-Verteilung angepasst werden. Da Teilsätze mit der Komplexität 0 nicht definiert werden, soll dieses Verteilungsmodell in der 1-verschobenen Form verwendet werden. Daher die Hypothese:

„Die Komplexität von Teilsätzen folgt in deren Verteilung der 1-vrechobenen Hyperpascal-Verteilung.“

Es wird angenommen, dass die Frequenz eines Teilsatzes von dem Bedürfnis nach Minimierung des Produktionsaufwands beeinflusst wird, analog zu dessen Einfluss in der Lexik, wo es die Länge betrifft. Im Falle von Teilsätzen ist der Produktionsaufwand durch die Anzahl der obligatorischen und fakultativen Aktanten des Verbs bestimmbar. Der mit der Artikulation verbundene Aufwand wird maximal reduziert, wenn Teilsätze mit der Komplexität 1 am häufigsten auftreten. Es wird also von einer gesetzmäßigen Häufigkeitsverteilung der Frequenzklassen und einer entsprechenden Ranghäufigkeitsverteilung, die dem Zipf-Mandelbrot-Gesetz folgt, ausgegangen. Das Frequenzspektrum soll genauso wie bei KÖHLER (1999) an die Waring-Verteilung angepasst werden. Diese Verteilung wird in der Linguistik in der 1-verschobenen Form benutzt, die geschrieben wird als

$$(16) \qquad f(x) = \frac{b}{(b+n)} \frac{n^{(x-1)}}{(b+n+1)^{x-1}} \;, x = 1,2,...n$$

3.2. Zusammenhänge zwischen sprachlichen Eigenschaften

3.2.1. Zusammenhang zwischen den Eigenschaften Polylexie, Länge und Frequenz

Bei der Überprüfung dieser Zusammenhänge soll analog zu KÖHLER die Funktionsgleichung

$$y = ax^b$$

angewendet werden, wobei y die abhängige Variable, x die unabhängige Variable, a und b Parameter sind.

3.2.2. Zusammenhang zwischen Satz- und Teilsatzlänge

Es wurden Arbeiten dargestellt, in denen die Hypothese

„Je länger ein Satz, desto kürzer die Teilsätze”

empirisch überprüft und bestätigt wird. Als Anpassungsmodelle werden dabei sowohl die allgemeine Funktion $y = ax^b e^{-cx}$ als auch die Sonderfälle $y = a\, e^{-cx}$ und $y = ax^b$ verwendet.

Für die Überprüfung dieser Hypothese in der vorliegenden Arbeit soll alle drei Formen des Menzerath-Altmann-Gesetzes angewendet werden.

3.2.3. Zusammenhang zwischen Frequenz und Komplexität von Teilsätzen

Der Zusammenhang zwischen Länge und Frequenz im Bereich der Lexik ergibt sich vor allem aus der Kürzung häufiger Wörter sowie daraus, dass kürzere Wörter längeren vorgezogen werden. Da es im Falle der Dependenzgrammatik um die Zwischenspeicherung von Knoten im Stemma und nicht um die der Wörter geht, wird anstelle der Länge die Komplexität angesetzt. Ähnlich wie im Falle lexikalischer Einheiten wird hier ein mit der Komplexität verbundener Kürzungseffekt auf die Verwendungshäufigkeit der Teilsätze angenommen. Die Auftretenshäufigkeit der Konstruktionen sinkt mit der Zahl der obligatorischen und fakultativen Aktanten: Je mehr unmittelbare Dependentien ein (Teilsatz)Verb hat, desto weniger häufig wird die Konstruktion verwendet.

Als mathematische Form dieses Zusammenhangs wird, ähnlich wie bei KÖHLER (1999), das Modell

$$y = ax^b e^{cx}$$

verwendet.

4. Statistische Testverfahren

Hier sollen die statistischen Testmethoden vorgestellt werden, die bei der empirischen Überprüfung der Häufigkeitsverteilungen und funktionalen Zusammenhänge verwendet werden. Bevor diesen Verfahren nachgegangen wird, ist ein kurzes Referat über das Prinzip des statistischen Testens geboten.

4.1. Prinzip des statistischen Testens

Ein statistischer Test ist ein Entscheidungsverfahren, das über die Gültigkeit von a priori formulierten Hypothesen nach der Sammlung von empirischen Daten eine konkrete Entscheidung trifft.

Bei der Überprüfung von Hypothesen können zwei Arten von Fehlentscheidungen getroffen werden: (a) Ablehnung der Nullhypothese (H_0), obwohl diese richtig ist, (b) Annahme der Nullhypothese, obwohl sie falsch ist. Im ersten Fall spricht man von dem Fehler 1. Art. Dessen Wahrscheinlichkeit wird durch die Irrtumswahrscheinlichkeit bzw. das Signifikanzniveau α bestimmt. Im zweiten Fall spricht man von dem Fehler 2. Art. Die Wahrscheinlichkeit für einen solchen Fehler wird mit β abgekürzt.

Der Nullhypothese entgegengesetzt ist die Alternativhypothese (H_1), welche die zu prüfende Aussage darstellt. Sie dient dazu, die Nullhypothese als ungeeignet zu verwerfen. Als H_0 wird die Behauptung eines Nicht-Unterschieds bzw. eines Nicht-Zusammenhangs zwischen zwei oder mehreren Variablen angesehen.

Beim statistischen Testen wird α („Signifikanzniveau“) vorgegeben und durch die Konstruktion des Entscheidungsverfahrens „kontrolliert“. Dies gilt nicht für β.

Wenn sich anhand der Stichprobe ein Resultat ergibt, das bei Gültigkeit der aufgestellten Hypothese unwahrscheinlich ist, wird die Nullhypothese verworfen. Das Signifikanzniveau wird in diesem Falle überschritten. Wird beispielsweise die Forderung aufgestellt, dass bei Gültigkeit der Hypothesen eine Wahrscheinlichkeit von 95% oder 99% vorhanden sein muss, dann beläuft sich die Irrtumswahrscheinlichkeit auf 5% bzw. 1%.

Führt eine Prüfung mit einem vorgegebenen Signifikanzniveau zur Feststellung eines Zusammenhangs bzw. Unterschieds zwischen zwei oder mehreren Variablen, so wird die Nullhypothese abgelehnt und die Alternativhypothese akzeptiert.

Die Konstruktion eines statistischen Prüfverfahrens zur Entscheidung über die Beibehaltung oder Zurückweisung einer Hypothese umfasst folgende Schritte[18]:

a. Formulierung der Nullhypothese (H_0)
b. Formulierung der logisch entgegengesetzten Alternativhypothese (H_1)
c. Festlegung der Irrtumswahrscheinlichkeit (Alpha-Fehler)
d. Bestimmung einer geeigneten theoretischen Prüfverteilung. Aus der passenden Tabelle wird der für den Fall der Gültigkeit der Nullhypothese passende theoretische Vergleichswert abgelesen
e. Erhebung der Daten und Bestimmung der geeigneten Prüffunktion
f. Entscheidung über die Beibehaltung oder Zurückweisung der Nullhypothese.

4.2. Angewandte Testverfahren

Zur Überprüfung der funktionalen Zusammenhänge zwischen zwei Systemgrößen soll der *Determinationskoeffizient bzw. das Bestimmtheitsmaß* berechnet werden, während für die Überprüfung der Häufigkeitsverteilungen der *Chiquadrat-Test* herangezogen werden soll.

4.2.1. Der Chiquadrat-Test

Der Chiquadrat-Test (χ^2-Test) dient der Überprüfung von Verteilungshypothesen, d.h. Hypothesen, die unbekannte Verteilungen von Grundgesamtheiten betreffen. Bei der Überprüfung

[18] Näheres zu den einzelnen Schritten vgl. KRUG & REHM (2003, 99ff.).

solcher Hypothesen wird untersucht, ob die in einer Stichprobe beobachtete Verteilung (empirische Häufigkeiten) mit den Annahmen über die unbekannte Verteilung (theoretische Häufigkeiten) der Grundgesamtheit übereinstimmt oder nicht, d.h. es wird untersucht, ob die Unterschiede zwischen theoretischen und empirischen Häufigkeiten zufällig sind oder nicht. Die Güte der Anpassung der theoretischen Verteilung an die empirischen Daten wird letztendlich überprüft. Man spricht daher auch von einem so genannten Anpassungstest.
Der Chiquadrat-Test, χ2, lässt sich nach der folgenden Formel berechnen:

$$\chi^2 = \frac{\sum_{i=1}^{n} (f_i - NP_i)^2}{NP_i} \qquad (17)$$

f_i: Empirische Häufigkeiten der i-Klasse
NP_i: Erwartete Häufigkeiten der i-Klasse
n: Gesamtzahl der Messwerte
Wie Gleichung (17) zeigt, ist der Wert von χ^2 abhängig von den Unterschieden zwischen den theoretischen und empirischen Häufigkeiten und der Zahl der Messwerte.

Um über die Güte der Anpassung entscheiden zu können, wird der Testwert χ^2 mit einem Wert aus der Chiquadrat-Verteilung mit FG = n –1 Freiheitsgraden verglichen. Falls für ein gegebenes Signifikanzniveau α der Wert aus der Tabelle größer als die Prüfgröße χ^2 ist, könnte die Nullhypothese nicht abgelehnt werden, d.h. die Anpassung kann als gut angesehen werden. Müssen für die Bestimmung der erwarteten Häufigkeiten aus der vorliegenden Stichprobe zunächst Parameter – dieser Verteilung – geschätzt werden, so reduziert sich die Zahl der Freiheitsgrade: Liegen *n* Klassen vor und werden aus der Stichprobe *m* Parameter geschätzt, dann vermindert sich die Zahl der Freiheitsgrade auf FG = n – m – 1.

Der Chiquadrat-Test ist unter bestimmten Voraussetzungen anzuwenden: Die Stichprobe muss vollständig in disjunkte, nichtleere Klassen zerlegt werden. Die Klasseneinteilung ist so zu wählen, dass die Bedingung n –1 > 0 bzw. n – m – 1 > 0 erfüllt ist. D.h. die Zahl der Klassen darf nicht zu klein sein. Ist diese Bedingung nicht erfüllt, so können vor Einsatz des Testes Merkmalsklassen zu stärker besetzten Klassen zusammengefasst werden.

Für die Berechnung des Chiquadrates wird in der vorliegenden Arbeit die Software *Altmann-Fitter* (2000) eingesetzt. Mit dieser Software lassen sich über 100 verschiedene Verteilungen an beliebige Dateien anpassen und prüfen. Die Software liefert als Ergebnis Informationen über u.a. zwei Testwerte: zum einen über ein mit *p* bezeichnetes maximales α, das aus dem berechneten χ^2 –Wert resultiert. Dieses Signifikanzniveau gilt als Gütekriterium bei einer Anpassung unter Verwendung eines χ2-Anpassungstests.

Bei der Abschätzung der Güte der Anpassung sollen hier die in Tabelle 4.1 enthaltenen Maßstäben aus ALTMANN & HAMMERL (1989, 21) herangezogen werden:

Tabelle 4.1: Anpassungskriterien nach ALTMANN & HAMMERL (1989, 21)

Signifikanzniveau α	> 0.50	0.50 – 0.20	0.19 – 0.05	< 0.05
Anpassung	gut	mäßig	schwach	schlecht

Der Altmann-Fitter liefert auch Auskunft über den Kontingenzkoeffizienten (C), auch Diskrepanzkoeffizienten genannt. Er wird nach der Formel $C = \chi^2/N$ berechnet und dann zur Beurteilung der Anpassung herangezogen, wenn die Prüfgröße *P* auf Grund des Stichprobeumfangs N oder mangels Freiheitsgraden nicht berechnet werden kann bzw. konnte. Der Diskrepanzkoeffizient wird bei $C < 0.02$ als Indiz einer guten, bei $C < 0.01$ als Indiz einer sehr guten Anpassung betrachtet. In letzterem Fall ist davon auszugehen, dass die theoretische Berechnung geeignet ist, um die empirisch ermittelten Werte in dem gegebenen Modell zu erfassen. Die Anpassung ist umso besser, je kleiner C ist.

4.2.2. Der Determinationskoeffizient

Die in der vorliegenden Arbeit zu überprüfenden funktionalen Zusammenhänge wurden in Anlehnung an KÖHLER (1986) und KÖHLER (1999) mit dem mathematischen Modell

$$y = ax^b$$

bzw.

$$y = ax^b e^{cx}$$

wiedergegeben. Bei diesen Formeln handelt es sich um nicht-lineare Gleichungen, die aber in lineare Funktionen ungewandelt werden können, z.B. als

$$ln(y) = ln(a) + b*ln(x)$$

bzw.

$$ln(y) = a + b * ln(x) - cx$$

Für die Überprüfung der Zusammenhänge an Daten besteht somit die Möglichkeit, eine lineare oder nicht-lineare Regression durchzuführen.

Nach GROJAHN (1992) ist die Linearisierung in vielen Fällen „a very convenient method which yields a good approximation to the original nonlinear model as fitted with the help of more sophisticated methods.” In der vorliegenden Arbeit werden aber die theoretischen Funktionsgleichungen nur mittels nicht-linearer Regression ermittelt. Auf die Verwendung des F-Testes wird verzichtet. Dieser Test ist nur für lineare Modelle einsetzbar. Für linearisierte Daten ist er gültig, für das diesen Daten entsprechende nichtlineare Modell aber nur „mit einem be-

stimmten (aber unbekannten) Fehler“ (HAMMERL 1990, 8). Anstelle des F-Testes soll der Determinationskoeffizient berechnet werden.

Der Determinationskoeffizient, D bzw. R^2, ist ein Maß für die Abweichungen der Vorhersagen eines Regressionsmodells von den empirischen Daten. Er gibt Auskunft darüber, ob der gewählte Ansatz den vermuteten funktionalen Zusammenhang mehr oder weniger gut wiedergibt, sprich ein Maß für die Güte der Anpassung der nach der mathematischen Funktion errechneten Daten an die empirischen Daten.
Konkret drückt der Determinationskoeffizient den Anteil der aufgeklärten Varianz aus, d.h. der Variation der Modellvorhersage an der Gesamtvarianz.

Zur Berechnung von R^2 wird die Gesamtvarianz der empirischen Daten in zwei Bestandteile zerlegt: in erklärte und nicht-erklärte Varianz. Der Determinationskoeffizient ergibt sich aus der Division von erklärter Streuung durch die Gesamtstreuung:

$$(18) \qquad R^2 = D = \frac{ErklärteStreuung}{Gesamtstreuung} = \frac{\sum_{i=1}^{n}\left(y_i^* - \bar{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}$$

Eine alternative Methode ist die Subtraktion des Verhältnisses der nicht erklärten Streuung (Restvarianz) zur Gesamtstreuung vom Maximalwert 1:

$$(19) \qquad R^2 = D = \frac{\mathrm{Re}\,st\,\mathrm{var}\,ianz}{Gesamt\,\mathrm{var}\,ianz} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}$$

R^2-Werte liegen im Intervall [0, 1]. Je besser die Anpassung der theoretischen Werte y_i^* an die empirischen Daten y_i ist, desto kleiner ist die nicht-erklärte Varianz, und desto größer ist der R^2-Wert. $R^2 = 0$ bedeutet, dass die unabhängige Variable keine Prognose der Ausprägung der abhängigen Variablen erlaubt, d.h. dass das mathematische Modell für die Beschreibung der vermuteten Abhängigkeiten ungeeignet ist. $R^2 = 1$ weist auf eine Erklärung der gesamten Varianz hin; die Modellfunktion gibt den Trend bzw. die Ausprägung der abhängigen Variablen hervorragend wieder, was auf eine perfekte Modellanpassung hinweist.

In der vorliegenden Arbeit wird für die empirische Überprüfung der funktionalen Zusammenhänge, d.h. für die Schätzung der Parameter der einzelnen Funktionen, die Berechnung der theoretischen Werte und des Determinationskoeffizienten, das Computerprogramm NLREG (Nonlinear Regression Analysis) in der Version 6.3 verwendet. Die Ergebnisse werden jeweils in Form von Tabellen und Abbildungen dargestellt.

5. Empirische Überprüfung der Häufigkeitsverteilungen

Die Ergebnisse der empirischen Überprüfung werden im Folgenden nur zusammenfassend wiedergegeben. Detaillierte Resultate können dem Anhang entnommen werden.

5.1. Häufigkeitsverteilung der Bedeutungen von Wortbildungsaffixen

Wortbildungsaffixe bzw. -morpheme dienen dazu, neue Wörter zu bilden. Sie ergeben in Verbindung mit einem Wortstamm bzw. Grundmorphem neue Wörter. Zu diesen Morphemtypen werden Präfixe, Infixe und Suffixe gerechnet.

In der vorliegenden Untersuchung wurden nur Suffixe berücksichtigt, da es bei der Suffigierung möglich ist, dass die Wortart des Derivats und die des Wortstamms identisch sind, wie die Wörter *ski* und *skieur* zeigen. Des Weiteren kann die Suffigierung auch die Wortklasse ändern, sodass aus einem Substantiv beispielsweise durch Anfügung eines Suffixes ein Adjektiv wird, wie an den Wörtern *aliment* und *alimentaire* zu erkennen ist.

Suffixe lassen sich je nach der Wortart der Derivate in Nominal-, Verbal- und Adjektivsuffixe einteilen. In dieser Arbeit wurden nur Nominalisierungen berücksichtigt, da diese zu den häufigsten Ableitungen im Französischen gehören (vgl. LÜDTKE 1978, 68).

Suffixe werden im Französischen ebenso wie in anderen Sprachen mit verschiedenen Bedeutungen verwendet. Diese Bedeutungen zu erfassen, ist Ziel der in diesem Abschnitt vorgenommenen Untersuchung. Dazu wurden die im Text „*Madame Bovary*“ von G. Flaubert vertretenen Suffixe auf ihre semantische Diversifikation untersucht, d.h. ihre Bedeutungsvariationen wurden festgestellt und die Menge gezählt. Zur Einteilung der Bedeutungsvariationen von Suffixen wurden die Klassifikationen von BÉCHADE (1992) und THIELE (1981) herangezogen. Aus mathematischen Gründen wurden nur solche Suffixe berücksichtigt, für die mindestens vier verschiedene Bedeutungen im Text vertreten waren. Tabelle 5.1. enthält die berücksichtigten Suffixe und deren Bedeutungen.

Anpassung

An die Häufigkeiten der Bedeutungen der Suffixe wurde in Anlehnung an andere Untersuchungen zur Diversifikation die negative Binomialverteilung in der 1-verschobenen und der 0-gestutzten Form angepasst. Hierbei wurde in zwei Schritten verfahren:

(a) Zunächst wurden sämtliche Bedeutungen gemeinsam angepasst, d.h. es wurde zwischen den Bedeutungen der einzelnen Suffixe nicht getrennt. Die Ergebnisse dieser Gesamtberechnung sind zusammenfassend in Tabelle 5.2 dargestellt.

Abbildungen 1 und 2 zeigen diese Resultate in anschaulicher Form. Eine Gegenüberstellung der empirischen und theoretischen Daten kann man dem Anhang, Tabelle 3.1, entnehmen.

Tabelle 5.1.: Verteilung der Vorkommen von Nominalsuffixen in „Madame Bovary“ von G. Flaubert

Suffixe	Bedeutungen	Beispiele	Vorkommen im Text
-ance/-ence	a) Bezeichnung von Eigenschaften	élégance	82
	b) Bezeichnung der Handlung oder des Ergebnisses der Handlung	résistance, existence	58
	c) Personenbezeichnung	(une) connaissance, (une) intelligence	6
	d) Angabe des Zustandes	Alliance	5
	e) Perioden	Enfance	3
-ement	a) Bezeichnung der Handlung und/oder des Ergebnisses der Handlung	équipement, accroissement	131
	b) Angabe des Zustands	adoucissement	28
	c) Gegenstands- oder Sachbezeichnung	vêtement	12
	d) Ortsbezeichnung	établissement	8
	e) Vorübergehender Zustand von Lebewesen	abattement	4
-erie	a) Bezeichnung der Handlung und /oder des Ergebnisses der Handlung	causerie	26
	b) Bezeichnung von Eigenschaften	coquinerie	21
	c) Ort der Herstellung oder des Verkaufs	mercerie	12
	d) Vorübergehender Zustand	rêverie	6
	e) Kollektivbezeichnung	sellerie	4
-esse(s)	a) Bezeichnung von Eigenschaften	délicatesse	57
	b) Bezeichnung der Handlungen, Worten oder Gedanken	délicatesses	11
	c) Kollektivbezeichnung	jeunesse	9

	d) Vorübergehender Zustand von Lebewesen	ivresse	5
-(u)ité	a) Bezeichnung von Eigenschaften	assiduité	20
	b) Personenbezeichnung	(une) célébrité	5
	c) Kollektivbezeichnung	(une) communauté	1
	d) Zustand	brochite	1
-(at)ure	a) Bezeichnung der Handlung und/oder des Ergebnisses der Handlung	égratignure	69
	b) Kollektivbezeichnung	nourriture	27
	c) Gegenstands- oder Sachbezeichnung	chaussure	10
	d) Ortsbezeichnung	filature	2

Tabelle 5.2: Vorkommen von Nominalsuffixen in „Madame Bovary“ von G. Flaubert. Anpassung der 1-verschobenen und der 0-gestutzen negativen Binomialverteilung (Gesamtberechnung)

1-verschobene Negative Binomialverteilung				0-gestutzte negative Binomialverteilung			
k	**p**	χ^2_{24}	**P(χ^2)**	**k**	**p**	χ^2_{24}	**P(χ^2)**
0.79	0.13	19.5229	0.7236	0.68	0.12	20.2858	0.6804

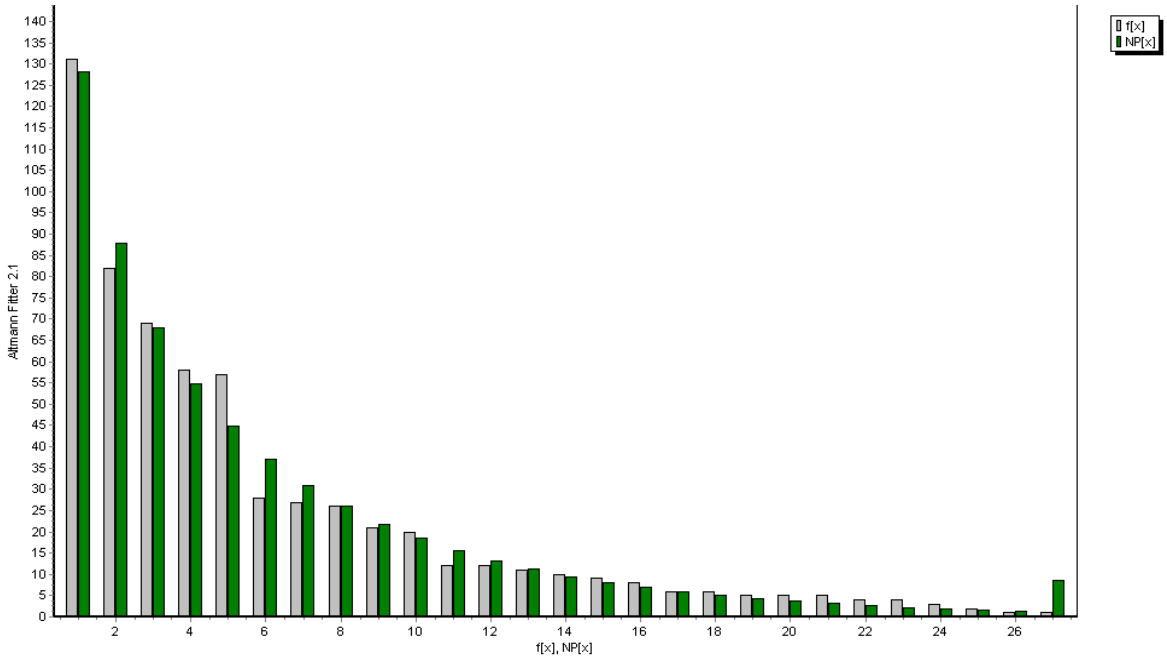

Abbildung 1: Anpassung der 1-verschobenen NB an die Diversifikation von Nominalsuffixen

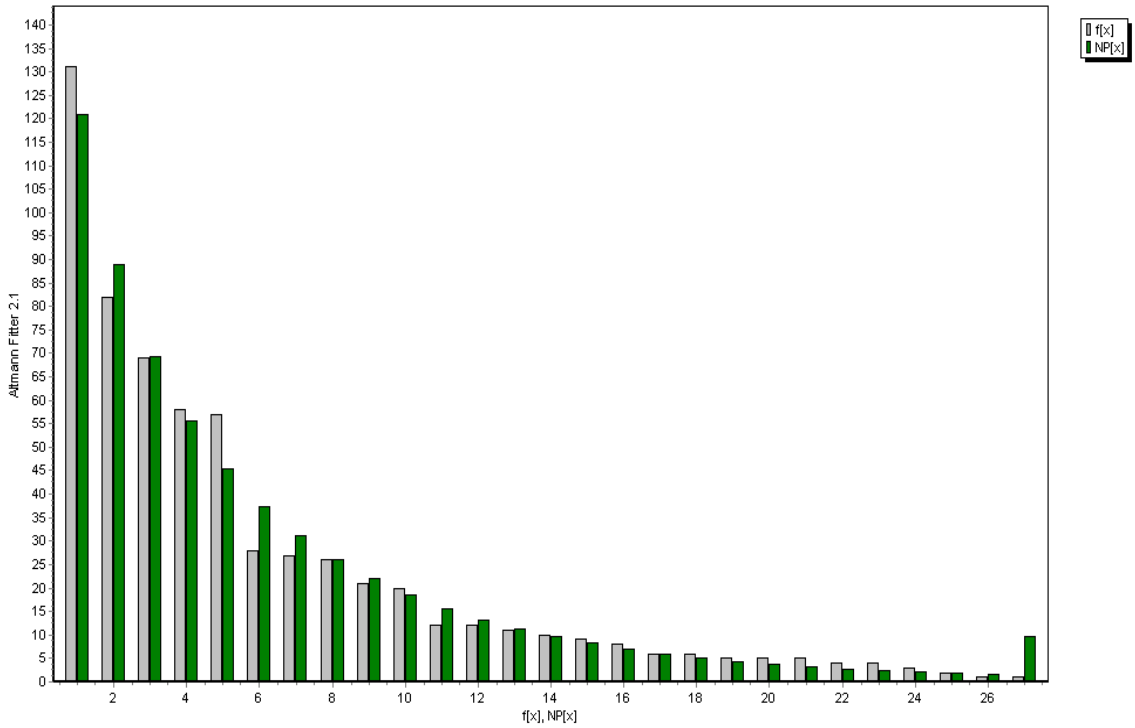

Abbildung 2: Anpassung der 0-gestutzten NB an die Diversifikation von Nominalsuffixen

Wie man sieht, zeigt die Untersuchung der Diversifikation französischer Wortbildungssuffixe eine Übereinstimmung mit den Modellen. D.h.: Die verschobene negative Binomialverteilung und die 0-gestutzte Variante erweisen sich als gute Modelle; die Anpassungsergebnisse sind sehr zufrieden stellend.; für beide Verteilungsmodelle wurde ein Chiquadrat-Wert erzielt, der weit über $P \geq 0.05$ liegt.

Die Parameter k und p haben in etwa gleiche Werte bei der ungestutzten und der gestutzten negativen Binomialverteilung. Dies deutet darauf hin, dass es für die analysierte Stichprobe nicht wesentlich ist, welches Verteilungsmodell bevorzugt wird.

(b) Im zweiten Schritt wurden zwischen den einzelnen Suffixen unterschieden. Die Bedeutungen innerhalb jedes Morphems wurden rangiert und erneut an die ungestutzte und die gestutzte negative Binomialverteilung angepasst. Tabelle 5.3 enthält die Zusammenfassung der Anpassungsergebnisse. Die empirischen und theoretischen Werte sind in Tabelle 3.2 im Anhang aufgeführt.

Wie aus Tabelle 5.3 hervorgeht, ergeben sich auch bei der Berechnung der Bedeutungen einzelner Suffixe gute Anpassungen. Beste Ergebnisse werden bei dem Suffix „**-(u)ité“** erzielt, mit einem Diskrepanzkoeffizienten von C ≈ 0.0000 für beide Verteilungen. Abbildungen 3 und 4 zeigen diese Resultate in anschaulicher Form.

In einem Fall konnten für beide Verteilungen keine guten Ergebnisse erzielt werden. Hier konkurrieren anscheinend unterschiedliche Stammmorpheme miteinander, was durchaus die Verzerrung bewirkt haben kann.

Insgesamt zeigt die Untersuchung, dass das Vorkommen der Bedeutungen von Suffixen im Französischen auch einer Gesetzmäßigkeit folgt.

Tabelle 5.3: Vorkommen der Bedeutungen einzelner Suffixe in „Madame Bovary“ von G. Flaubert. Anpassung der 1-verschobenen und der 0-gestutzen negativen Binomialverteilung

	Negative Binomialverteilung					0-gestutzte negative Binomialverteilung				
Suffixe	**k**	**p**	**DF**	**X^2**	**$P(\chi^2)$**	**k**	**p**	**DF**	**χ^2**	**$P(\chi^2)$**
-ance/-ence	3.64	0.84	2	10.6769	***C = 0.0048***	16	0.93	2	11.8972	***C = 0.0026***
-ement	0.56	0.49	2	2.9540	**0.2283**	0.01	0.45	2	3.0538	**0.2172**
-erie	5.12	0.81	2	0.5886	**0.7451**	8	0.82	2	0.008	**0.9960**
-esse(s)	0.88	0.58	1	4.2312	**0.0397**	0.23	0.48	1	3.2967	**0.0694**
-(u)ité	0.86	0.70		0.0032	**C = 0.0001**	0.53	0.67		0.0005	**C = 0.0000**
-(at)ure	4.33	0.89	1	0.7379	**0.3903**	40.21	0.97	1	0.4737	**0.4913**

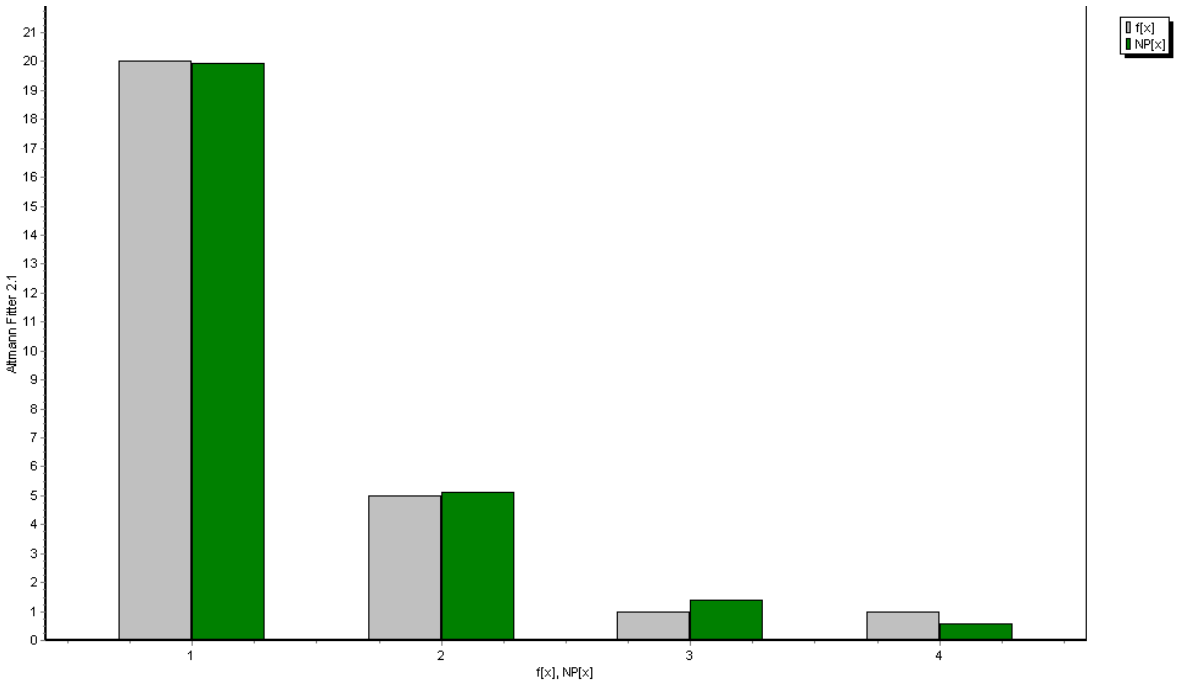

Abbildung 3: Anpassung der 1-verschobenen NB an die Diversifikation des Nominalsuffixes *„(u)ité”*

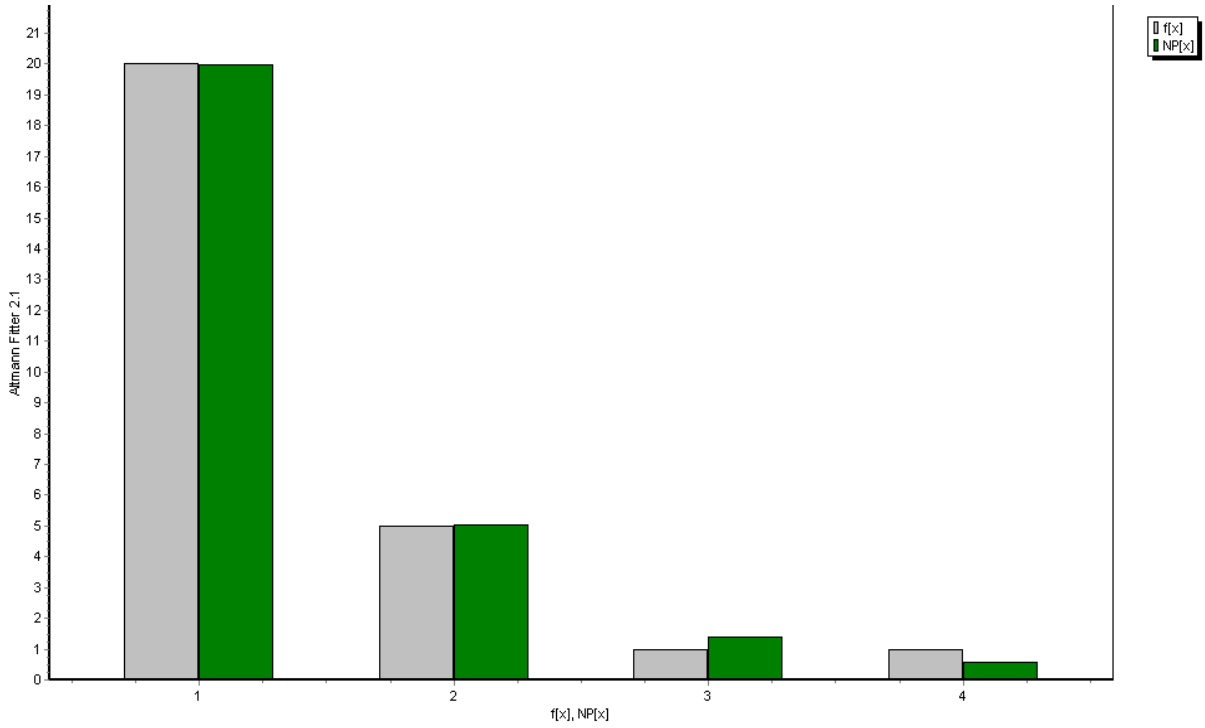

Abbildung 4: Anpassung der 0-gestutzten NB an die Diversifikation des Nominalsuffixes *„(u)ité"*

5.2. Ranghäufigkeitsverteilung von Wörtern

An die Untersuchung der Worthäufigkeiten in einem Text kann man auf zweierlei Weise herangehen: Entweder unterscheidet man zwischen den verschiedenen Formen eines Wortes und zählt jede Wortform separat oder man lemmatisiert zuerst den Text und zählt nur die Lemmata. Untersucht wurde hier die Ranghäufigkeitsverteilung von Wörtern.

Als Datenbasis dienten die Texte 1 bis 18, also alle Zeitungstexte, und die literarischen Texte 21, 23, 25, 27 bis 32; 38 und 39. Die Texte wurden einzeln bearbeitet.

Zunächst wurde das Vorkommen jedes Wortes in dem jeweiligen Text gezählt. Anschließend wurden die so ermittelten Wortlisten in Rangfolgen mit absoluter Frequenz jedes Wortes gebracht. Die erzeugten Ranghäufigkeitslisten wurden mit dem Zipf-Mandelbrot-Gesetz modelliert.

a) Modellierung der Ranghäufigkeitsverteilung von Wörtern in Zeitungstexten

In eine Rangfolge gemäß ihrer Häufigkeit gebracht, ergaben sämtliche Wortformen eines Textes eine Verteilung mit mehreren Datenpaaren. Abbildung 5 zeigt die empirische Ranghäufigkeitsverteilung von Text 5, der kürzesten Stichprobe, und ihre Erfassung durch die Zipf-Mandelbrot-Verteilung. Eine Gegenüberstellung der empirischen und theoretischen Wer-

te wird im Anhang, Tabelle 4.1, aufgeführt. Diese Gegenüberstellung steht stellvertretend für alle anderen Stichproben.

Eine Übersicht über die Anpassungsergebnisse aller 18 Zeitungstexte enthält Tabelle 5.4.

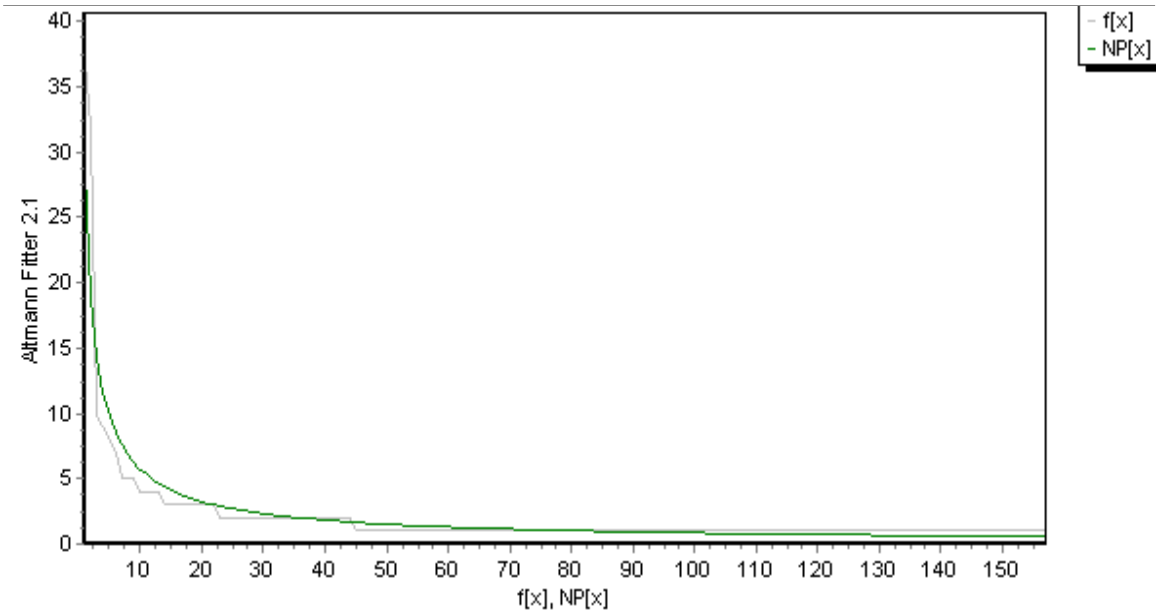

Abbildung 5: Anpassung der Zipf-Mandelbrot-Verteilung an die Ranghäufigkeitsverteilung von Wörtern im Text 5

Tabelle 5.4: Anpassung der Zipf-Mandelbrot-Verteilung an die empirischen Ranghäufigkeitsverteilungen der Wörter in den Zeitungstexten

Text 1	**Text 2**	**Text 3**	**Text 4**	**Text 5**	**Text 6**
a = 0.8443 b = 1.2461 n = 593 X^2_{451}=72.5731 **P(X^2) ≈ 1**	a = 0.8962 b = 1.5538 n = 1039 X^2_{821}=138.6395 **P(X^2) ≈ 1**	a = 0.9034 b = 1.4614 n = 718 X^2_{583}=59.1516 P(X^2) ≈ 1	a = 0.8337 b = 1.7537 n = 501 X^2_{383}=59.3488 **P(X^2) ≈ 1**	a = 0.8385 b = 1,8050 n = 287 X^2_{224} =23.09 **P(X^2) ≈ 1**	a = 0.8453 b = 1.8060 n = 383 X^2_{299}= 34.9654 **P(X^2) ≈ 1**
Text 7	**Text 8**	**Text 9**	**Text 10**	**Text 11**	**Text 12**
a = 0.7715 b = 1.1065 n = 427 X^2_{320}=53.9456 **P(X^2) ≈ 1**	a = 0.8076 b = 1.4111 n = 440 X^2_{331}= 57.7978 **P(X^2) ≈ 1**	a = 0.8373 b = 1.3613 n = 385 X^2_{294}=43.0852 **P(X^2) ≈ 1**	a = 0.8184 b = 1.5113 n = 484 X^2_{370}=50.152 **P(X^2) ≈ 1**	a = 0.8409 b = 1.5329 n = 478 X^2_{373}=51.7619 **P(X^2) ≈ 1**	a = 0.8084 b = 1.8694 n = 330 X^2_{251}= 37.2964 **P(X^2) ≈ 1**
Text 13	**Text 14**	**Text 15**	**Text 16**	**Text 17**	**Text 18**
a = 0.8469 b = 1.8109 n = 231 X^2_{182}=15.7845 **P(X^2) ≈ 1**	a = 0.8277 b = 1.9256 n = 466 X^2_{364}= 42.6932 **P(X^2) ≈ 1**	a = 0.8458 b = 0.5713 n = 157 X^2_{115}= 31.0501 **P(X^2) ≈ 1**	a = 0.9109 b = 0.5516 n = 293 X^2_{215}= 93.0253 **P(X^2) ≈ 1**	a = 0.9239 b = 0.9094 n = 300 X^2_{224}= 48.6417 **P(X^2) ≈ 1**	a = 0.8760 b = 0.9682 n = 180 X^2_{132}= 34.6874 **P(X^2) ≈ 1**

Bewertung der Ergebnisse

Als Ergebnis der Untersuchung kann festgestellt werden: In allen bearbeiteten Zeitungstexten konnte nachgewiesen werden, dass die Häufigkeiten der Wortformen – in eine absteigende Rangfolge gebracht – im Französischen gemäß der Zipf-Mandelbrot-Verteilung vertreten sind. Das Modell bewährt sich für alle Texte mit hervorragenden Ergebnissen. Dies wird deutlich durch die Prüfgröße *P* hervorgehoben, die in allen Einzelstichproben einen Wert von ungefähr 1 hat. Der Altmann-Fitter, mit dem die Berechnungen durchgeführt wurden, hat Rundungen vorgenommen. Das Ergebnis zeigt jedoch sehr gute Anpassungen, die alle Kriterien zur Abschätzung der Güte erfüllen. Der Verlauf der graphischen Darstellung der empirischen Verteilung von Text 5, die hier stellvertretend für die Abbildungen anderer Texte steht, entspricht vollkommen den Erwartungen; die Gegenüberstellung der empirischen Daten und der theoretischen Funktion zeigt eine gute Übereinstimmung der Verteilungen. Die Gültigkeit des Zipf-Mandelbrot-Gesetzes kann somit für die untersuchten Stichproben als nachgewiesen betrachtet werden.

Wie den in Tabelle 5.4 enthaltenen Anpassungsergebnissen zu entnehmen ist, erweist sich nicht nur der Wahrscheinlichkeitswert *P* in allen Einzeltexten als stabil; auch die Parameter a und b, in erster Linie a, erscheinen als ziemlich konstant. Die Werte für *a* liegen im Intervall [0.7715, 0.9239] und die für *b* im Intervall [0.5516, 1.9256].

Es hat sich bei einer weiteren Überprüfung ergeben, dass die Ranghäufigkeitsverteilung von Wörtern in den behandelten Zeitungstexten mit der negativen hypergeometrischen Verteilung modelliert werden kann. Eine Zusammenfassung der Testergebnisse ist Tabelle 5.5 zu entnehmen. Wie die Werte dieser Tabelle zeigen, stellt die negative hypergeometrische Verteilung ein gutes Modell für die Daten von 16 Texten dar, was sich auch im Wert von *P* deutlich ausdrückt: Diese Stichproben zeigen mit $P \approx 1$ hervorragende Anpassungsergebnisse. Die Texte 2 und 3 weisen sowohl einen schlechten *P*- als auch einen *C*-Wert auf. Sie sind somit mit der negativen hypergeometrischen Verteilung nicht modellierbar, im Unterschied zu der Zipf-Mandelbrot-Verteilung.

Tabelle 5.5: Güte der Anpassung der negativen hypergeometrischen Verteilung an die empirische Ranghäufigkeitsverteilung in Zeitungstexten

Text 1	**Text 2**	**Text 3**	**Text 4**	**Text 5**	**Text 6**
K = 1.3215 M = 0.3261 n = 592 X^2_{483}=126.266 **$P(X^2) \approx 1$**	K = 1.4813 M = 0.3049 n = 1038 X^2_{863}=515.7599 **C = 0.1738**	K= 1.6911 M = 0.3472 n = 752 X^2_{605}=287.5415 C = **0.1333**	K = 1.3695 M = 0.3521 n = 500 X^2_{409}=161.6747 **$P(X^2) \approx 1$**	K = 1.4623 M = 0.3836 n = 286 X^2_{239} = 55.1376 **$P(X^2) \approx 1$**	K = 1.4393 M = 0.3664 n = 382 X^2_{319} = 95.0553 **$P(X^2) \approx 1$**

Text 7	**Text 8**	**Text 9**	**Text 10**	**Text 11**	**Text 12**
K = 1.2765 M = 0.3611 n = 426 X^2_{336}=50.3520 **$P(X^2) \approx 1$**	K = 1.2973 M = 0.3521 n = 439 X^2_{351} = 71.8748 **$P(X^2) \approx 1$**	K = 1.3537 M = 0.3497 n = 384 X^2_{315}=72.1915 **$P(X^2) \approx 1$**	K = 1.3529 M = 0.3531 n = 483 X^2_{394}=78.4438 **$P(X^2) \approx 1$**	K = 1.4513 M = 0.3599 n = 478 X^2_{398}= 84.6145 **$P(X^2) \approx 1$**	K = 1.3758 M = 0.3813 n = 329 X^2_{267} = 65.4434 **$P(X^2) \approx 1$**
Text 13	**Text 14**	**Text 15**	**Text 16**	**Text 17**	**Text 18**
K = 1.4948 M = 0.3934 n = 230 X^2_{193}=41.5682 **$P(X^2) \approx 1$**	K = 1.4415 M = 0.3689 n = 465 X^2_{387} = 98.6936 **$P(X^2) \approx 1$**	K = 1.2445 M = 0.3413 n = 156 X^2_{124}=22.1618 **$P(X^2) \approx 1$**	K = 1.1848 M = 0.2881 n = 292 X^2_{238}=49.5196 **$P(X^2) \approx 1$**	K = 1.3200 M = 0.3054 n = 299 X^2_{243}= 80.9032 **$P(X^2) \approx 1$**	K = 1.2538 M = 0.3365 n = 179 X^2_{141} = 26.0100 **$P(X^2) \approx 1$**

Im Hinblick auf die bisher dargestellten Ergebnisse kann festgehalten werden, dass das Zipf-Mandelbrot-Gesetz ein sehr gutes Modell für die Ranghäufigkeitsverteilung von Wörtern in französischen Pressetexten darstellt. Ob dieses Gesetz auch zur Erfassung der empirischen Rangverteilung von in literarischen Texten vorkommenden Wörtern geeignet ist, darauf wird im Folgenden eingegangen.

b) Modellierung der Ranghäufigkeitsverteilung von Wörtern in literarischen Texten

Erhalten wurden hier wiederum Häufigkeitslisten, die weit mehr als 99 Datenpaare enthalten. Die Ergebnisse der Anpassung der Zipf-Mandelbrot-Verteilung an die Daten sind in Tabelle 5.6 zusammenfassend dargestellt.

Tabelle 5.6: Anpassung der Zipf-Mandelbrot-Verteilung an die empirischen Ranghäufigkeitsverteilungen der Wörter in literarischen Texten

Text 21	**Text 23**	**Text 25**	**Text 27**	**Text 28**	**Text 29**
a = 0.8991 b = 1.2337 n = 369 X^2_{298}=26.9883 **$P(X^2) \approx 1$**	a = 1.0967 b = 1.3048 n = 2083 X^2_{1677}=432.2128 **C = 0.0351**	a = 0.9289 b = 1.0565 n = 437 X^2_{343}=42.9410 **$P(X^2) \approx 1$**	a = 1.0226 b = 1.4020 n = 1509 X^2_{1217}=217.2464 **C = 0.0326**	a = 0.9887 b = 0,5178 n = 1104 X^2_{763}=486.1230 **C = 0,1503**	a = 0.9692 b = 0.4167 n = 1734 X^2_{1336}=306.7487 **C = 0.0474**

Text 30	**Text 31**	**Text 32**	**Text 38**	**Text 39**
a = 0.9699 b = 1.3195 n = 586 X^2_{471} = 54.3974 **$P(X^2) \approx 1$**	a = 1.0026 b = 0.9579 n = 1077 X^2_{859}=142.4249 **$P(X^2) \approx 1$**	a = 0.9041 b = 1.0689 n = 473 X^2_{362} = 64.5346 **$P(X^2) \approx 1$**	a = 0.9853 b = 0.8884 n = 951 X^2_{760}=97.0385 **$P(X^2) \approx 1$**	a = 1.0458 b = 2.7297 n = 549 X^2_{435} = 58,6741 **$P(X^2) \approx 1$**

Wie dieser Tabelle entnommen werden kann, erweist sich bei 7 Texten die angepasste Verteilung als sehr geeignetes Modell; diese Dateien sind mit $P(X^2) \approx 1$ ein Beleg für sehr gute Anpassungsergebnisse. Auch der empirische Kurvenverlauf jeder dieser Stichproben entspricht den Erwartungen. Diese Kurven werden hier stellvertretend durch die Kurve von Text 21, der

kürzesten Stichprobe, dargestellt. In Tabelle 4.2 im Anhang findet man auch eine Gegenüberstellung der empirischen und theoretischen Werte.

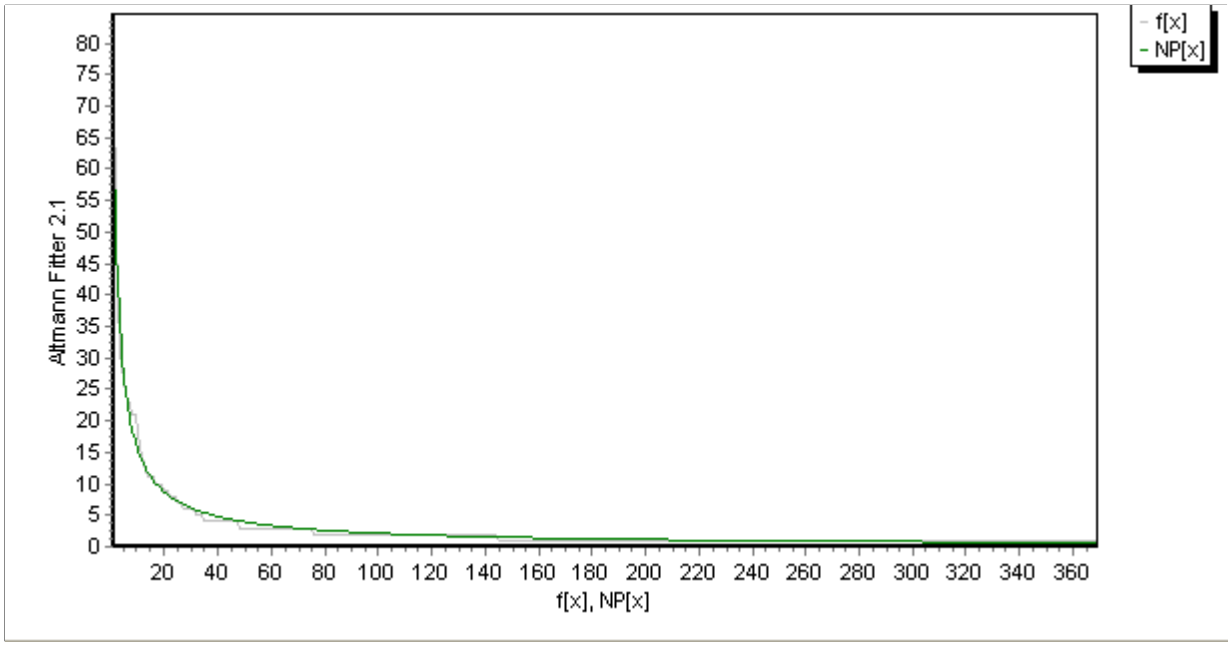

Abbildung 6: Anpassung der Zipf-Mandelbrot-Verteilung an die Ranghäufigkeitsverteilung von Wörtern im Text 21

In vier Texten, nämlich 23, 27, 28 und 29, versagt die Verteilung; das Gesetz stellt für die Daten dieser Texte kein gutes Modell dar. Dies drückt sich sowohl im *P*- als auch im *C*-Wert deutlich aus: Keine dieser Stichproben kommt auf einen Wert von $P \geq 0.05$ bzw. $C < 0.02$. Auch die Häufigkeitskurven dieser Texte zeigen relativ große Abweichungen zwischen den empirischen und den berechneten Werten. Text 23 repräsentiert diese Kurven hier stellvertretend.

Eine Erklärung, warum das Modell für vier Texte nicht bestätigt werden konnte, kann man in der Länge dieser Dateien finden. Die Texte, deren Daten an die Zipf-Mandelbrot-Verteilung mit Erfolg angepasst werden konnten, haben eine durchschnittliche Länge von 2314.12 Wortformen, wobei der längste Text und der kürzeste Text jeweils 4316 und 1023 Wortformen umfassen. Die durchschnittliche Länge anderer Texte liegt hingegen bei 8487.66 Wortformen, wobei der kürzeste Text 6473 und der längste 12329 Wortformen umfassen. In den behandelten Zeitungstexten ist die durchschnittliche Textlänge noch geringer.

Bei der Anpassung des Modells an die Daten scheint also der Aspekt der Länge eine bedeutsame Rolle gespielt zu haben; statistische Gesetzmäßigkeiten gelten nicht für Texte jeder beliebigen Länge.

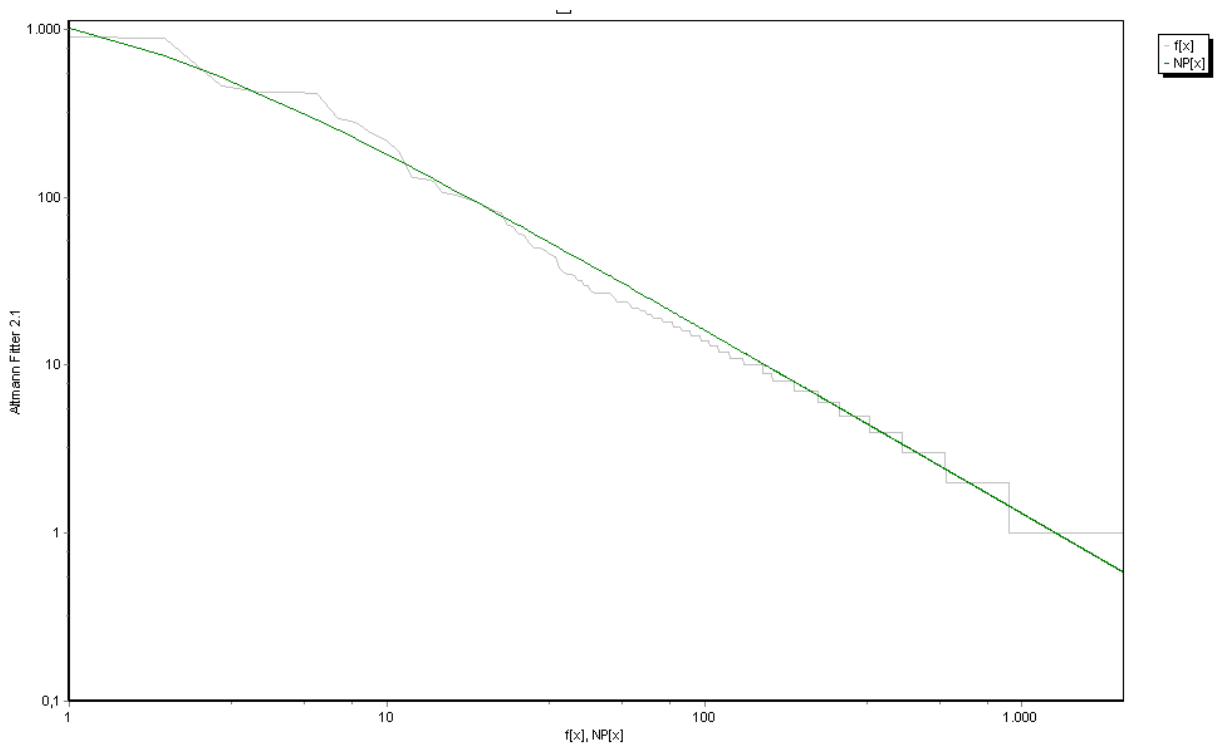

Abbildung 7: Anpassung der Zipf-Mandelbrot-Verteilung an die Ranghäufigkeitsverteilung von Wörtern im Text 23 (doppelt logarithmisches Koordinatensystem)

Der Einfluss der Textlänge auf die - statistische und stilistische - Textstruktur wurde bereits von anderen Autoren untersucht. ORLOV (1982a, 135) zeigt, dass mit der geplanten Textlänge der Informationsfluss unterschiedlich ausfällt:

> „Es besteht ein Zusammenhang zwischen dem Vokabularumfang in einem und der Häufigkeitsstruktur in einem beliebigen anderen Punkt des Textes. Die Häufigkeitsstruktur ist dynamisch und hängt von dem Textumfang ab.“

HAMMERL (1990, 151) begründet Schwierigkeiten bei der Anpassung von Wortlängenverteilungen mit dem Hinweis auf kurze Texte:

> „Unser Modell der Wortartenverteilung konnte jedoch für drei Texte nicht bestätigt werden. Es fällt aber sofort auf, daß es sich hier im Vergleich zu anderen Texten um wesentlich längere Texte handelt (mit einem N von 2000 bis über 3000 Wörtern).“

ALTMANN (1992) weist darauf hin, dass viele Texte wegen mangelnder Homogenität keine befriedigende Übereinstimmung mit theoretischen Modellen zeigen. Er schlägt vor, kurze Texte zu verwenden und Texte auszuwählen, die in einem Zug geschrieben und nicht nachträglich überarbeitet wurden:

„Für eine Texteigenschaft, wie z.B. Satzlänge, ist nicht einmal ein einziger Text, z.B. ein Roman, homogen. Pausen beim Schreiben rufen womöglich Rhythmusveränderungen hervor, so daß man in dieser Hinsicht nur Textteile als homogen betrachten kann.“ (S. 288)

Bei längeren Texten besteht somit die Gefahr, dass man bei deren Abfassung unterbrochen wird. Dies kann zur Folge haben, dass „der ursprünglich gefasste Plan der Textabfassung vergessen wird und von der Stelle der Unterbrechung an mit einem neuen Konzept weitergearbeitet wird.“ (BEST 2001a, 40)

Längere Texte sind demnach stilistisch und thematisch weniger homogen als kürzere Texte oder bloße Textfragmente. Hierin wird auch die Ursache gesehen, warum die hier behandelten Texte je nach Länge verschieden gut zu der Zipf-Mandelbrot-Verteilung passen.

An die Ranghäufigkeitsverteilungen in literarischen Texten wurde auch die negative hypergeometrische Verteilung angepasst. Die Anpassungsergebnisse sind im Unterschied zu denen von Zeitungstexten schlechter: Nur drei Stichproben belegen mit $P(X^2)\approx 1$ hervorragende Anpassungsergebnisse. In den restlichen Texten zeigt weder *P* noch *C* ein zufrieden stellendes Ergebnis.

Tabelle 5.7: Anpassung der negativen hypergeometrischen Verteilung an die empirischen Ranghäufigkeitsverteilungen der Wörter in literarischen Texten

Text 21	**Text 23**	**Text 25**	**Text 27**	**Text 28**	**Text 29**
K = 1.5161 M = 0.3415 n = 369 X^2_{314}=85.9498 **P(X²) ≈ 1**	K = 2.0022 M = 0.2407 n = 2367 X^2_{1735}=3541.6487 **C = 0.2873**	K = 1,4576 M = 0,3078 n = 436 X^2_{366}=150.9774 **P(X²) ≈ 1**	K = 1.7276 M = 0.2667 n = 1616 X^2_{1284}=1365.4762 C = 0.2050	K = 1.1274 M = 0.2149 n = 1103 X^2_{876}=431.2608 C = 0.1333	K = 1.3581 M = 0.2211 n = 1733 X^2_{1443}=1258.4173 C = 0.1944

Text 30	**Text 31**	**Text 32**	**Text 38**	**Text 39**
K = 1.7137 M = 0.3236 n = 620 X^2_{494} =282.3227 C = 0.1432	K = 1.7601 M = 0.2816 n = 1168 X^2_{903}=880.2120 C = 0.2039	K = 1.3823 M = 0.3091 n = 472 X^2_{389}=145.1667 **P(X²) ≈ 1**	K = 1.6827 M = 0.2827 n = 1013 X^2_{800}=545.3493 C = 0.1505	K = 1.8851 M = 0.3454 n = 602 X^2_{460} =323,5003 C = 0.1715

Die Häufigkeitsverteilung von Wörtern in Zeitungs- und literarischen Texten wurde auch aus der Sicht der Belegung der unterschiedlichen Frequenzen mit Wortvorkommen betrachtet. D.h.: Die Rangverteilung wurde in eine Klassenhäufigkeitsverteilung umgewandelt. Nach ZÖRNIG & BORODA (1990; auch ZÖRNIG & BORODA 1992) lässt sich diese Transformation auch mit der Zipf-Mandelbrot-Verteilung beschreiben. Dies wurde hier überprüft, konnte jedoch leider nicht in allen Fällen bestätigt werden: In 8 Zeitungstexten erwies sich das Modell als geeignet. Die Wahrscheinlichkeit *P* liegt in diesen Dateien im Intervall $0.7376 \geq P \geq 0.0633$.

In 7 literarischen Texten konnte die Zipf-Mandelbrot-Verteilung erfolgreich angewendet werden. In diesen Dateien liegt P im Intervall $0.9152 \geq P \geq 0.1792$.

Eine gelungene Anpassung an die empirischen Klassenhäufigkeitsverteilungen in allen Zeitungstexten und literarischen Texten ergab sich durch die Waring-Verteilung. Es handelt sich hierbei um jenes Modell, das von HERDAN (1964) eingeführt und u.a. von CHITASHVILI & BAAYEN (1993), MULLER (1965, 1968, 1977) und TEŠITELOVÁ (1967) erfolgreich angewendet wurde. Die mit dieser Verteilung erzielten P-Werte liegen für die Zeitungstexte im Intervall $0.7202 \geq P \geq 0.0848$; die entsprechenden Werte für die literarischen Texte liegen im Intervall $0.9245 \geq P \geq 0.0767$. Über alle Einzeltexte hinweg erweist sich somit P als stabil.

Auf Grund der empirischen Befunde kann festgehalten werden:

a) Als ausgezeichnetes Modell für die Anpassung rangierter Worthäufigkeiten in Zeitungstexten eignet sich die Zipf-Mandelbrot-Verteilung, mit der sich hervorragende Anpassungsergebnisse erzielen lassen. Diese Verteilung hat sich auch mit sehr guten Ergebnissen in sieben literarischen Texten bewährt. Für vier Stichproben konnte sie nicht bestätigt werden;

b) Die Anpassung der Worthäufigkeiten in Zeitungstexten durch die negative hypergeometrische Verteilung erbrachte bei 16 der 18 Texte ausgezeichnete P-Werte. Bei den literarischen Texten kamen demgegenüber nur drei Stichproben auf einen Wert von $P > 0.05$;

c) Für die Erfassung der Klassenhäufigkeiten erwies sich die Waring-Verteilung als geeignetes Modell, und zwar sowohl für Zeitungstexte als auch für literarische Texte.

5.3. Ranghäufigkeitsverteilung von Buchstaben und Phonemen

5.3.1 Ranghäufigkeitsverteilung von Buchstaben

Wir haben Arbeiten vorgestellt, in denen die Buchstabenhäufigkeiten an verschiedenen Sprachen systematisch untersucht worden sind. Im Folgenden sollen entsprechende Untersuchungen zum Französischen unternommen werden. Dabei richtet sich das Interesse auf die Suche nach einem einheitlichen Modell, dem ungeachtet der Häufigkeit der individuellen Buchstaben bzw. Grapheme verschiedene Stichproben unterliegen[19].

Für die Untersuchung wurden die literarischen Texte 19 bis 45 berücksichtigt. Diese wurden einzeln und unter zwei verschiedenen Bedingungen ausgewertet, die sich durch den jeweils angesetzten Inventarumfang unterschieden: Einmal beruhte die Buchstabenzählung auf den 26 Buchstaben des lateinischen Alphabets, wobei Buchstaben mit diakritischen Zeichen

[19] Wie bereits gesagt, sollen folgende Modelle auf ihre Adäquatheit hin überprüft werden: Zeta-Verteilung, geometrische Verteilung, Zipf-Mandelbrot-Verteilung, Good-Verteilung und die negative hypergeometrische Verteilung.

nicht als eigentliche Einheiten, sondern als Allographe aufgefasst, d.h. unter den entsprechenden Standardbuchstaben eingeordnet wurden. Zum anderen wurde bei der Zählung ein Inventarumfang von 38 Einheiten unter Einbeziehung der 26 Standardbuchstaben und der zwölf Diakriten à, â, é, è, ê, ë, î, ï, ô, ù, û, ç zu Grunde gelegt.
Damit ist aber der Inventarumfang längst nicht ausgeschöpft. Die Problematik des zu Grunde zu legenden Inventars ist viel komplexer als hier dargestellt. Darauf wird hier aber nicht eingegangen.

Empirische Überprüfung der Verteilungsmodelle

Wie es bereits an anderer Stelle erwähnt wurde, wird der Diskrepanzkoeffizient *C* zur Beurteilung der Güte der Anpassung verwendet, wenn die Wahrscheinlichkeit *P* auf Grund des Stichprobeumfangs N oder mangels Freiheitsgraden nicht berechnet werden kann. Da wir bei den Buchstabenhäufigkeiten meistens mit großen Stichproben zu tun haben, wird bei der Bewertung der Güte der einzelnen Anpassungen der Diskrepanzkoeffizient verwendet.

In den unten aufgeführten Tabellen werden die Ergebnisse der Anpassungen dargestellt. Die Tabellen enthalten neben der Textnummer die Parameterwerte für die einzelnen Verteilungsmodelle sowie den Chiquadrat-Wert und den Wert des Diskrepanzkoeffizienten *C*. In Tabellen 5.1 und 5.2 im Anhang finden sich die empirischen Buchstabenhäufigkeiten.

Die Anpassung der rechts-gestutzten Zeta-Verteilung an die Daten der Einzelstichproben erbrachte für die Bedingungen I = 26 und I = 38 die Ergebnisse, die in Tabelle 5.8. ersichtlich sind. Wie den Werten in dieser Tabelle zu entnehmen ist, stellt diese Verteilung kein gutes Modell dar. Dies drückt sich in den Werten der Prüfgröße *C* deutlich aus: Die *C*-Werte liegen für I = 26 im Intervall $0.2193 \geq C \geq 0.1623$ und für I = 38 im Intervall $0.\ 3565 \geq C \geq 0.2484$ und keine einzige Stichprobe weist einen Wert von $C < 0.02$ auf.

Diese negative Einschätzung gilt auch für die Zipf-Mandelbrot-Verteilung, deren Ergebnisse der Anpassung der Tabelle 5.9 entnommen werden können. Hier bewegen sich die Resultate für I = 26 im Intervall von $0.1417 \geq C \geq 0.0698$ und für I = 38 im Intervall $0.1978 \geq C \geq 0.0674$. In keiner Datei kann die festgelegte Signifikanzgrenze von $C < 0.02$ festgestellt werden.

Tabelle. **5.8**: Parameter und Anpassungsergebnisse der rechts-gestutzten Zeta-Verteilung an die Buchstabenhäufigkeit

	I = 26			**I = 38**		
	Rechts-gestutzte Zeta-Verteilung R = 26			**Rechts-gestutzte Zeta-Verteilung** R = 38		
Texte	**a**	X^2_{23}	**C**	**a**	X^2_{35}	**C**
19	0.78	984.0730	0.1684	0.86	1686.1639	0.2885
20	0.77	1585.1613	0.1880	0.87	2658.7952	0.3151
21	0.78	811.6288	0.1953	0.86	1320.7586	0.3178
22	0.77	10507.3844	0.2016	0.85	17349.1797	0.3329
23	0.78	10106.0460	0.1882	0.86	16804	0.3130
24	0.76	7631.8467	0.1876	0.85	12939.7072	0.3180
25	0.80	935.7672	0.1623	0.87	1604.7687	0.2784
26	0.76	9215.9593	0.2193	0.84	14981.9956	0.3565
27	0.78	5992.3516	0.1976	0.86	10024.6814	0.3306
28	0.81	2627.3061	0.1824	0.88	4232.3648	0.2938
29	0.77	5639.1126	0.1891	0.85	9413.3726	0.3157
30	0.77	1844.3673	0.2104	0.85	3030.0913	0.3457
31	0.76	4097.4100	0.2161	0.84	6731.9677	0.3550
32	0.76	1004.9400	0.1895	0.84	1672.5226	0.3153
33	0.75	6980.7421	0.1909	0.84	11859.7558	0.3244
34	0.77	3262.4338	0.2005	0.85	5420.6976	0.3331
35	0.76	6165.7193	0.2002	0.85	10193.6206	0.3310
36	0.77	11175.1439	0.2088	0.86	18188.2552	0.3411
37	0.77	10370.5457	0.1815	0.85	17452.0228	0.3054
38	0.78	3423.2380	0.2014	0.85	5417.8100	0.3188
39	0.75	1441.7178	0.1977	0.84	2398.9076	0.3290
40	0.73	12732.4626	0.2078	0.84	20266.0125	0.3252
41	0.78	10515.2371	0.1993	0.85	17334.8217	0.3286
42	0.78	9543.8942	0.1936	0.86	15761.5680	0.3197
43	0.77	5905.3638	0.1912	0.85	9934.3853	0.3205
44	0.78	6997.4296	0.1939	0.85	11692.4406	0.3241
45	0.79	8654.9362	0.1828	0.87	14282.3321	0.3016

Tabelle 5.9: Parameter und Anpassungsergebnisse der Zipf-Madelbrot-Verteilung an die Buchstabenhäufigkeit

	I = 26				**I = 38**			
	Zipf-Mandelbrot-Verteilung n = 26				**Zipf-Mandelbrot-Verteilung n = 38**			
Texte	**a**	**b**	X^2_{22}	**C**	**a**	**b**	X^2_{34}	**C**
19	1.33	2.39	734.5908	0.1257	1.45	2.39	1131.7097	0.1936
20	1.91	5.54	946.2518	0.1121	2.37	7.16	1225.2941	0.1452
21	2.35	7.88	444.4158	0.1069	2.90	10.38	519.0161	0.1249
22	11.99	69.97	4031.0846	0.0773	8.86	48-05	4352.1567	0.0835
23	7.66	41.63	4516.3485	0.0841	11.99	85.98	4316.0289	0.0804
24	5.19	26.58	3423.7547	0.0841	5.84	29.44	3828.1329	0.0941
25	1.58	3.65	698.7644	0.1212	1.82	4.29	952.6615	0.1652
26	11.99	78.29	3459.9799	0.0823	11.99	85.01	3540.6221	0.0843

27	11.99	83.46	2311.0329	0.0762	11.99	80.12	2067.1539	0.0682
28	1.36	2.26	2040.6750	0.1417	1.50	2.44	1849.1229	0.1978
29	4.08	18.93	2738.0043	0.0918	11.99	80.53	2112.3709	0.0708
30	11.99	73.50	638.6566	0.0729	11.99	81.09	653.2122	0.0745
31	11.99	72.99	1587.7052	0.0837	11.99	85.29	1648.3563	0.0869
32	11.99	84.34	429.8268	0.0810	11.99	89.18	434.4854	0.0819
33	10.64	63.12	2727.4551	0.0746	11.99	82.68	2675.8558	0.0732
34	11.99	72.06	1174.5528	0.0722	11.99	69.94	1177.9203	0.0724
35	5.05	24.95	2764.2341	0.0898	11.99	94.18	2961.8835	0.0962
36	11.99	69.67	4036.8282	0.0754	11.99	81.60	4028.1089	0.0755
37	3.68	16.24	4604.4874	0.0806	7.66	41.50	4035.9203	0.0706
38	1.52	3.32	2395.9808	0.1410	2.38	7.42	2526.3473	0.1487
39	5.74	30.32	609.9063	0.0837	11.98	85.06	534.6308	0.0733
40	3.86	18.04	5502.9862	0.0898	7.14	38.05	5115.6492	0.0821
41	5.64	28.60	4675.3914	0.0886	4.70	21.54	5577.0727	0.1057
42	1.71	4.37	6423.3085	0.1303	1.98	5.04	8586.5850	0.1742
43	8.45	46.71	2155.9023	0.0698	11.25	64.65	2089.9633	0.0674
44	11.99	81.48	3132.6964	0.0868	11.99	81.18	2767.6657	0.0767
45	4.65	21.66	3691.1790	0.0779	5.85	28.61	3731.9568	0.0788

In Anbetracht der Befunde scheiden die Zeta-Verteilung und die Zipf-Mandelbrot-Verteilung für die weiteren Betrachtungen aus, in deren Verlauf wir uns als Nächstes der geometrischen und der Good-Verteilung zuwenden. Die Anpassungsergebnisse für beide Modelle finden sich in den Tabellen 5.10 und 5.11. Deutlich ist zu sehen, dass beide Verteilungen für die hier untersuchten Daten gänzlich ungeeignet sind; für diese Modelle liegen alle Stichproben oberhalb der Signifikanzschwelle. Für die Bedingungen I = 26 und I = 38 bewegen sich die *C*-Werte für die geometrische Verteilung jeweils im Intervall von $0.0833 \geq C \geq 0.0567$ und $0.0654 \geq C$ 0.0437; für die Good-Verteilung liegen sie jeweils im Intervall von $0.4808 \geq C \geq 0.0763$ und $0.4661 \geq C \geq 0.0582$.

Tabelle 5.10: Parameter und Anpassungsergebnisse der rechts-gestutzten geometrischen Verteilung an die Buchstabenhäufigkeit

	I = 26			**I = 38**		
	Rechts-gestutzte geometrische Verteilung R = 26			**Rechts-gestutzte geometrische Verteilung** R = 38		
Texte	**q**	$\mathbf{X^2_{23}}$	**C**	**q**	$\mathbf{X^2_{35}}$	**C**
19	0.86	390.5170	0.0668	0.86	315.8847	0.0540
20	0.86	507.2295	0.0601	0.86	420.6752	0.0498
21	0.86	262.5039	0.0632	0.86	195.9868	0.0472
22	0.86	3709.9216	0.0712	0.86	2926.4214	0.0562
23	0.86	4062.5942	0.0757	0.86	3142.2975	0.0585
24	0.86	2814.1417	0.0692	0.87	2351.3607	0.0578
25	0.86	480.4002	0.0833	0.86	376.8578	0.0654

26	0.86	3144.3560	0.0748	0.86	2527.4414	0.0601
27	0.86	1905.0670	0.0628	0.86	1549.7758	0.0511
28	0.85	1164.9041	0.0809	0.86	867.4000	0.0602
29	0.86	2146.4260	0.0720	0.86	1669.4109	0.0560
30	0.86	589.3062	0.0672	0.86	480.2957	0.0548
31	0.86	1482.1133	0.0782	0.86	1223.9369	0.0645
32	0.86	383.3014	0.0723	0.87	302.0633	0.0570
33	0.86	2499.0849	0.0684	0.87	2101.9488	0.0575
34	0.86	1094.0465	0.0672	0.86	887.0736	0.0545
35	0.86	2183.1329	0.0709	0.86	1765.3419	0.0573
36	0.86	3678.3293	0.0687	0.86	2941.3057	0.0552
37	0.86	3242.9369	0.0567	0.87	2505.6653	0.0438
38	0.86	1202.6259	0.0708	0.86	886.4944	0.0522
39	0.86	491.6813	0.0674	0.87	396.1092	0.0543
40	0.86	3675.7280	0.0600	0.87	3121.2392	0.0501
41	0.86	3829.0811	0.0726	0.86	3010.9535	0.0571
42	0.86	3688.4314	0.0748	0.86	2924.6077	0.0593
43	0.86	1849.5469	0.0599	0.86	1456.2899	0.0470
44	0.86	2764.2563	0.0766	0.86	2153.0660	0.0597
45	0.86	2821.0683	0.0596	0.86	2070.7958	0.0437

Tabelle 5.11: Parameter und Anpassungsergebnisse der Good-Verteilung an die Buchstabenhäufigkeit

	I = 26				**I = 38**			
	Good-Verteilung				**Good-Verteilung**			
Texte	**a**	**p**	X^2_{23}	**C**	**a**	**p**	X^2_{35}	**C**
19	8.93	0.85	501.9936	0.0859	3	0.86	340.2128	0.0582
20	0.89	0.98	3700.0870	0.4385	0.87	0.98	3508.9850	0.4158
21	0.89	0.98	1839.7591	0.4427	0.86	0.98	1448.8850	0.4206
22	1.15	0.85	4649.2021	0.0892	0.86	0.98	2287.7403	0.4392
23	4.67	0.85	5047.4604	0.0940	5.56	0.86	3367.0037	0.0627
24	3.31	0.85	3643.9874	0.0896	0.85	0.98	17363.1680	0.4267
25	0.89	0.98	2331.2514	0.4044	0.87	0.98	2201.7309	0.3819
26	0	0.85	3935.4850	0.0937	0.85	0.98	19512.6908	0.4644
27	3.71	0.85	2421.5038	0.0798	0.86	0.98	13141.3590	0.4333
28	4.48	0.85	1383.1348	0.0960	0.31	0.89	1354.0436	0.0940
29	9.34	0.85	2719.3642	0.0912	0.86	0.98	12588.9834	0.4222
30	2.94	0.85	745.4528	0.0850	0.86	0.98	3943.0543	0.4498
31	3.59	0.85	1836.9109	0.0969	0.85	0.98	8839.1090	0.4661
32	7.68	0.85	488.7857	0.0922	0.85	0.98	2253.8079	0.4249
33	2.24	0.86	3264.3290	0.0893	0.85	0.98	15889.3414	0.4346
34	7.78	0.85	1389.9763	0.0854	0.86	0.98	7123.5655	0.4377
35	1.44	0.85	2768.0087	0.0899	0.86	0.98	13499.1916	0.4384
36	0.89	0.98	24727.5547	0.4621	0.87	0.98	23626.9091	0.4430
37	0.89	0.98	24584.9128	0.4302	0.86	0.98	23520.5019	0.4116
38	0.89	0.98	7678.9532	0.4518	0.86	0.98	7224.9710	0.4251
39	0.87	0.98	3389.6956	0.4449	0.85	0.98	3202.5603	04392
40	0.86	0.98	29461.1614	0.4808	0.83	0.98	27166.6206	0.4360

41	1.51	0.85	4809.3861	0.0912	0.86	0.98	22926.8485	0.4346
42	0.89	0.98	21807.2985	0.4424	0.87	0.98	20851.3680	0.4230
43	4.98	0.85	2420.9295	0.0784	0.86	0.98	13229.5276	0.4267
44	9.24	0.85	3390.2736	0.0939	0.86	0.98	15569.9755	0.4316
45	2.46	0.85	3612.1456	0.0763	0.87	0.98	19050.2136	0.4023

Auf Grund der bisherigen Befunde können die vier genannten Verteilungsmodelle als ungeeignet für die Modellierung der Buchstabenhäufigkeiten in den untersuchten französischen Texten angesehen werden. Wenden wir uns als letztes der negativen hypergeometrischen Verteilung zu. Dieses Modell wurde für die Modellierung von Ranghäufigkeitsverteilungen unterschiedlicher Einheiten herangezogen, d.h. nicht nur sprachlicher Entitäten. Verwendet wurde es beispielsweise zur Modellierung der Ranghäufigkeiten, mit denen Töne einer gegebenen Tonhöhe in den Werken von BACH und BEETHOVEN vorkommen (vgl. WIMMER & ALTMANN 2001). Die Anpassung mit diesem Modell erbrachte in den meisten Fällen gute Resultate.

Die Anpassung der negativen hypergeometrischen Verteilung an die Buchstabenhäufigkeitsverteilung in den untersuchten Einzeltexten führte zu den in Tabelle 5.12. aufgeführten Ergebnissen. Wie dieser Tabelle zu entnehmen ist, ergeben sich für diese Verteilung im Vergleich zu den anderen Modellen zwar bessere, aber auch keine überzeugenden Resultate; es liegen partiell annähernd akzeptable Anpassungen vor. Dies gilt für beide Untersuchungsbedingungen. Für I = 26 bewegen sich die Ergebnisse im Intervall von $0.0399 \geq C \geq 0.0232$, wobei in fünf Stichproben der C-Wert knapp über $C < 0.02$ liegt. Für die Bedingung I = 38 bewegen sich die Resultate im Intervall von $0.0360 \geq C \geq 0.0196$, wobei nur in einer Stichprobe ein $C < 0.02$ festgestellt werden kann. Für die restlichen Stichproben ist die Anpassung zwar nicht völlig fehlgeschlagen, aber auch nicht zufrieden stellend. Insgesamt ergibt sich damit für die negative hypergeometrische Verteilung kein überzeugendes Resultat.

Tabelle 5.12: Parameter und Anpassungsergebnisse der negativen hypergeometrischen Verteilung an die Buchstabenhäufigkeit

	I = 26				**I = 38**			
	Negative hypergeometrische Verteilung n = 25				**Negative hypergeometrische Verteilung n = 37**			
Texte	**K**	**M**	X^2_{22}	**C**	**K**	**M**	X^2_{34}	**C**
19	3.33	0.76	189.4132	0.0324	4.98	0.85	179.4015	0.0307
20	3.64	0.82	222.4366	0.0264	5.49	0.91	207.8219	0.0246
21	3.71	0.83	118.6078	0.0285	5.55	0.93	88.7784	0.0214
22	3.68	0.83	1770.4294	0.0340	5.58	0.94	1422.8530	0.273
23	3.45	0.78	2070.9694	0.0386	5.25	0.89	1704.5889	0.0317
24	4.15	0.84	1624.1306	0.0399	5.29	0.91	1250.6882	0.0307
25	3.17	0.72	221.8580	0.0385	4.86	0.82	199.9560	0.0347

26	3.88	0.86	1461.9250	0.0348	5.84	0.98	1176.3541	0.0280
27	3.58	0.80	1036.1108	0.0342	5.50	0.92	919.5912	0.0303
28	3.48	0.75	551.0092	0.0382	5.10	0.83	507.8234	0.0353
29	3.49	0.79	1099.7505	0.0369	5.08	0.87	1072.1335	0.0360
30	3.81	0.85	272.4783	0.0311	5.79	0.96	223.1253	0.0255
31	3.75	0.85	745.2369	0.0393	5.67	0.96	630.4522	0.0332
32	3.45	0.80	194.7806	0.0367	5.17	0.89	169.1348	0.0319
33	3.45	0.81	1320.7330	0.0361	5.20	0.90	1232.9411	0.0337
34	3.76	0.84	487.9935	0.0300	5.70	0.95	408.1287	0.0251
35	3.60	0.82	1089.5595	0.0354	5.43	0.92	940.0242	0.0305
36	3.81	0.85	1711.9079	0.0320	5.71	0.94	1503.1485	0.0282
37	3.62	0.82	1327.7327	0.0232	5.31	0.91	1122.5779	0.0196
38	3.72	0.82	559.7193	0.0329	5.44	0.92	430.8528	0.0254
39	3.61	0.84	236.8033	0.0325	5.41	0.94	216.6444	0.0297
40	3.85	0.89	1439.8179	0.0235	5.37	0.93	1556.0018	0.0250
41	*3.64*	*0.82*	*1871.5712*	*0.0355*	*5.47*	*0.92*	*1573.0970*	*0.0298*
42	3.53	0.79	1877.2902	0.0381	5.35	0.89	1622.7426	0.0329
43	3.63	0.82	849.3615	0.0275	5.44	0.93	691.1842	0.0223
44	3.57	0.80	1307.4291	0.0362	5.49	0.92	1041.0362	0.0289
45	3.63	0.80	1209.2320	0.0255	5.35	0.90	945.0515	0.0200

Resümierend lassen sich die erzielten Ergebnisse zu den einzelnen Verteilungsmodellen nicht in das Bild bisheriger Untersuchungen zu Graphem- und Buchstabenhäufigkeiten einfügen. Bei den Untersuchungen der Buchstabenhäufigkeiten slawischer Sprachen (Russisch, Slowenisch, Ukrainisch und Slowakisch) und des Deutschen hat sich, wie wir bereits gesehen haben, die negative hypergeometrische Verteilung als geeignetes Modell erwiesen. In den hier untersuchten Texten konnte aber die Gültigkeit dieses Modells nicht nachgewiesen werden. Unter den beiden Untersuchungsbedingungen haben wir es mit Anpassungsergebnissen zu tun, welche die Befunde aus vorherigen Untersuchungen nicht vollkommen bestätigen. Die Rangverteilungen von Buchstaben lassen sich durch keines der betrachteten Wahrscheinlichkeitsmodelle beschreiben. Ob dies nur für die hier untersuchten Stichproben gilt oder mit den zu Grunde gelegten Inventarumfängen in Zusammenhang steht, kann erst nach weiteren Untersuchungen geklärt werden.

5.3.2. Ranghäufigkeitsverteilung von Phonemen

In den vorausgehenden Analysen ist der Frage nachgegangen worden, welchem Gesetz die Buchstaben- bzw. Graphemhäufigkeiten in französischen Texten unterliegen. Es hat sich dabei herausgestellt, dass keines der Modelle, die für ähnliche Fragestellungen oft mit guten Ergebnissen verwendet wurden bzw. werden, geeignet ist.

Ausgehend von Untersuchungen zu Gesetzmäßigkeiten, denen Buchstaben-, Laut- und Phonemhäufigkeiten unterliegen, wird im Folgenden der Frage nachgegangen, welches Mo-

dell sich an die Phonemhäufigkeiten im Französischen mit guten Ergebnissen anpassen lässt. Es wurde angenommen, dass Laute und Phoneme sich nicht anders verhalten als Buchstaben. Dies konnte in einigen Fällen bestätigt werden (vgl. z.B. GRZYBEK & KELIH 2005, KELIH 2007). Vor der eigentlichen Modellierung der Phonemhäufigkeiten wird kurz auf das Verhältnis zwischen Phonem und Buchstaben eingegangen.

Zwischen Phonemen und Buchstaben bzw. Graphemen besteht keine 1:1-Entsprechungen. Ein Phonem entspricht nicht immer einem einzigen Buchstabe bzw. Graphem. Phoneme können durch Buchstabengruppen wiedergegeben werden, so beispielsweise <ch> oder <sch> für /ʃ/, wie z.B. in dem französischen Wort "chou" oder dem deutschen Wort "Schein". Auch kommt es vor, dass ein Graphem zwei Phonemen zugeordnet wird: In “die Flucht” (von “fliehen”) und “er flucht” (von “fluchen”) ist das Graphem “u” enthalten. Dieses Graphem entspricht auf der lautlichen Ebene zwei verschiedenen Phonemen; im ersten Fall wird es zu /u/, im zweiten Fall zu /u:/. Der Bedeutungsunterschied der Wörter "Flucht" und "flucht" ist auf die Divergenz beider Phoneme zurückzuführen.

Ein und dasselbe Phonem kann durch mehrere Buchstaben repräsentiert werden. Das französische Phonem /s/ beispielsweise weist 8 graphische Repräsentanten auf: in**s**ister, a**ss**i, he**x**amètre, **c**ible, fa**ç**on, **sc**ience, po**ti**on, hert**z.** Keinem Graphem entspricht exakt ein einziges Phonem. Der Buchstaben "x" deckt im Französischen 4 Phonemqualitäten ab: /k+s/ (z.B. in axe), /g+z/ (z.B. in exact), /s/ (z.B. in six), /z/ (z.B. in sixième).

Empirische Modellierung der Phonemhäufigkeiten

Für die Untersuchung der Phonemhäufigkeitsverteilung wurden die Daten aus der Arbeit von De KOCK (1983) verwendet. In dieser Studie geht es um die Abweichungen der relativen Häufigkeiten französischer Phoneme in einem Korpus von deren mittleren Häufigkeiten in mehreren Korpora. De KOCK analysiert zu diesem Zweck 11 Phonem-Häufigkeitslisten und vergleicht sie miteinander. Die Häufigkeitslisten wurden auf der Basis von 11 phonetisch transkribierten Korpora erstellt. In Tabelle 5.13 sind diese Korpora sowie deren Umfang (N) aufgeführt.

Bei der Erstellung der Phonem-Häufigkeitslisten verwendete De KOCK zwei verschiedene Inventarumfänge (I) von Phonemen: Bei den ersten 6 Korpora aus Tabelle 5.13 ging er von einem Umfang von 36 Phonemen aus, bei den restlichen 5 Korpora verwendete er einen Umfang von 35 Phonemen.

Tabelle 5.13: Transkribierte Korpora aus De KOCK (1983)

Korpora	N (in Phonemen)
J.A. BAUDOT	50189
L. SZKLARCZYK	286946
F. WIOLAND	77702
J.P. HATON/M. LAMOTTE	50033
J. De KOCK 1	1390900
J. De KOCK 2	8001
AUV1, 1	17582
AUV1, 2	10977
AUV1, 3	8511
AUV1, 4	3018
ORL 14	21298

Für die 11 Stichproben hat De KOCK Phonemhäufigkeiten (prozentualer Anteil) ermittelt und angeführt. Diese Daten wurden in der vorliegenden Arbeit für die Modellierung der Phonemhäufigkeiten herangezogen[20]. Die in Prozentzahlen angegebenen Häufigkeiten wurden in absolute Häufigkeiten umgewandelt. Auf Grund der ungenaueren Prozentangaben ergaben sich aus der Transformation nicht immer die tatsächlichen Stichprobengrößen.

Die bei der Buchstabenhäufigkeitsverteilung verwendeten Verteilungsmodelle wurden an die gewonnenen absoluten Häufigkeiten der Phoneme angepasst. Die Werte der absoluten Häufigkeiten können den Tabellen 6.1 und 6.2 im Anhang entnommen werden, während die Anpassungsergebnisse in den Tabellen 5.14 und 5.15 enthalten sind. Diese Resultate stellen sich folgendermaßen dar:

Sowohl die Zeta-Verteilung als auch die Good-Verteilung scheiden für die untersuchten Phonemhäufigkeiten gänzlich aus; in keinem einzigen Fall ist $C < 0.02$. Die Zipf-Mandelbrot-Verteilung zeigt kein besseres Ergebnis: Nur in einer einzigen Datei kann ein $C < 0.02$ festgestellt werden. Insgesamt ergibt sich für diese drei Verteilungen kein gutes Resultat. Demgegenüber zeigen die geometrische Verteilung und vor allem die negative hypergeometrische Verteilung recht bessere Resultate. Die geometrische Verteilung ist durchaus gelungen; für dieses Modell ergaben sich partiell sehr gute Anpassungen: Die C-Werte bewegen sich im Intervall von $0.0296 \geq C \geq 0.0162$, wobei in 4 Fällen $C < 0.02$ ist; für alle anderen Stichproben ist diese Grenze knapp überschritten. Die Ergebnisse sind hier zwar nicht überzeugend,

[20] Die Phonemhäufigkeiten für das Korpus J.A. BAUDOT wurden außer Acht gelassen.

aber annähernd akzeptabel. Für die negative hypergeometrische Verteilung lassen sich geradezu hervorragende Ergebnisse feststellen: In allen Stichproben liegen die *C*-Werte unterhalb der Signifikanzschwelle. Sie bewegen sich im Intervall von $0.0135 \geq C \geq 0.0039$, wobei diese in neun Fällen sogar unter 0.01 liegen. Damit können sowohl die negative hypergeometrische Verteilung als auch die geometrische Verteilung für die Modellierung der Phonemhäufigkeiten in den untersuchten Stichproben als geeignet angesehen werden.

Während bei der Modellierung der Buchstabenhäufigkeiten keine Verteilung passte, kommen für die Phonemhäufigkeiten zwei Modelle in Frage. Somit ist weder eine Zusammenführung der Buchstaben- und Phonemhäufigkeiten noch eine vergleichende Darstellung der jeweiligen C- und Parameterwerte möglich.

Offenbar haben die Buchstaben im Französischen eine völlig andere Ordnung als die Phoneme, was einige Rückschlüsse über die Graphem-Phonem-Relation und die Schrifteffizienz in dieser Sprache zulässt. Die Ergebnisse lassen die Schlussfolgerung zu, dass die Graphem-Phonem-Korrespondenz im Französischen nicht stark ausgeprägt ist. Es ist auch möglich, dass die Buchstaben im Französischen gar keinem bekannten Modell folgen.

Tabelle 5.14: Parameter und Anpassungsergebnisse der Verteilungen an die Phonemhäufigkeiten in den Korpora (I = 35)

Verteilungsmodelle		**Korpora**				
		AV1, 1	AV 1, 2	AV 1, 3	AV 1, 4	ORL14
Zeta-Verteilung R = 35	**a**	0.57	0.59	0.61	0.58	0.63
	$\mathbf{X^2_{32}}$	2517.2194	1401.1844	1107.1145	385.5786	2008.6018
	C	0.1432	0.1276	0.1301	0.1278	0.0943
Good-Verteilung	**a**	2.23	9.12	0.67	0.65	0.68
	p	0.91	0.91	0.98	0.98	0.98
	$\mathbf{X^2_{32}}$	1317.6395	811.8957	3233.4562	1192.1097	7203.8431
	C	0.0749	0.0740	0.3800	0.3950	0.3382
Zipf-Mandelbrot-Verteilung n = 35	**a**	5.89	11.99	1.52	11.99	2.19
	b	59.01	134.67	7.20	151.68	14.53
	$\mathbf{X^2_{31}}$	561.4231	245.1224	464.8867	59.1606	585.6927
	C	0.0319	0.0223	0.0546	0.0196	0.0275
Geometrische Verteilung R = 35	**q**	0.92	0.92	0.92	0.92	0.92
	$\mathbf{X^2_{32}}$	425.2141	255.0122	137.9009	59.4942	385.2732
	C	0.0242	0.0232	0.0162	0.0197	0.0181
Negative hypergeometrische Verteilung	**K**	3.44	3.10	3.13	3.15	2.79
	M	0.96	0.89	0.88	0.90	0.78
	n	36	34	35	35	34
	$\mathbf{X^2_{31}}$	69.3491	56.4188	81.8963	20.0767	95.8554
	C	0.0039	0.0051	0.0096	0.0067	0.0045

Tabelle 5.15: Parameter und Anpassungsergebnisse der Verteilungen an die Phonemhäufigkeiten in den Korpora (I = 36)

Verteilungsmodelle		**Korpora**				
		De Kock 1	De Kock 2	Haton/Lamotte	SZKLARCYK	Wioland
Zeta-Verteilung R = 36	**a**	0.63	0.64	0.65	0.58	0.61
	$\mathbf{X^2_{33}}$	178015.4444	980.5051	9172.2401	38063.9782	10506.1876
	C	0.1422	0.1229	0.1836	0.1327	0.1352
Good-Verteilung	**a**	0.68	2.73	7.29	0.64	0.66
	p	0.98	0.91	0.90	0.98	0.98
	$\mathbf{X^2_{33}}$	468367.8505	509.9787	3033.1429	112492.4780	29579.3441
	C	0.3743	0.0639	0.0607	0.3921	0.3807
Zipf-Mandelbrot-Verteilung n = 36	**a**	11.99	1.20	5.42	1.22	1.12
	b	251.90	3.91	44.41	5.11	3.73
	$\mathbf{X^2_{32}}$	146412.4743	517.1447	2186.4296	20167.9554	6159.2566
	C	0.1170	0.0648	0.0438	0.0703	0.0793
Geometrische Verteilung R = 36	**q**	0.92	0.92	0.91	0.92	0.92
	$\mathbf{X^2_{33}}$	23335.7431	182.8915	1481.1601	5987.7296	1779.9224
	C	0.0186	0.0229	0.0296	0.0209	0.0229
Negative hypergeometrische Verteilung	**K**	3.24	2.95	3.61	3.14	3.28
	M	0.88	0.82	0.92	0.90	0.89
	n	36	35	35	36	36
	$\mathbf{X^2_{32}}$	17438.6936	107.5914	393.3620	2680.6479	391.1370
	C	0.0139	0.0135	0.0079	0.0093	0.0050

5.4. Häufigkeitsverteilung der Satzlängen

In diesem Abschnitt soll die Clause/Satz-Verteilung empirisch überprüft werden. Als Datenbasis dienen 3172 Sätze und 7960 Teilsätze, die aus 13 literarischen Texten ausgewählt wurden, nämlich aus den Texten 23, 24, 25, 28, 29, 30, 31, 32, 33, 35, 37, 43 und 45. Die Textauswahl wurde relativ willkürlich getroffen.

Bei der Auswertung des Sprachmaterials wurde ausschließlich der laufende Text berücksichtigt. Die Erhebung der Daten erfolgte so, dass jeweils die Sätze eines Textes nacheinander ausgezählt wurden. Die Untersuchung beschränkte sich auf die Auswertung der Einzeltexte; es wurde keine Zusammenfassung von Texten vorgenommen.

Die erhobenen Daten wurden mit der negativen Binomialverteilung in deren 1-verschobenen und 0-gestutzten Formen und der 1-verschobene Hyperpoisson-Verteilung modelliert.

Ergebnisse und Bewertung

Die Ergebnisse der Berechnungen sind in den Tabellen 5.16 bis 5.18 zusammenfassend aufgeführt. Eine Gegenüberstellung der empirischen und theoretischen Werte für die jeweilige Verteilung findet sich im Anhang, Tabelle 7.

Betrachtet man die Ergebnisse der empirischen Überprüfung kann man festhalten:
Die 1-verschobene negative Binomialverteilung ist zum größten Teil gelungen. Sie stellt bei 11 von 13 Texten ein geeignetes Modell dar. 10 Texte konnten mit der 0-gestutzten negativen Binomialverteilung und der 1-verschobenen Hyperpoisson-Verteilung erfolgreich modelliert werden[21]. Bei Text 35 sind für beide Verteilungen die Anpassungen mit $P = 0.0228$ bzw. $P = 0.0253$ gerade noch akzeptabel.

Die Texte 31 und 45 konnten mit keiner der Verteilungen erfolgreich modelliert werden. Die *C*-Werte zeigen, dass die Anpassungen als völlig missglückt anzusehen sind. Dass beide Dateien nicht angepasst werden konnten, liegt möglicherweise an stilistischen Besonderheiten der Texte, da diese zu Abweichungen führen können. Für beide Texte lassen sich bereits anhand der empirischen Daten Abweichungen zu den Verteilungen von Satzlängen der übrigen Texte feststellen. Die Häufigkeitsverteilungen nehmen einen Kurvenverlauf an, der den Schluss zulässt, dass die Texte den Modellen nicht folgen bzw. keinen Trend erkennen lassen. Abbildungen 8 bis 13 zeigen in anschaulicher Form die empirischen und theoretischen Werte dieser Dateien. Hier können mehrere Häufigkeitsgipfel beobachtet werden.

Wie bereits gesehen stellt sich die 1-verschobenen Hyperpoisson-Verteilung als bestes Modell für die Clause/Satz-Verteilung in deutschen Texten dar. Das Gleiche lässt sich für die hier untersuchten Texte nicht belegen. Dieses Verteilungsmodell hat sich in 10 der 13 Einzelanalysen bewährt, wobei die erzielten Ergebnisse gegenüber den anderen Verteilungen nicht besser sind. Es gibt keine Datei, die sich nur nach der Hyperpoisson-Verteilung, nicht aber nach der negativen Binomialverteilung modellieren lässt.

Die 1-verschobene negative Binomialverteilung ermöglichte deutlich bessere Anpassungen; sie erwies sich als besseres Modell. Somit können die Annahmen von ALTMANN (1988a, 1988b) vorläufig als bestätigt angesehen werden.

Unter Berücksichtigung bisher durchgeführter Untersuchungen zur Satzlängenverteilung in Texten lassen die hier erzielten Ergebnisse nachfolgende Interpretation zu:
In den analysierten Texten ist die Verteilung der Satzlängen nicht als willkürlich oder gar als chaotisch zu betrachten; sie hat eine Ordnung, die durch theoretische Modelle beschreibbar ist. Die Existenz dieser Modelle ist ein Hinweis darauf, dass stochastische Gesetzmäßigkeiten, denen man bei der Textproduktion unbewusst folgt, den Text beeinflussen.

[21] Anders als in allen anderen Texten wurde an die Textdatei 32 die 0-gestutzte negative Binomialverteilung ohne Zusammenfassung von NP_x-Werten angepasst.

Es bleibt aber zu untersuchen, ob die erzielten Befunde nur für die hier bearbeiteten Textsorten gelten, oder ob bei der Berücksichtigung anderer wissenschaftlicher Texte nicht noch weitere Modelle in Betracht gezogen werden können oder gar müssen.

Tabelle 5.16: Parameter und Anpassungsergebnisse der 1-verschobenen negativen Binomialverteilung an die Satzlängenhäufigkeit in den Einzeltexten

Texte	k	p	DF	X^2	P (X^2)
23	6.2586	0.7461	6	7.2169	0.3013
24	0.9444	0.4434	7	10.9140	0.1424
25	3.2281	0.6949	3	0.7599	0.8590
28	0.84092	0.2192	10	7.3814	0.6890
29	2.0551	0.5592	6	7.1410	0.3080
30	1.4915	0.6236	3	0.3879	0.9427
31	2.3708	0.3992	10	26.1854	C = 0.2518
32	3.4754	0.7658	2	2.4691	0.2910
33	1.4832	0.5993	5	7.8835	0.1628
35	7.3948	0.8241	4	8.3890	0.0783
37	10.3523	0.8534	5	2.6855	0.7483
43	2.2846	0.6178	5	2.7419	0.7382
45	3.3588	0.6564	6	18.1921	C = 0.0437

Tabelle 5.17: Parameter und Anpassungsergebnisse der 0-gestuzten negativen Binomialverteilung an die Satzlängenhäufigkeit in den Einzeltexten

Texte	k	p	DF	X^2	P (X^2)
23	22.1462	0.8831	6	3.8426	0.6980
24	1.0659	0.4495	8	9.4058	0.3092
25	7.3776	0.7860	3	0.5418	0.9096
28	0.8495	0.2247	10	7.0389	0.7218
29	4.7315	0.6913	5	6.2913	0.2789
30	3.2791	0.7253	3	0.1130	0.9902
31	3.3807	0.4432	10	26.2249	C = 0.2522
32	5.7011	0.7745	4	4.7239	0.3168
33	2.1048	0.6305	5	7.1601	0.2090
35	113.1277	0.98	4	11.3610	0.0228
37	462.5538	0.9944	5	3.3547	0.6455
43	7.6474	0.7899	5	2.8171	0.7282
45	8.0405	0.7681	6	24.0978	C = 0.0579

Tabelle 5.18: Parameter und Anpassungsergebnisse der 1-verschobenen Hyperpoisson-Verteilung an die Satzlängenhäufigkeit

Texte	a	b	DF	X^2	P (X^2)
23	3.5669	2.7489	6	3.4075	0.7562
24	319.2448	587.9562	7	10.0496	0.1858
25	3.5095	4.1412	3	0.4987	0.9192
28	2267.8195	3049.6375	9	7.1096	0.6257
29	3.2512	3.4871	4	6.5787	0.1599
30	3.4328	5.8704	3	0.2508	0.9690
31	12.9981	12.1110	9	25.5229	C = 0.2454
32	1.8265	2.3546	2	2.2713	0.3212
33	8.6259	15.7793	5	6.2767	0.2802
35	2.1723	1.8187	4	11.1151	0.0253
37	2.3583	1.7287	4	2.7484	0.6008
43	3.2469	3.7460	4	2.5401	0.6375
45	3.2543	2.9892	5	28.4497	C = 0.0684

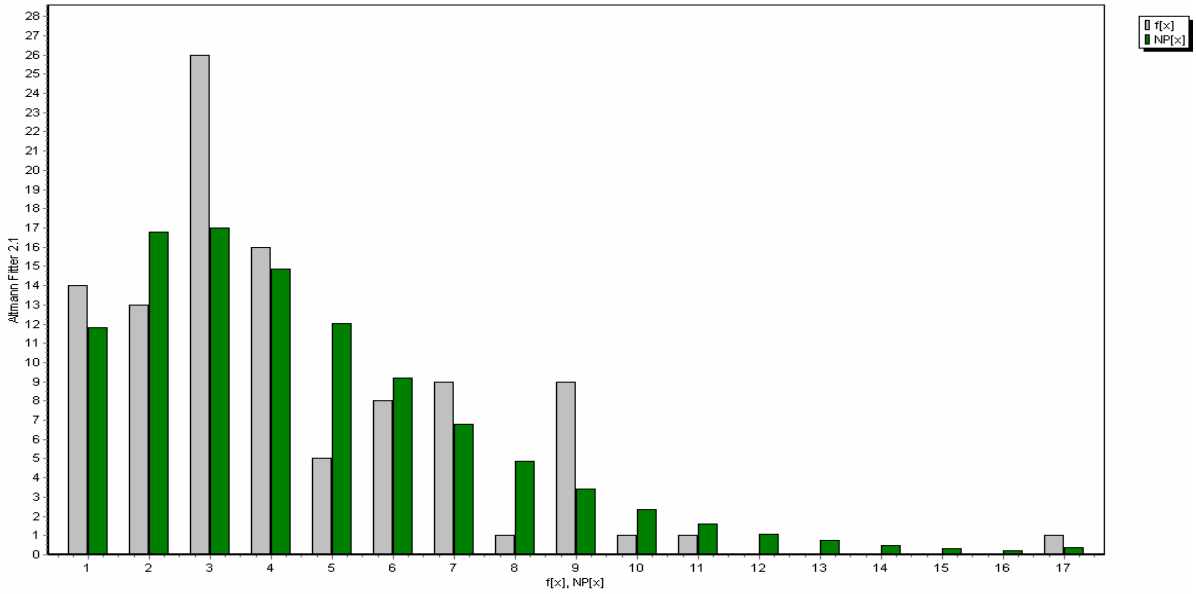

Abbildung 8: Anpassung der 1-verschobenen negativen Binomialverteilung an die Verteilung der Satzlänge im Text 31

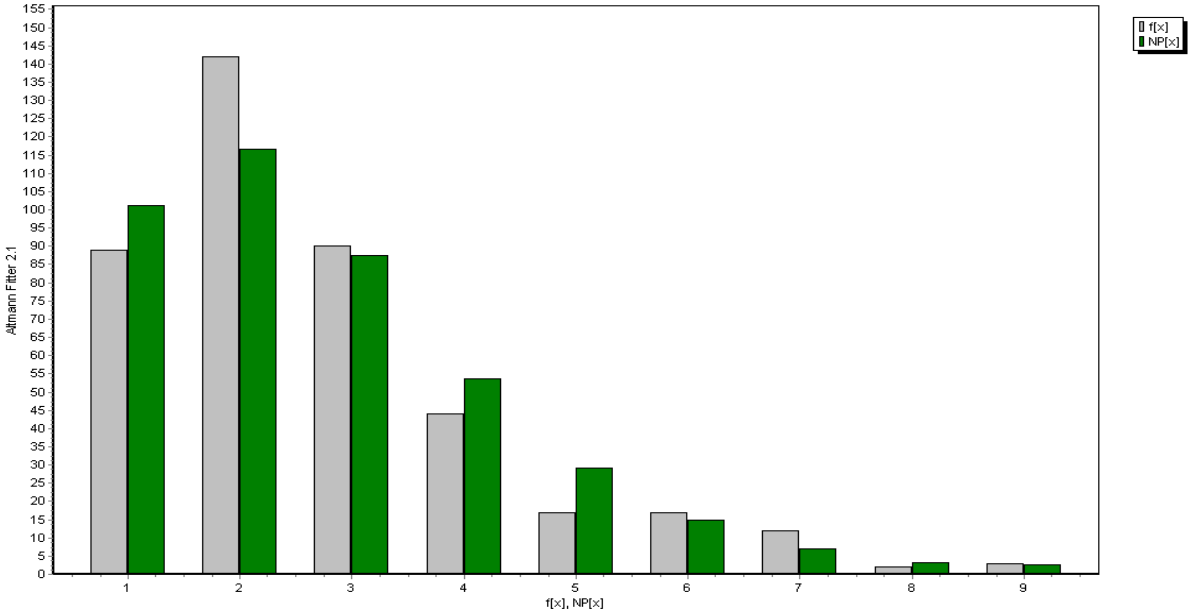

Abbildung 9: Anpassung der 1- verschobenen negativen Binomialverteilung an die Verteilung der Satzlänge im Text 45

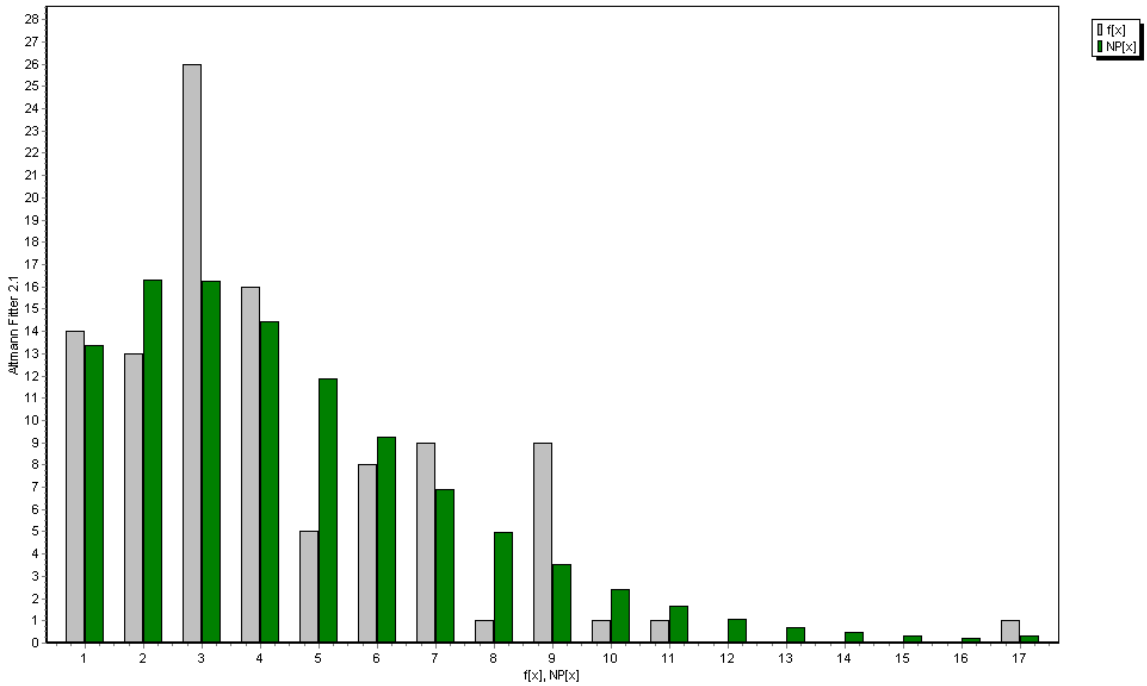

Abbildung 10: Anpassung der 0-gestutzten negativen Binomialverteilung an die Verteilung der Satzlänge im Text 31

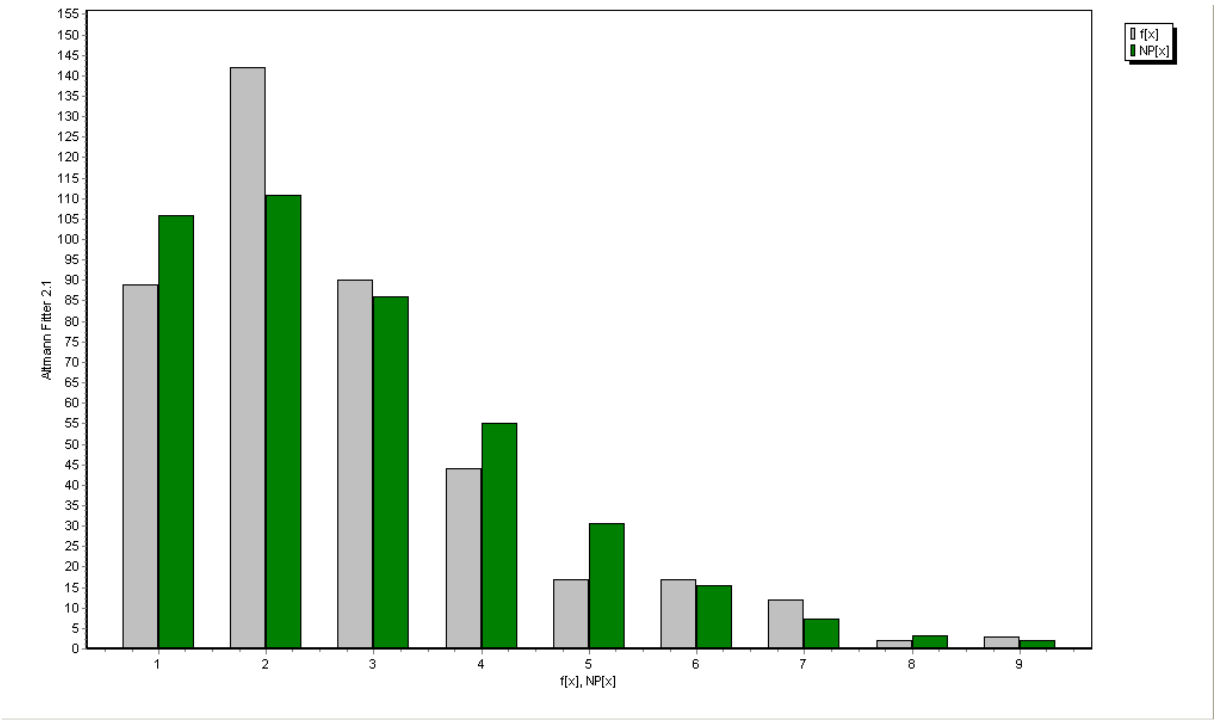

Abbildung 11: Anpassung der 0-gestutzten negativen Binomialverteilung an die Verteilung der Satzlänge im Text 45

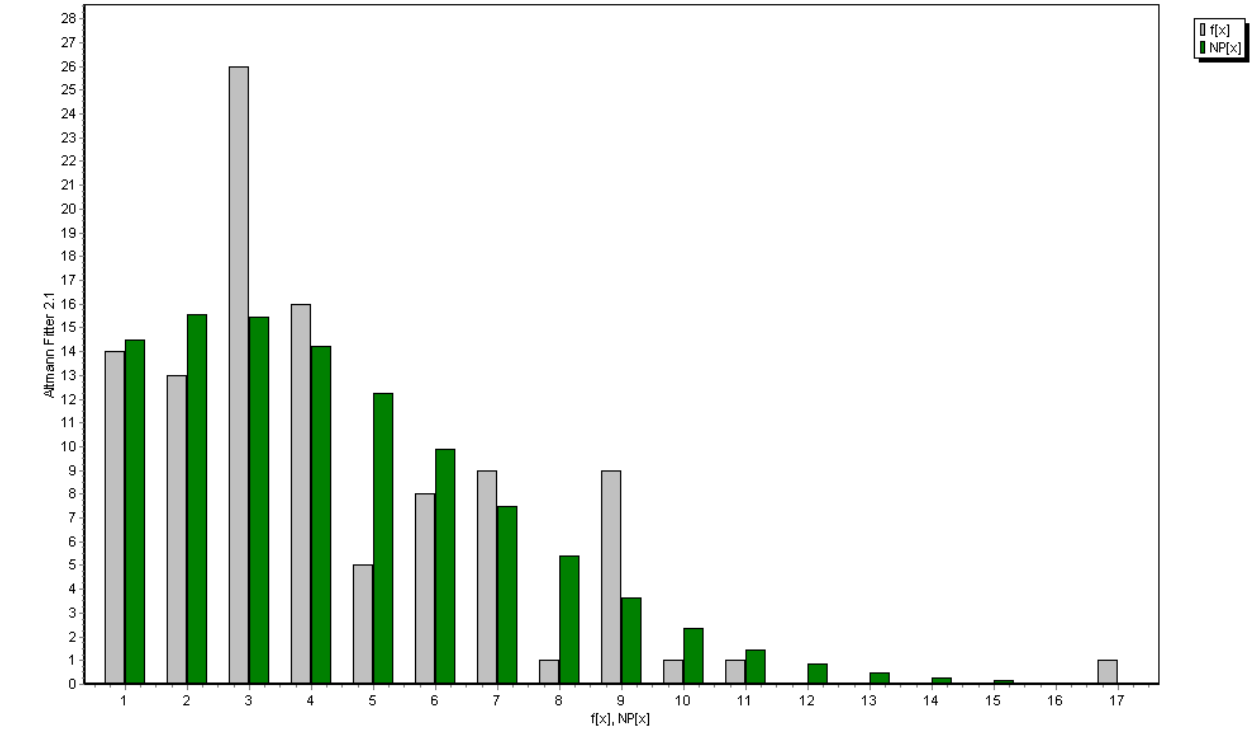

Abbildung 12: Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Verteilung der Satzlänge im Text 31

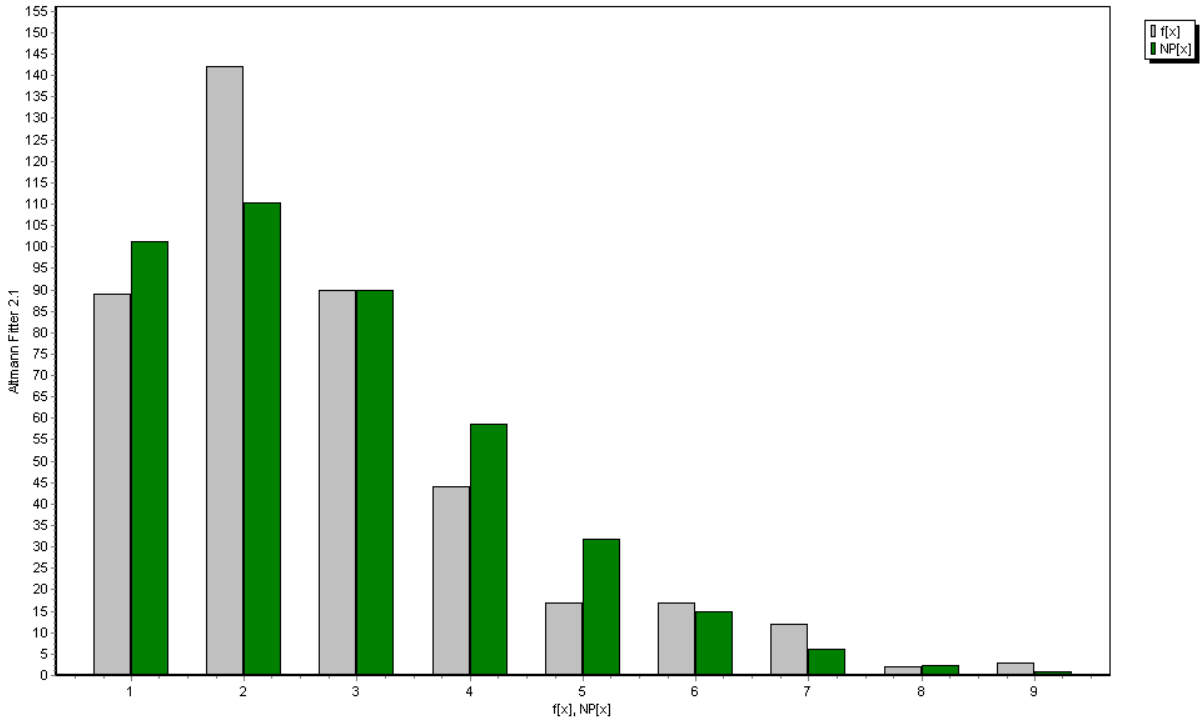

Abbildung 13: Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Verteilung der Satzlänge im Text 45

5.5. Häufigkeitsverteilung in Textblöcken

Die Hypothese, dass linguistische Einheiten in Textblöcken bestimmter Länge einer bestimmten Gesetzmäßigkeit folgen, wird im Folgenden überprüft. Untersucht wird die Verteilung von Wortarten in 100-Wort-Textblöcken.

Es existieren verschiedene Verfahren zur Wortartenklassifikation. In der quantitativen Linguistik finden hauptsächlich syntaktisch-distributionelle und morphologische Ansätze Anwendung. Die Zuordnung von Texttoken zu Wortartenklassen basierte hier auf syntaktisch-distributionellen Kriterien. Berücksichtigt wurden nur Wörter, die zu der Kategorie *Adjektiv* gehören.

Als Datengrundlage zur Überprüfung der Hypothese wurden die Texte 27 bis 33; 35, 37, 40 und 46 verwendet. Jeder Text wurde in Textblöcke mit je 100 Wörtern geteilt und innerhalb jedes Blocks wurde das Vorkommen der Wortart *Adjektiv* ausgezählt. An die so ermittelten Häufigkeitslisten wurde in Anlehnung an ALTMANN & BURDINSKI (1982) die negative hypergeometrische Verteilung angepasst. Eine Zusammenfassung der Testergebnisse enthält Tabelle 5.19. Eine Gegenüberstellung der empirischen und theoretischen Werte jedes Textes kann dem Anhang, Tabelle 8.1, entnommen werden. Dabei bedeutet x = 0 die Textblöcke, in

denen kein Adjektiv vorkommt; x = 1 meint Textblöcke, in denen ein Adjektiv vorkommt; usw.

Tabelle 5.19: Anpassung der negativen hypergeometrischen Verteilung an das Vorkommen von Adjektiven in 100-Wort-Textblöcken

Texte	K	M	n	X^2	DF	$P(X^2)$
27	6.1873	2.2611	18	12.3248	12	0.42
28	4.3696	2.1854	15	9.1739	10	0.5157
29	7.3995	2.8882	20	10.8408	12	0.5426
30	17.9738	9.1841	10	3.3389	3	0.3423
31	4.3095	1.6950	8	11.9624	4	0.0176
32	3.9008	1.4866	14	3.3943	4	0.4941
33	14.5486	5.7091	14	7.5965	7	0.3695
35	18.1347	6.3785	14	10.0319	6	0.1233
37	11.1176	4.1717	15	8.0942	7	0.3244
40	22.17923	6.4073	15	5.2200	6	0.5159
46	36.8660	12.5341	16	6.6657	4	0.1546

Anders als in den Texten 33 und 35 wurde in allen anderen Stichproben die negative hypergeometrische Verteilung in der 1-verschobenen Form angepasst, da kein Textabschnitt vorlag, in dem die Wortart *Adjektiv* nicht vorkam.

Tabelle 5.19 kann man entnehmen, dass das angepasste Modell sich in 10 von 11 Textdateien als geeignet erweist. Der in diesen Dateien von dem χ^2-Test gelieferte *P*-Wert liegt im Intervall $0.5426 \geq P \geq 0.1233$. Es deutet sich folglich an, dass die negative hypergeometrische Verteilung ein gutes Modell für das Vorkommen von Adjektiven in Textblöcken ist.

Für Text 31 führte die Anwendung dieser Verteilung zu einem unbefriedigenden Ergebnis. Die Anpassung ergab mit $P = 0.0176$ kein gutes Ergebnis, das sich auch nicht nennenswert verbessern ließ: Eine Anpassung ohne Zusammenfassung der Häufigkeitsklassen x = 9 und x = 10 erbrachte lediglich ein $P = 0.0309$. Dies muss aber nicht als Modifikation des Modells angesehen werden, denn

> „eine Einheit kann in einem Text nach dem Grundmodell verteilt sein, während sie in einem anderen Text ein bestimmtes Grenzverhalten, in einem dritten ein anderes Grenzverhalten aufweist.“ (ALTMANN 1988, 177)

Es sei hier noch angegeben, was die Anpassung der Grenzfälle an die einzelnen Textdateien ergab:

Die Anpassung der negativen Binomialverteilung an die empirischen Daten ergab bessere Ergebnisse; der von dem χ^2-Text gelieferte *P*-Wert liegt in allen Dateien über der Signifikanzschwelle. Es muss aber angemerkt werden, dass die Anpassung des Modells an die Textdatei

30 ohne Zusammenfassung der Häufigkeitsklassen durchgeführt wurde. Die Ergebnisse der Binomialverteilung und der Poisson-Verteilung sind im Vergleich zu den Resultaten der negativen Binomialverteilung etwa schlechter: Beide Modelle konnten an Daten von 3 Texten nicht erfolgreich angepasst werden.

In den Tabellen 5.20 und 5.21 sind die Anpassungsergebnisse der Grenzfälle zusammenfassend aufgeführt.

Tabelle 5.20: Anpassung der Binomial- und negativen Binomialverteilung an das Vorkommen von Adjektiven in 100-Wort-Textblöcken

Texte	**Binomialverteilung**					**Negative Binomialverteilung**				
	n	p	X2	DF	P(X2)	k	p	X2	DF	P(X2)
27	6522	0.0009	41.8522	7	0	4.9135	0.4206	13.2505	13	0.4287
28	7887	0.0010	8.2524	8	0.4092	20.1753	0.7205	5.1161	8	0.7451
29	7858	0.0010	24.2512	9	0.0039	7.9688	0.5112	9.0239	13	0.7711
30	268	0.0194	2.9916	5	0.7013	*10.7266*	*0.6937*	*7.3540*	*8*	*0.4990*
31	3229	0.0010	14.2411	5	0.0141	7.3925	0.6904	11.9999	6	0.0620
32	5398	0.0010	1.2486	4	0.87	5.7979	0.5310	1.4419	5	0.9197
33	5467	0.0009	10.4550	8	0.2345	29.3603	0.8422	9.3895	9	0.4021
35	4949	0.0010	10.0605	7	0.1852	18.0295	0.7798	10.3174	8	0.2435
37	5944	0.0010	6.9741	7	0.4316	15.5940	0.7244	6.2557	9	0.7141
40	4282	0.0009	5.8112	6	0.4447	7.5312	0.6447	7.6611	8	0.4673
46	5948	0.0010	6.5221	6	0.3673	15.2965	0.7449	3.9303	6	0.6861

Tabelle 5.21: Anpassung der Poisson-Verteilung an das Vorkommen von Adjektiven in 100-Wort-Textblöcken

Texte	**Poisson-Verteilung**			
	a	**X2**	**DF**	**P(X2)**
27	3.7849	173.3022	7	0
28	7.9713	8.2317	9	0.5110
29	8.1140	31.3685	11	0.0010
30	5.2021	3.0117	6	0.8074
31	3.3141	14.2131	6	0.0273
32	2.3599	18.0711	3	0.0004
33	5.4433	10.4401	9	0.3160
35	4.9666	10.0517	8	0.2614
37	6.0321	6.9603	8	0.5409
40	4.2318	5.8006	7	0.5632
46	6.0024	6.5230	7	0.4802

Die bisherigen Ergebnisse wurden ausschließlich an Einzeltexten gewonnen. Um zu prüfen, ob die Gesetzmäßigkeit des Auftretens von Adjektiven in Textblöcken auch im Korpus nachweisbar ist, wurden alle hier untersuchten Texte in ein Textkorpus gemischt. Dieses Korpus wurde dann in Textblöcke mit je 100 Wörtern geteilt. Es entstanden so 496 Textblöcke. An das Vorkommen der Adjektive in diesen Textblöcken wurden die negative hypergeometri-

sche Verteilung und deren Grenzfälle angepasst. Die Anpassung des Grundmodells ergab mit $P = 0.0151$ ein schlechtes Ergebnis, das sich auch nicht verbessern ließ. Von den Grenzfällen konnte nur die negative Binomialverteilung mit Erfolg angewendet werden; die Anpassung ergab mit $P = 0.3807$ ein zufrieden stellendes Ergebnis. Damit deutet sich an, dass die Gesetzmäßigkeit des Vorkommens linguistischer Einheiten in Textblöcken auch in Korpora nachzuweisen sein wird.

In Tabelle 8.2 im Anhang finden sich die empirischen und die entsprechenden theoretischen Werte.

5.6. Häufigkeitsverteilung der Komplexität und Frequenz von Teilsätzen

5.6.1. Die Dependenzgrammatik als Bezugs- und Beschreibungsrahmen

Der vorliegende Überblick bezieht sich in erster Linie auf die Dichotomie Aktanten - Angaben (Circumstanten) in der Dependenzgrammatik von TESNIÈRE.

Wesentlich für die Dependenzgrammatik (DG) sind nicht nur Wörter bzw. Nuclei (Knoten), sondern auch die Dependenzrelationen zwischen den einzelnen Knoten. Die DG geht davon aus, dass syntaktische Einheiten hierarchisch strukturiert sind, d.h. Sätze werden als Gefüge von Relationen betrachtet. Als Grundrelation gilt die Konnexion bzw. Dependenz: Es gibt im Satz eine Einheit, die der anderen unter- oder übergeordnet ist, wobei die einzelnen Einheiten in Beziehung zueinander stehen. Die Einheiten können unterschiedlicher Art sein (Prädikat, Subjekt oder Objekt) und aus einzelnen (z.B. Verben, Substantive) oder komplexen (z.B. Relativsätze) Wörtern bestehen.

Von dem Verb hängen direkt oder indirekt alle anderen Einheiten ab, d.h. das Verb wird als die Einheit betrachtet, die den Satz regiert. Die Anzahl der im Satz auftretenden Einheiten ist auch durch das Verb bedingt. Es bestimmt somit die Satzstruktur schon im Vorhinein. Diese Fähigkeit des Verbs, die Satzstruktur vorherzubestimmen, wird von TESNIÈRE (1980, 161) als Valenz bezeichnet:

> „Man kann so das Verb mit einem Atom vergleichen, an dem Häkchen angebracht sind, so daß es – je nach Anzahl der Häkchen – eine wechselnde Zahl von Aktanten an sich ziehen und in Abhängigkeit halten kann. Die Anzahl der Häkchen, die ein Verb aufweist, und dementsprechend die Anzahl der Aktanten, die es regieren kann, ergibt das, was man die Valenz des Verbs nennt. “

Die dem Verb direkt untergeordneten Nuclei werden als Aktanten und Circumstanten bezeichnet. Aktanten bestimmt TESNIÈRE (1980, 93) wie folgt:

> „Der verbale Nexus [Knoten], der bei den meisten europäischen Sprachen im Zentrum steht,

lässt sich mit einem kleinen Drama vergleichen. Wie das Drama umfasst er notwendig ein Geschehen und meist auch noch Akteure und Umstände. Wechselt man aus der Wirklichkeit des Dramas auf die Ebene der strukturalen Syntax über, so entspricht dem Geschehen das Verb, den Akteuren die Aktanten und den Umständen die Angaben. [...] Die Aktanten sind Wesen oder Dinge, die auf irgendeine Art, sei es auch nur passiv, gewissermassen als blosse Statisten, am Geschehen teilhaben."

Aktanten sind somit Einheiten, die für die Handlung konstitutiv sind. Sie sind „Leerstellen, die von einem Verb eröffnet sind" (WEBER 1997, 34). Aktanten dienen dazu, die Bedeutung des Verbs zu vervollständigen und können nicht weggelassen werden. Circumstanten sind demgegenüber nicht valenzgebunden: Sie bezeichnen die näheren Umstände des Geschehens und sind somit nicht obligatorisch (vgl. ebd. S. 35).

TESNIÈRE (1980) lässt 3 Typen von Aktanten zu: 1. Aktant (Subjekt), 2. Aktant (Akkusativobjekt) und 3. Aktant (andere Objekte). Präpositionalgruppen sieht er nicht als Aktanten an, was in machen Fällen problematisch sein kann. Der folgende Satz dient der Verdeutlichung:

(1) L'archiduc résidait à Milan.

In diesem Satz kann die Präpositionalgruppe "à Milan" nicht ohne weiteres weggelassen werden. In semantischer Hinsicht ist sie obligatorisch und somit als Aktant zu betrachten. Unter syntaktischen Gesichtspunkten wäre sie laut der Argumentation von TESNIÈRE jedoch als Angabe einzustufen. Die Tesnière'schen Kriterien sind also nicht immer haltbar, auch nicht im Deutschen, wie die folgenden Sätze aus STORRER (1992, 57) zeigen:

(2) a. Der Cowboy steigt auf das Pferd.

b. Der Cowboy besteigt das Pferd.

Semantisch gesehen sind die PP "auf das Pferd" und die NP "das Pferd" als Ergänzungen zu betrachten. Syntaktisch ist laut TESNIÈRE nur die NP als Aktant anzusehen.

Um die Unzulänglichkeiten der Tesnière'schen Kriterien und das Problem der Festsetzung der Wertigkeit eines Verbs zu umgehen, wurde bei der Untersuchung der Verteilung der Komplexität und Frequenz von Teilsätzen nicht zwischen Aktanten und Circumstanten unterschieden. Folgende (Teil)Sätze sollen der Veranschaulichung dienen:

(3) a. La maman s'occupe brillamment des enfants.

b. La maman s'occupe des enfants.

In (3a) sind die Wörter "maman", "brillamment" und "enfants" direkt dem Verb untergeordnet. Dieser (Teil)Satz hat somit die Komplexität 3, während in (3b) nur 2 Wörter direkt vom Verbknoten abhängig sind. Der (Teil)Satz hat somit die Komplexität 2.

Als „strukturgleich" angesehen wurden Teilsätze mit ein und demselben Verb sowie der gleichen Komplexität. Dies möge das folgende Satzpaar zeigen:

(4) a. Il accueille ses parents.

b. Il accueille ses amis.

5.6.2. Häufigkeitsverteilung der Komplexität von Teilsätzen

Für die Untersuchung der Verteilung der Komplexität von Teilsätzen wurden 1505 Teilsätze aus den Texten 24, 35, 37, 43 und 45 ausgewählt. Anschließend wurde die Komplexität dieser Einheiten ermittelt und an die gewonnenen Daten in Anlehnung an KÖHLER & ALTMANN (2000) die Hyperpascal-Verteilung angepasst. Folgende Werte wurden dabei erzielt: k = 1.49, m = 0.06, q = 0.21, X^2_3 = 60.1604, bei einem P(X^2) = 0 bzw. C = 0.4. Im Anhang, Tabelle 9.1, werden die aus den Parametern errechneten theoretischen Werte den empirischen Werten gegenübergestellt.

Wie man sieht, lieferte die Anpassung kein zufrieden stellendes Ergebnis. Auf der Basis von *P* bzw. *C* stellt sich die Hyperpascal-Verteilung für die getesteten Daten nicht als geeignetes Modell dar. Dennoch ist die Anpassung nicht als völlig misslungen zu betrachten. Abbildung 14 zeigt bei zunehmender Komplexität zunächst einen starken Anstieg und weist dann einen raschen Abstieg in den Teilsatz-Häufigkeiten auf. Um diese Unregelmäßigkeiten auszugleichen, wurden die erarbeiteten Daten durch Zusammenfassungen "geglättet"; es wurde eine Zusammenfassung der Teilsätze mit der Komplexität 1-2 zu Klasse 1, die mit der Komplexität 3-4 zu Klasse 2, etc durchgeführt. An die so erhaltenen Daten wurde erneut die Hyperpascal-Verteilung angepasst. Die Anpassung erfolgte jedoch nicht unter der Voraussetzung, dass die theoretische Frequenz einer Klasse mindestens bei 1 liegt. Außerdem wurde den Daten eine neue Klasse, Klasse 5, angehängt, damit die Anpassung möglich wird. Dadurch wurde ein Freiheitsgrad gewonnen. Die Anpassung ergab in diesem Fall ein deutlich besseres Ergebnis. Folgende als (sehr) gut zu bewertende Werte wurden dabei ermittelt: X^2_1 = 0.1278, P(X^2) = 0.7208, C = 0.0001.

Abbildung 15 zeigt die Ergebnisse in anschaulicher Form.

Für die Modellierung der Komplexität von Teilsätzen wurde die Hyperpascal-Verteilung deshalb herangezogen, weil sie sich bei den quantitativen Analysen zur Syntax in KÖHLER & ALTMANN (2000) bewährt hatte. Da sich Dependenzgrammatik und Phrasenstrukturgrammatik unterscheiden, wurden mittels einer automatischen Anpassung auch andere Verteilungen an die Daten angepasst. Das beste Ergebnis ergab sich dabei für die 1-*verschobene Dacey-Poisson-Verteilung*

$$(19) \qquad P_x = \frac{(1-\alpha)\, a^{x-1} e^{-a}}{(x-1)!} + \frac{\alpha(x-1) a^{x-2} e^{-a}}{(x-1)!}, \; x = 1, 2, \cdots n,$$

die eine Art gemischter Poisson-Verteilung darstellt. Folgende Parameterwerte wurden ermittelt: a = 0.61, $\alpha = 0.84$, bei einem sehr guten Diskrepanzkoeffizienten von $C = 0.0067$.

Man kann annehmen, dass sich auf die Verteilung der Komplexität von Teilsätzen zusätzliche Faktoren auswirken. Tabelle 9.2 im Anhang stellt den empirischen Werten der Untersuchung die theoretischen Werte gemäß der Dacey-Poisson-Verteilung gegenüber. Abbildung 16 zeigt diese Werte in anschaulicher Form.

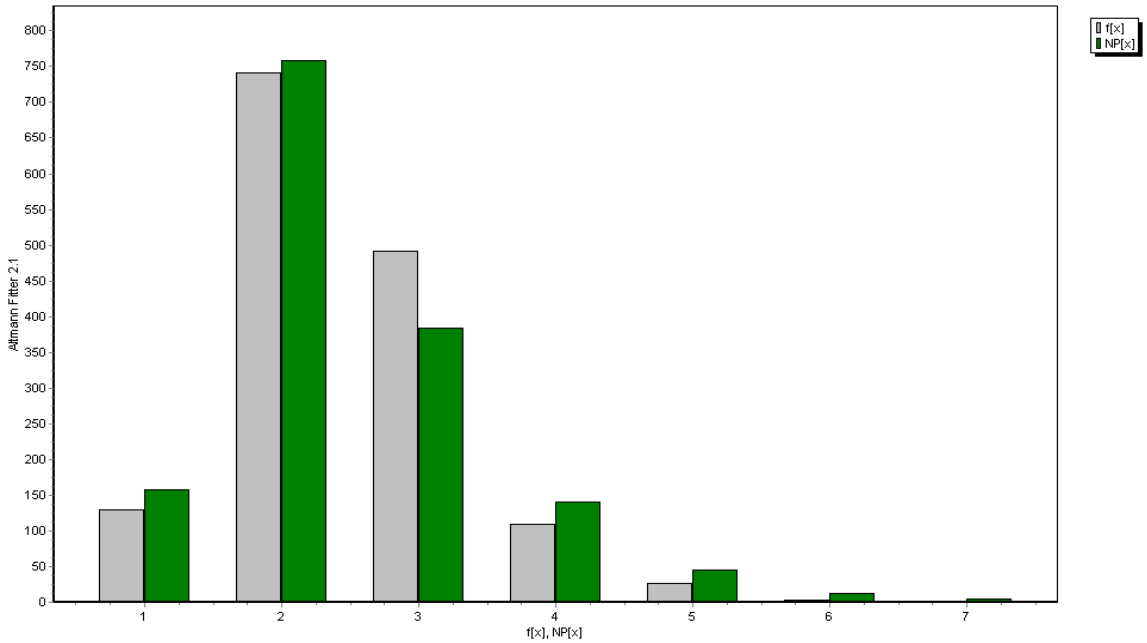

Abbildung 14: Anpassung der Hyperpascal-Verteilung an die Verteilung der Komplexität von Teilsätzen

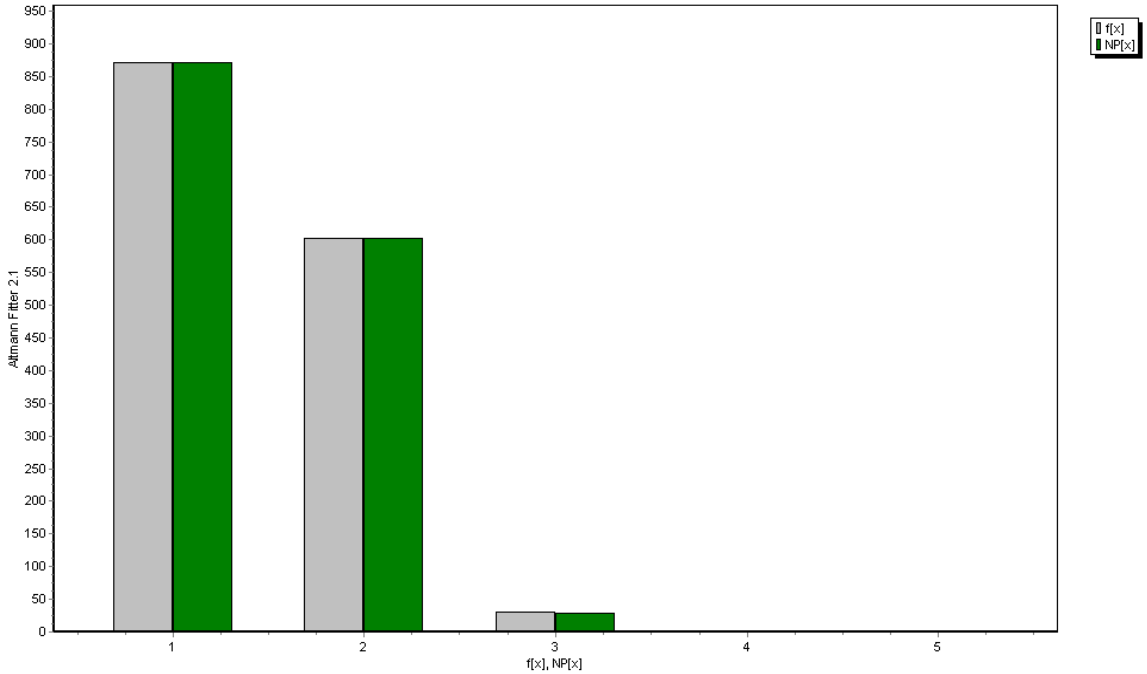

Abbildung 15: Anpassung der Hyperpascal-Verteilung an die Komplexität von Teilsätzen (Zusammenfassung von Komplexität zu Klassen)

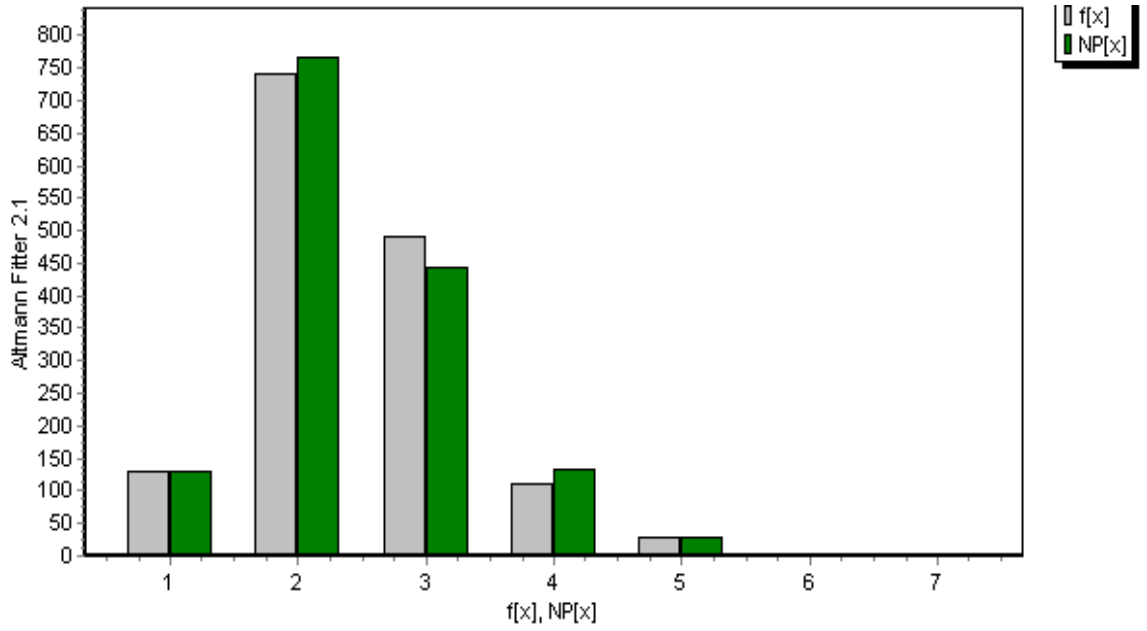

Abbildung 16: Anpassung der Dacey-Poisson-Verteilung an die Komplexität von Teilsätzen

5.6.3. Häufigkeitsverteilung der Teilsätze

In den vorausgehenden Ausführungen wurde die Verteilung der Komplexität von 1505 Teilsätzen untersucht. Im Folgenden soll der Frage nachgegangen werden, wie die Frequenz dieser Einheiten verteilt ist. Dabei wird zwischen Rangverteilung und Frequenzspektrum unterschieden.

Zur Untersuchung der Rangfrequenzverteilung der 1505 Teilsätze wurde zunächst das Vorkommen jedes Teilsatzes gezählt, wobei Teilsätze mit (1) ein und demselben Verb und (2) der gleichen Komplexität eine Einheit bildeten. Anschließend wurde die so ermittelte Teilsatz-Liste nach absoluter Häufigkeit geordnet und daraus eine Liste mit Rangnummer und absoluter Frequenz erstellt.

Eine Durchführung der Anpassung der Zipf-Mandelbrot-Verteilung an die erzeugte Ranghäufigkeitsverteilung führte zu hervorragenden Ergebnissen. Folgende Werte für die Parameter der Verteilung wurden dabei ermittelt:

$a = 0.67$, $b = 4.92$, $n = 881$, $X^2_{683} = 108.7106$, bei einem $P(X^2) \approx 1$.

Auch die Gegenüberstellung des theoretischen Modells und der empirischen Daten in Abbildung 17 spiegelt die gute Übereinstimmung der Verteilungen wider. Der Kurvenverlauf entspricht den Erwartungen. Die Angemessenheit des Zipf-Mandelbrot-Gesetzes kann somit für die untersuchte Häufigkeitsverteilung vorläufig angenommen werden.

Die Rangverteilung der 1505 Teilsätze wurde in einem zweiten Schritt in eine Klassenhäufigkeitsverteilung umgewandelt und mit der Waring-Verteilung modelliert. Für die Parameter dieser Verteilung wurden folgende Werte ermittelt: b = 2.13, n = 0.85, bei einem *P = 0.7391* bzw. *C = 0.0117*. Somit kann die Waring-Verteilung für die Modellierung der Frequenzbelegung von Teilsätzen in französischen Texten vorläufig als geeignet angesehen werden. Diese Annahme wird auch durch die Abbildung 18 gestützt.

Im Anhang, Tabellen 10.1 und 10.2, werden die empirischen Häufigkeitsverteilungen und die gemäß der Zipf-Mandelbrot- und Waring-Verteilung errechneten theoretischen Werte aufgeführt.

Den Befunden, vor allem dem Frequenzspektrum, kommt eine praktische Bedeutung zu, z.B. beim Fremdsprachenunterricht. Von den 881 verschiedenen Teilsätzen mit 1505 Vorkommen sind 633 Typen nur einmal belegt. Von den restlichen 248 Typen kommen 137 zweimal, von den dann übrigen 111 Typen 53 dreimal vor etc. Nur 58 Typen sind mehr als viermal belegt, was 6.58 % des Inventars repräsentiert. Die dargestellten Ergebnisse könnten bei der Strukturierung von Sprachlehrmaterial eine große Hilfestellung sein.

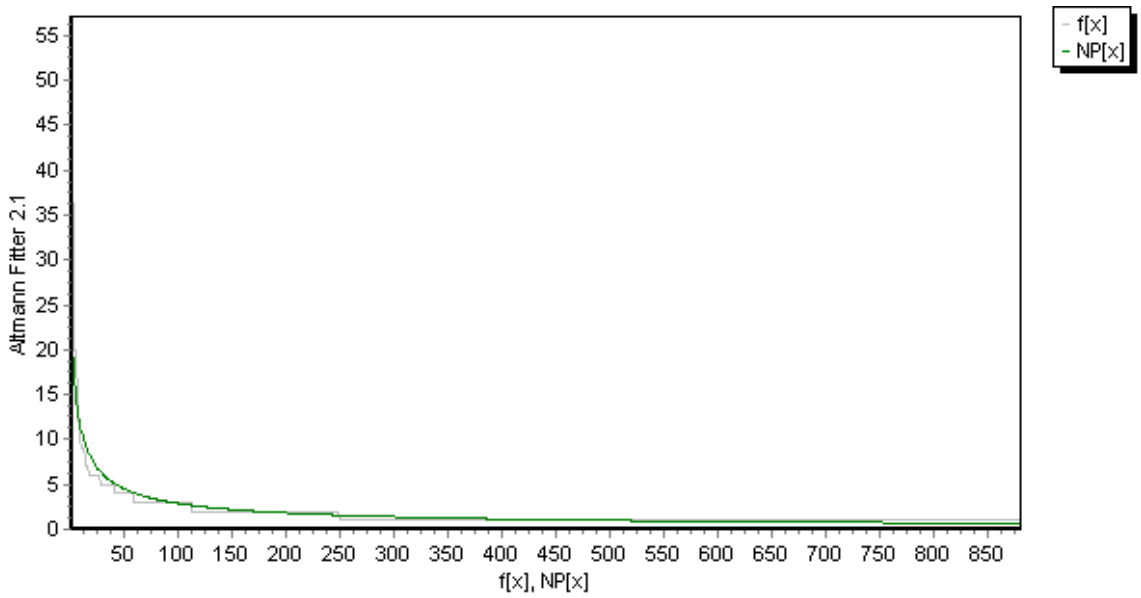

Abbildung 17: Anpassung der Zipf-Mendelbrot-Verteilung an die Ranghäufigkeitsverteilung von Teilsätzen

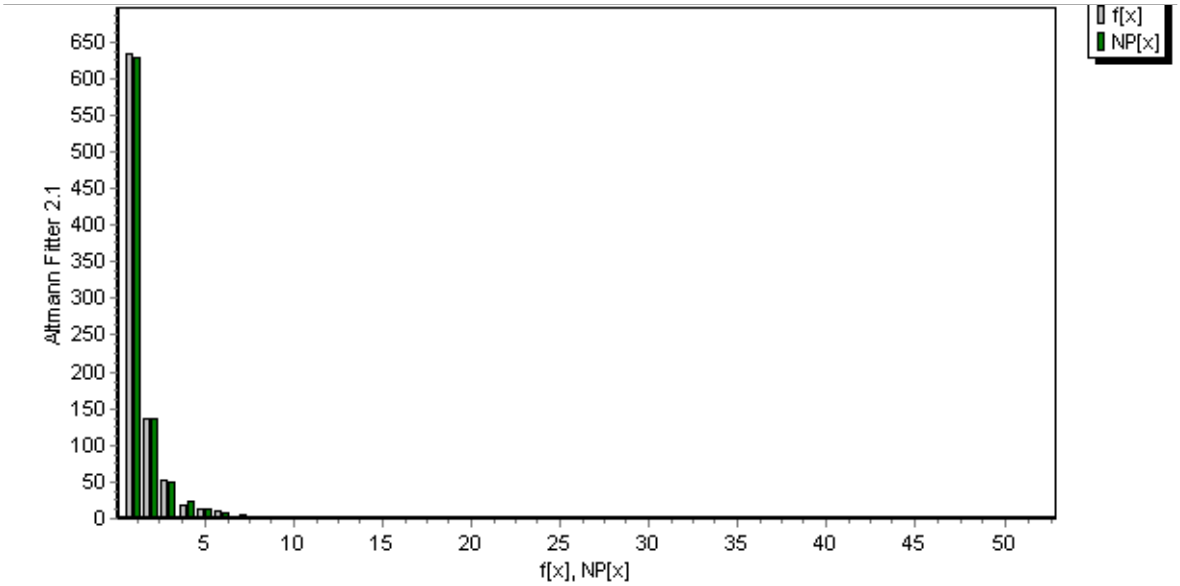

Abbildung 18: Anpassung der Waring-Verteilung an das Frequenzspektrum von Teilsätzen.

5.7. Zusammenfassung der Testergebnisse

Die Verteilung einiger Eigenschaften sprachlicher Einheiten wurde an französischen Texten und Korpora untersucht. Dabei konnte die allgemeine Hypothese, dass sprachliche Erscheinungen einer Gesetzmäßigkeit folgen, im Großen und Ganzen bestätigt werden.

Die bei der Verteilung der Bedeutungen von Affixen, der Satzlängenverteilung, der Ranghäufigkeitsverteilung von Wörtern sowie der Häufigkeitsverteilungen in Textabschnitten erzielten Ergebnisse sind als sehr gut bis befriedigend zu bewerten. Die aufgestellten Hypothesen können vorläufig als bestätigt angesehen werden, sollten aber nochmals an umfangreicheren Daten überprüft werden.

Die auf der Phonemebene erzielten Resultate stehen in deutlichem Gegensatz zu den Ergebnissen auf der Buchstabenebene: Während die Häufigkeitsverteilung der Phoneme mit zwei Modellen erfolgreich erfasst werden kann, lässt sich für die Modellierung der Häufigkeit von Buchstaben keine über alle Stichproben hinweg gültige Verteilung finden.

Die Anpassung der Hyperpascal-Verteilung an die Komplexität von Teilsätzen erbrachte wegen lokaler Unregelmäßigkeiten zunächst kein akzeptables Ergebnis. Eine Zusammenfassung der Teilsätze in Klassen und eine erneute Anpassung der Verteilung an die "geglätteten" Daten erwies sich dann auf Grund des erzielten Wertes von $C = 0.0001$ als sehr zufrieden stellend. Die automatische Anpassung an die Daten ergab einen sehr guten *C*-Wert für die 1-verschobene Dacey-Poisson-Verteilung.

6. Empirische Überprüfung der funktionalen Zusammenhänge

Die Ergebnisse der Anpassung der Funktionen an die empirischen Daten werden im Folgenden nur zusammenfassend wiedergegeben. Detaillierte Resultate können dem Anhang entnommen werden.

6. 1. Zusammenhänge zwischen den Eigenschaften Polylexie, Länge und Frequenz

Die funktionalen Zusammenhänge zwischen Polylexie und Länge, Polylexie und Frequenz und Frequenz und Länge von lexikalischen Einheiten sollen im Folgenden ein weiteres Mal empirisch überprüft werden. Die Überprüfung erfolgt an:

(a) einem Textkorpus. Dieses besteht aus allen Zeitungstexten, d.h. aus den Texten 1 bis 18 aus der Tabelle 1 im Anhang. Die durchschnittliche Länge dieser Texte liegt bei 916 Lexemen. Alle Artikel sind abgeschlossene Texte und entstammen verschiedenen Themenbereichen;

(b) einem Text: Analysiert wurden die Kapitel 1 bis 4 aus *„Les Trois Mousquetaires“* von A. Dumas[22]. Aus diesen Kapiteln wurden nur solche Lexeme berücksichtigt, die zu der Wortklasse *Substantiv*, *Adjektiv* oder *Verb* gehören. Eigennamen wurden nicht mitgezählt. Dadurch entstanden 5434 Wortformen bzw. 1600 Lemmata.

Bei der Überprüfung der Zusammenhänge an diesem Material wird in zwei Schritten verfahren: Zunächst werden alle Wortarten gemeinsam angepasst. In einem zweiten Schritt werden sie dann in Klassen eingeteilt und getrennt berechnet;

(c) dem Lexikon: Für die Berechnungen wurde eine dritte Stichprobe von 1000 Wörtern erstellt. Die Wörter wurden aus dem Wörterbuch *"Le Petit Robert Illustré 2007"* ausgewählt, wobei nur Wörter mit der Silbenlänge 1, 2, 3, 4 oder ≥ 5 berücksichtigt wurde. Pro Längenklasse wurden 200 Wörter ausgewählt.

Bei der Überprüfung der Zusammenhänge wird auf zweierlei Weise verfahren:

Erstens wird analog zu KÖHLER mit Mittelwerten der abhängigen Variablen gearbeitet. Da die Zuverlässigkeit der Mittelwerte davon abhängig ist, wie groß die ihnen zu Grunde liegenden Gruppen sind, wurden zahlenmäßig unterbelegte Klassengrößen von den Berechnungen ausgeschlossen, d.h.: Um Verzerrungen der Ergebnisse auf Grund zu weniger Ausgangsdaten zu vermeiden, wurden nur Klassengrößen mit mindestens 10 Belegen mit in die Berechnung einbezogen. Zweitens wird die Regression unter der Verwendung der Rohdaten durchgeführt. In den beiden Fällen wird die Funktionsgleichung: $y = ax^b$ an die Daten angepasst.

[22] Siehe Text 40

6.1.1. Zusammenhang zwischen Polylexie und Länge

a) Länge-Polylexie: Untersuchung im Korpus und Text

Bei der Überprüfung der Abhängigkeit der Wortpolylexie von der Wortlänge im Korpus und Text wurde die Eigenschaft *Länge* nur in Buchstabenanzahl gemessen. Für die Parameter **a** und **b** der angepassten Funktionsgleichung wurden bei der Untersuchung im Korpus folgende Werte ermittelt:

$$a = 11.4455$$

$$b = -0.5759$$

bei einem Determinationskoeffizienten von

$$R^2 = 0.8337$$

In Tabelle 11.1 im Anhang werden die aus diesen Parametern errechneten theoretischen Werte den empirischen Werten gegenübergestellt. Abbildung 19 zeigt anschaulich die empirisch gewonnenen Datenpunkte sowie die theoretische Kurve

$$y = 11.4455x^{-0.5759}$$

Der Determinationskoeffizient ist hier relativ zufrieden stellend. Daher kann für die untersuchte Stichprobe angenommen werden, dass die Abhängigkeit der Polylexie von der Länge bestätigt ist. Dies wird auch durch die empirisch gewonnenen Werte für die Polylexie gestützt, denn je länger ein Wort, desto geringer dessen mittlere Bedeutungszahl.

Der in Abbildung 19 von den empirischen Datenpunkten vermittelte optische Eindruck leistet aber keine Unterstützung für die angenommene Abhängigkeit. Die empirischen Daten führen zu keiner monoton fallenden Kurve, wie es das angepasste Modell voraussagt. Stattdessen steigt die Kurve an, um dann wieder abzufallen. Eine Erklärung für das Auftreten dieser nicht-monotonen empirischen Kurve, die dem Modell widerspricht, könnte der zu große Wert für das Vorkommen der Lexeme mit einer kleinen Anzahl von Bedeutungen sein. Es handelt sich bei diesen Lexemen in erster Linie um Funktionswörter, die im Wörterbuch nicht in so viele Bedeutungsnuancen unterschieden werden. Obwohl sie im Korpus mit einem großen Wert vertreten sind, besitzen Pronomina wie beispielsweise *il, sa, son* nur eine kleine Anzahl an Bedeutungen. Eine weitere Erklärung für diesen empirischen Kurvenverlauf, der den Erwartungen nicht entspricht, wäre die Oszillation der Lexik im Sinne von KÖHLER (1986). Darauf wird in den Abschnitten 6.1.2 und 6.1.5 noch mal zurückgekommen.

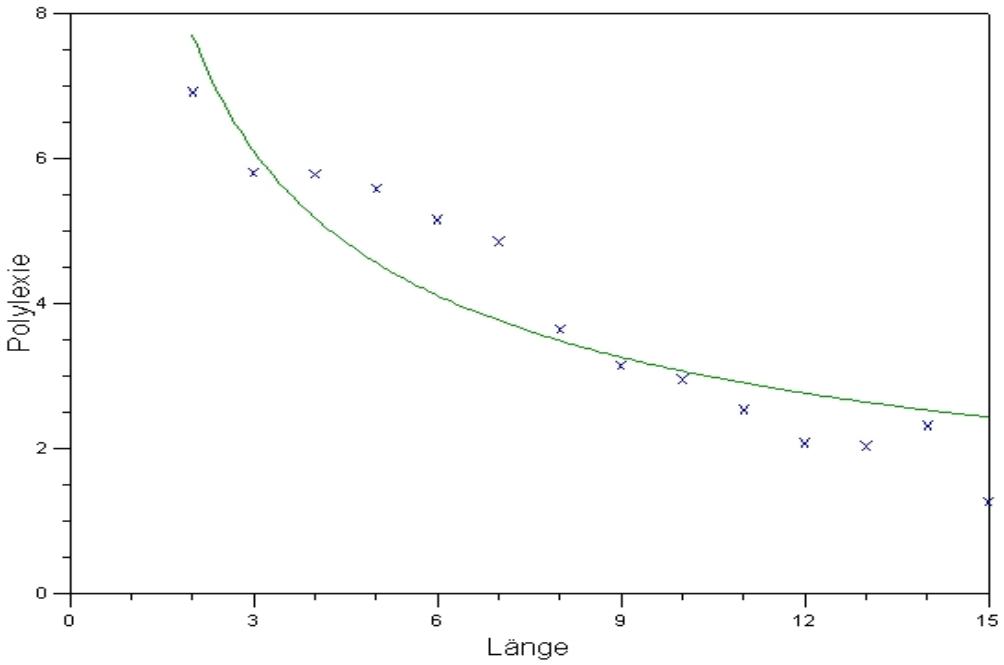

Abbildung 19: Abhängigkeit der Polylexie von Buchstabenlänge im Korpus

Die Untersuchung der Polylexie in Abhängigkeit von der Länge im Text erbrachte bei der gemeinsamen Betrachtung der Wortarten folgende Werte für die Parameter **a** und **b**:

a = 15.3487

b = -0.7662

bei einem sehr guten Determinationskoeffizienten von

$R^2 = 0.9863$

Die den empirischen Werten entsprechenden, aus den Anpassungsparametern errechneten theoretischen Werte kann man Tabelle 11.2 im Anhang entnehmen. Abbildung 20 stellt den empirisch ermittelten Datenpunkten die theoretische Kurve $y = 15.3487x^{-0.7662}$ gegenüber. Der Wert des Determinationskoeffizienten R^2 zeigt, dass die Anpassung hervorragend gelungen ist. Dies wird auch durch den von der empirischen Kurve vermittelten optischen Eindruck bestätigt: Der Kurvenverlauf entspricht vollkommen dem erwarteten Trend.

Sowohl auf der Basis des Determinationskoeffizienten als auch auf der Grundlage der empirischen Kurve kann hier die Abhängigkeit der Polylexie von der Länge als bestätigt angesehen werden.

Die Untersuchung dieser Abhängigkeit bei getrennter Berechnung der Wortarten führte zu den Ergebnissen in Tabelle 6.1. Für eine Gegenüberstellung der empirischen und theoretischen Werte sei auf Tabelle 11.3 im Anhang hingewiesen.

Die einzelnen Werte des Determinationskoeffizienten zeigen, dass die Abhängigkeit der Polylexie von der Länge auch für die einzelnen Wortklassen haltbar ist.

Diese Abhängigkeit wird auch durch die Abbildungen 21 bis 23 gestützt, welche die Daten und die theoretischen Kurven für alle drei Wortartklassen zeigen. An diesen graphischen Darstellungen ist eine relativ große Übereinstimmung zwischen empirischen und theoretischen Werten erkennbar. Bei Abbildung 23 kommt es im Bereich der kleinen und großen Längen zu einer relativ großen Streuung. Diese Schwankungen lassen sich anhand der relativ kleinen Klassengrößen erklären

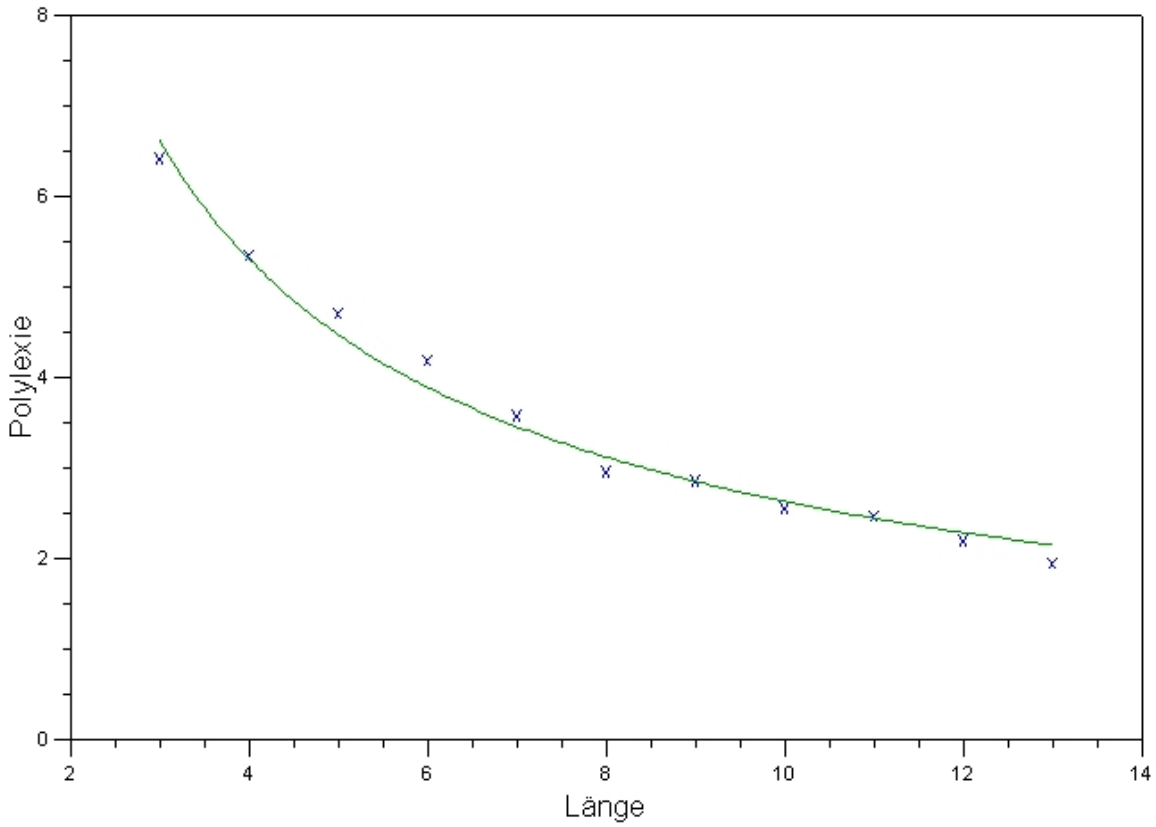

Abbildung 20: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (alle Wortarten)

Tabelle 6.1: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (einzelne Wortarten)

Wortarten	**a**	**b**	**R^2**
Substantive	15.1942	-0.7529	0.9791
Adjektive	16.6825	-0.8913	0.9358
Verben	13.6036	-0.6542	0.8456

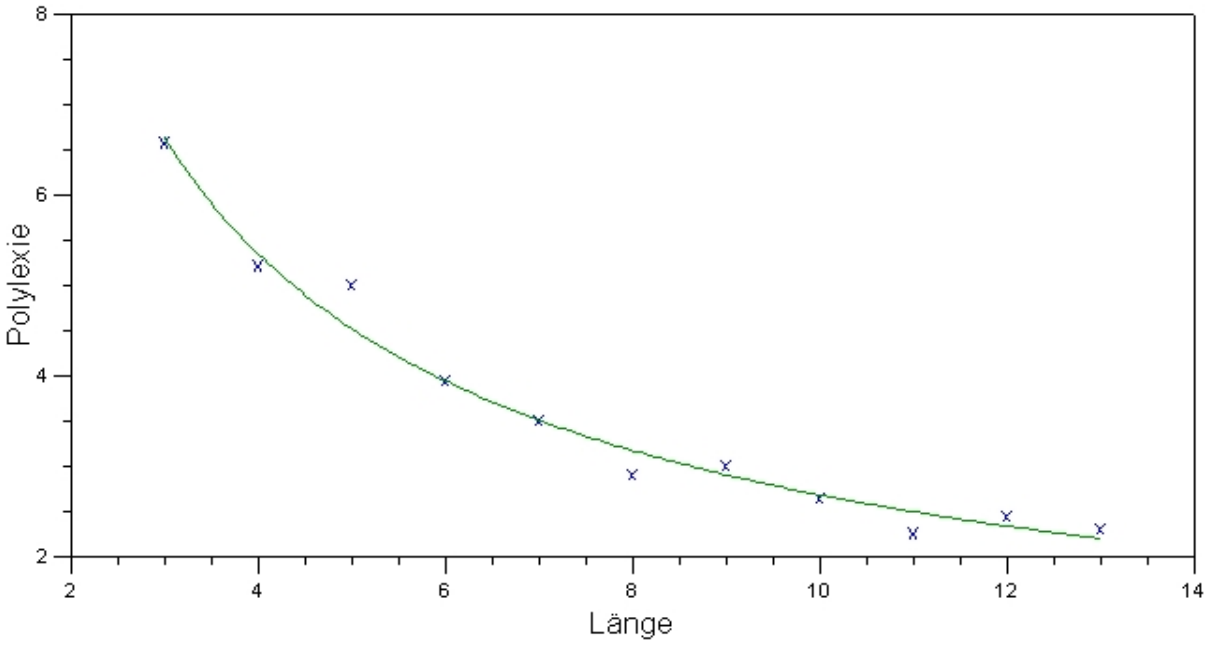

Abbildung 21: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (Substantive)

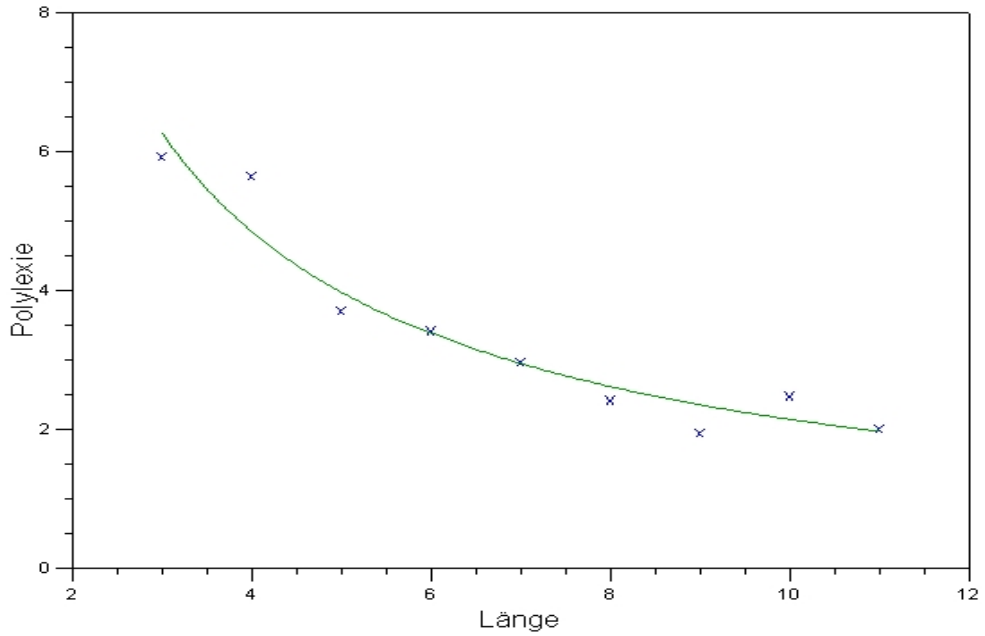

Abbildung 22: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (Adjektive)

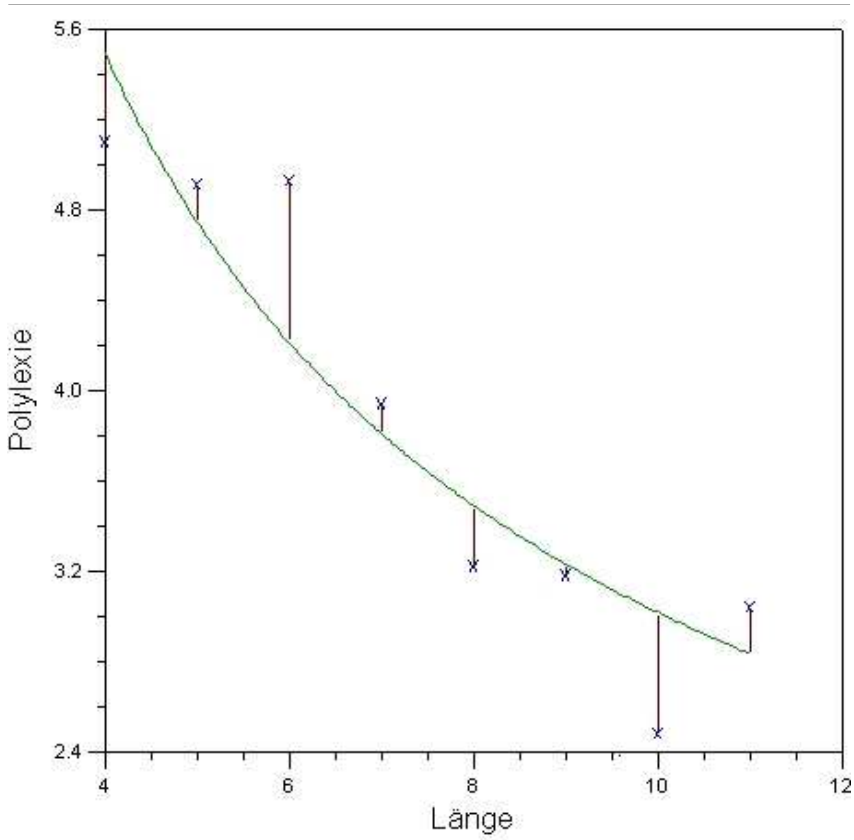

Abbildung 23: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (Verben)

b) Länge-Polylexie: Untersuchung im Lexikon

Bei der Untersuchung des Zusammenhangs zwischen Länge und Polylexie im Lexikon wurde die Länge in Buchstaben- und Silbenanzahl gemessen. Der Wortlänge wurde ihre jeweilige mittlere Polylexie gegenübergestellt.

Operationalisierung der Länge in Buchstaben

Die Anpassung der Funktion $y = ax^b$ an die empirischen Daten lieferte für die Parameter **a** und **b** folgende Werte:

$$a = 6.2439$$

$$b = -0.3463$$

bei einem Determinationskoeffizienten von

$$R^2 = 0.7144$$

Die den empirischen Ergebnissen entsprechenden theoretischen Werte sind in Tabelle 11.4 im Anhang aufgeführt. Abbildung 24 zeigt die empirischen Werte gemeinsam mit der theoretischen Kurve $y = 6.2439x^{-0.7144}$.

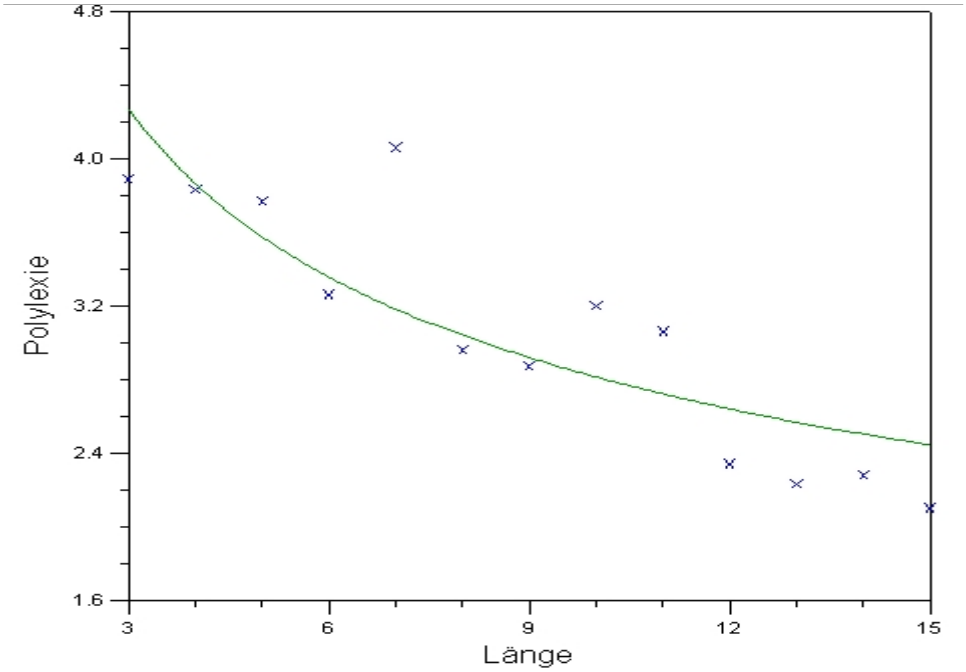

Abbildung 24: Abhängigkeit der Polylexie von der Buchstabenlänge im Lexikon

Der Wert des Determinationskoeffizienten ist hier im Vergleich zu den vorher erzielten Werten relativ niedrig; die Anpassung ist nicht so gut gelungen wie bei der Korpus- und Textuntersuchung. Dies wird durch den Kurvenverlauf der empirischen Datenpunkte unterstützt: Die Anpassung ist im Bereich der Längen mit niedrigen Werten (bis 6) sehr gut; die Tendenz zu fallender Polylexie bei steigender Länge ist hier erkennbar. Im Bereich der Längenwerte ab 7 bzw. 10 kommt es aber zu einer relativ großen Varianz. Daher erreichte der Determinationskoeffizient trotz des guten optischen Eindrucks der Kurve in deren ersten Teil keinen hoch signifikanten Wert.

Die Abweichung von einer gedachten Idealkurve scheint jedoch nicht erheblich, sodass die Abhängigkeit der Polylexie von der Länge als erwiesen betrachtet werden kann

Operationalisierung der Länge in Silben

Bei der Untersuchung der mittleren Polylexie in Abhängigkeit von der Silbenlänge wurden für die angepasste Funktionsgleichung folgende Parameterwerte ermittelt:

$$a = 4.1748$$

$$b = -0.3124$$

bei einem Determinationskoeffizienten von

$$R^2 = 0.8621$$

In Tabelle 11.5 im Anhang werden die aus diesen Parametern errechneten theoretischen Werte den empirischen Werten gegenübergestellt. Abbildung 25 zeigt die empirisch gewonnenen Datenpunkte zusammen mit der theoretischen Kurve $y = 4.1748\ x^{-0.3124.}$

Der Wert des Determinationskoeffizienten weist hier auf eine relativ gut gelungene Anpassung hin. Daher kann man die Abhängigkeit der Polylexie von der Silbenlänge als nachgewiesen ansehen.

Diese Annahme wird aber von der Abbildung 25 nicht voll gestützt. Diese graphische Darstellung zeigt, dass der optische Eindruck der Anpassung nicht so gut ist, wie es der R^2-Wert vermuten lässt. Die empirischen Datenpunkte liegen nur im Bereich der kleinen Längen sehr nahe an der theoretischen Funktion. Je größer die Längen werden, umso größer sind die Entfernungen zu der Funktion. Die Schwankungen bei großen Längenwerten lassen sich durch die relativ kleinen Klassengrößen erklären: Mit wachsender Länge werden weniger Wörter gefunden und deshalb wird die Streuung automatisch größer.

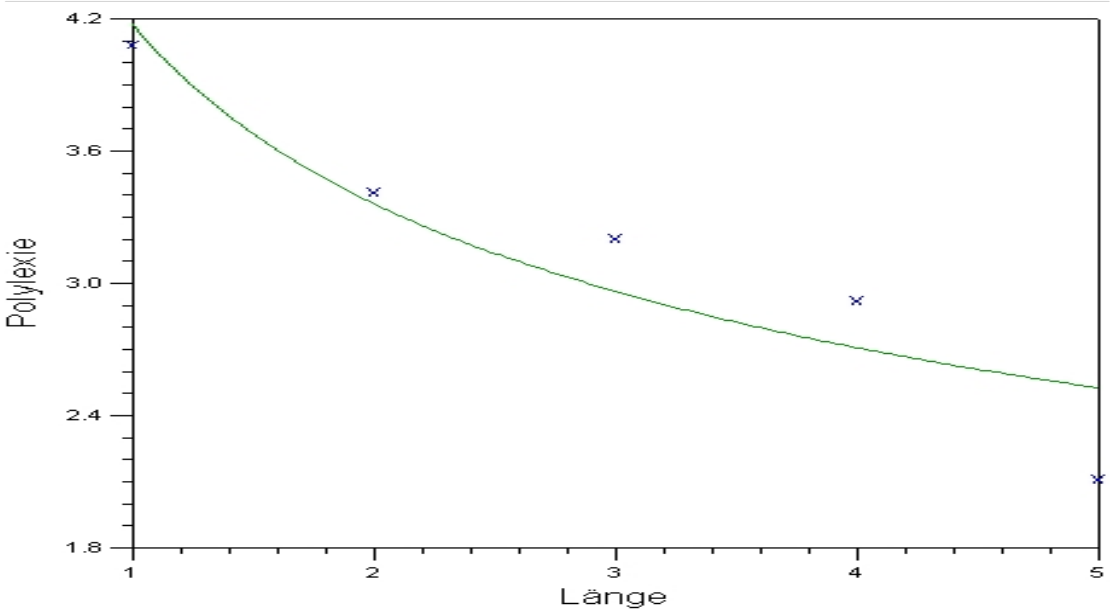

Abbildung 25: Abhängigkeit der Polylexie von der Silbenlänge im Lexikon

Die mittlere Polylexie in Abhängigkeit von der Länge wurde im Korpus, Text und Lexikon untersucht. Vergleicht man die Ergebnisse der Erhebungen, so kann man feststellen, dass die Untersuchung im Text die besten Ergebnisse darstellt. Dies ist eigentlich nicht erstaunlich, da die theoretische Funktionsgleichung an Daten angepasst wurde, die eine relativ große Homogenität aufweisen. Berücksichtigt wurden ja nur Lexeme, die zu einer bestimmten Wortklasse gehörten. Die Untersuchung im Lexikon führte bei der Buchstabenlänge zum schlechtesten R^2-Wert.

Es zeigen sich auch Unterschiede in den Parametern, die aus der sehr unterschiedlichen Datengrundlage resultieren und nur schlecht miteinander verglichen werden können.
In nachfolgender Tabelle werden alle Untersuchungsergebnisse zusammenfassend aufgeführt.

Tabelle 6.2: Abhängigkeit der Polylexie von der Länge

Länge	Untersuchung im Korpus			Untersuchung im Text (Alle Wortarten)			Untersuchung im Lexikon		
	a	b	R^2	a	b	R^2	a	b	R^2
Buchstabenlänge	11.4455	-0.5719	0.8337	15.3487	-0.7662	0.9863	6.2439	-0.3463	0.7144
Silbenlänge							4.1748	-0.3124	0.8621

6.1.2. Zusammenhang zwischen Länge und Frequenz

Bei der Überprüfung des Zusammenhangs zwischen Länge und Frequenz hat KÖHLER (1986) festgestellt, dass eine relativ starke Abweichung der empirischen Datenpunkte um die theoretische Kurve auftritt. Diese Datenstreuung führte er (S. 137ff.) auf die „Oszillation der Lexik" zurück, die in Form einer „um die theoretische hyperpolische Funktion schwingende[n] Kurve" an die Frequenzachse in Erscheinung tritt, wenn die Daten mit Hilfe gleitender Mittel geglättet werden.

Die Oszillation der Lexik lässt sich laut KÖHLER (1986, 144ff.) aus der Annahme einer Kürzungsrate lexikalischer Einheiten erklären, die von der Länge und Frequenz abhängt, d.h. aus der Annahme, dass ein Wort umso stärker gekürzt wird, je länger es ist.

Zur Beschreibung der von ihm festgestellten Oszillation der Länge in der Frequenz, die mit dem von ihm abgeleiteten Modell nicht beschrieben werden konnte, schlägt KÖHLER (1986, 144ff.) eine Differentialgleichung zweiter Ordnung vor, die aus der Annahme einer Kürzungsrate in die Länge-Frequenz-Relation abzuleiten ist. Eine solche Funktion wird in ZÖRNIG et al. (1990) abgeleitet und lautet:

$$L(F) = aF^b + a'F^{b'} + c\, e^{d(F-F_0)2} \sin(\alpha F)$$

Die Anpassung dieser Differentialgleichung an die Daten zur Länge-Frequenz-Beziehung, die KÖHLER (1986, 141) mit Hilfe der Methode gleitender Mittel (Intervall 50) erzeugte, liefert ein hervorragendes Ergebnis. Eine linguistische Interpretation für diese Funktion wird aber nicht geliefert.

In diesem Abschnitt wird nur die Abhängigkeit der Länge von der Frequenz lexikalischer Einheiten untersucht. Im Abschnitt 4.1.5. wird die von KÖHLER beobachtete Oszillation der Länge in der Frequenz überprüft.

a) Länge-Frequenz: Untersuchung im Korpus und Text

Hier wurde die Länge abermals nur in der Anzahl der Buchstaben gemessen. Die Anpassung der Funktionsgleichung $y = ax^b$ erbrachte für die Untersuchung im Korpus folgende Parameterwerte:

$$a = 8.6688$$

$$b = -0.1384$$

bei einem Determinationskoeffizienten von

$$R^2 = 0.6770$$

Eine Gegenüberstellung der empirischen und theoretischen Werte findet man in Tabelle 12.1 im Anhang.

Der Wert für den Determinationskoeffizienten ist hier etwas niedriger als der entsprechende Wert für die Relation zwischen Polylexie und Länge im Textkorpus. R^2 liegt nur bei 0.6770, was ein Hinweis darauf ist, dass der Anteil der erklärten Varianz der abhängigen Variablen relativ gering ist, d.h. die Anpassung ist nur mäßig gut.

Den niedrigen R^2-Wert kann man auf die Datenmenge zurückführen: Das untersuchte Korpus besteht aus verschiedenen Texten, die nicht einzeln, sondern gemeinsam bearbeitet wurden. Obwohl alle Texte der Textsorte *Zeitungsartikel* angehören, weist jeder Text individuelle Eigenschaften auf; der Zusammenhang zwischen Frequenz und Länge ist in jedem Text anders. Das Anwendungsbedürfnis für bestimmte Wörter ist in jedem Text unterschiedlich, was dazu führen kann, dass in einer Textdatei ein bestimmtes Lexem häufiger, in einem anderen jedoch seltener verwendet wird. Durch Addieren der Häufigkeiten ergibt sich wahrscheinlich eine schlechte Anpassung.

An der Abbildung 26, welche die empirisch ermittelten Daten zusammen mit der theoretischen Kurve $y = 8.6688\ x^{-0.1384}$ zeigt, ist im Vergleich zu den anderen Anpassungen eine auffallend große Streuung der Datenpunkte um den Graphen der Funktion zu erkennen. Diese Datenstreuung bestätigt die Beobachtungen von KÖHLER (19986). Sie wurde auch von KROT (2002) auf der Ebene der Morphe festgestellt, bei der Untersuchung der Abhängigkeit der Morphlänge von der Morph-Token-Frequenz.

Es ist aber auch zu erkennen, dass im Bereich der kleinen Frequenz-Werte, welche den größten Teil der Daten repräsentieren, die Datenpunkte nahe an der theoretischen Funktionsgleichung liegen.

Die Überprüfung der Abhängigkeit der Länge von der Frequenz im Text erbrachte keine besseren Ergebnisse. Für die Parameter **a** und **b** des angepassten theoretischen Modells wurden folgende Werte ermittelt:

a = 7.5091

b = -0.0804

Erzielt wurde ein Determinationskoeffizient von $R^2 = 0.6136$, was wiederum auf keine gute Anpassung hinweist. Die den empirischen Werten entsprechenden theoretischen Werte, die aus den Parameterwerten errechnet wurden, findet man im Anhang, Tabelle 12.2.

Abbildung 27 zeigt in anschaulicher Form die empirischen Daten gemeinsam mit der theoretischen Kurve $y = 7.5091x^{-0.0804}$. Der durch diese Abbildung vermittelte optische Eindruck zeigt keine gute Übereinstimmung der theoretischen Funktion mit den empirischen Daten: Im Bereich der mittleren Frequenz-Werte kommt es zu einer recht großen Streuung der empirischen Daten um die theoretische Funktion.

Eine nach Wortklassen homogenisierte Untersuchung der Stichprobe führte, wie Tabelle 6.3 entnommen werden kann, nur bei der Wortart *Adjektiv* mit $R^2 = 0.9174$ zu einem besseren Ergebnis. Der entsprechende empirische Kurvenverlauf (vgl. Abbildung 28) zeigt einen eindeutigen Trend. Abbildungen 29 und 30 geben die Daten und die theoretischen Kurven für die Wortarten *Substantive* und *Verben* wieder. An den beiden graphischen Darstellungen ist wieder eine auffallend große Streuung der Datenpunkte erkennbar, die bei Steigung der Frequenzwerte zunächst abnimmt (bei Substantiven bis zu der Frequenz 4 und bei Verben bis zu der Frequenz 2) sich bei hohen Werten jedoch verstärkt.

Als Ursache für die bei der Überprüfung der Abhängigkeit zwischen Länge und Frequenz im Korpus und Text festgestellten Abweichungen der empirischen Datenpunkte von der theoretischen Kurve sehen wir in Anlehnung an KÖHLER (1986) die Oszillation der Lexik. Darauf kommen wir im Abschnitt 4.1.5. der vorliegenden Arbeit zurück.

Tabelle 6.3: Abhängigkeit der Buchstabenlängen von der Frequenz im Text (einzelne Wortarten)

Wortarten	**a**	**b**	**R^2**
Substantive	7.4261	-0.0906	0.6976
Adjektive	29.5165	-1.5058	0.9174
Verben	8.0630	-0.0966	0.3959

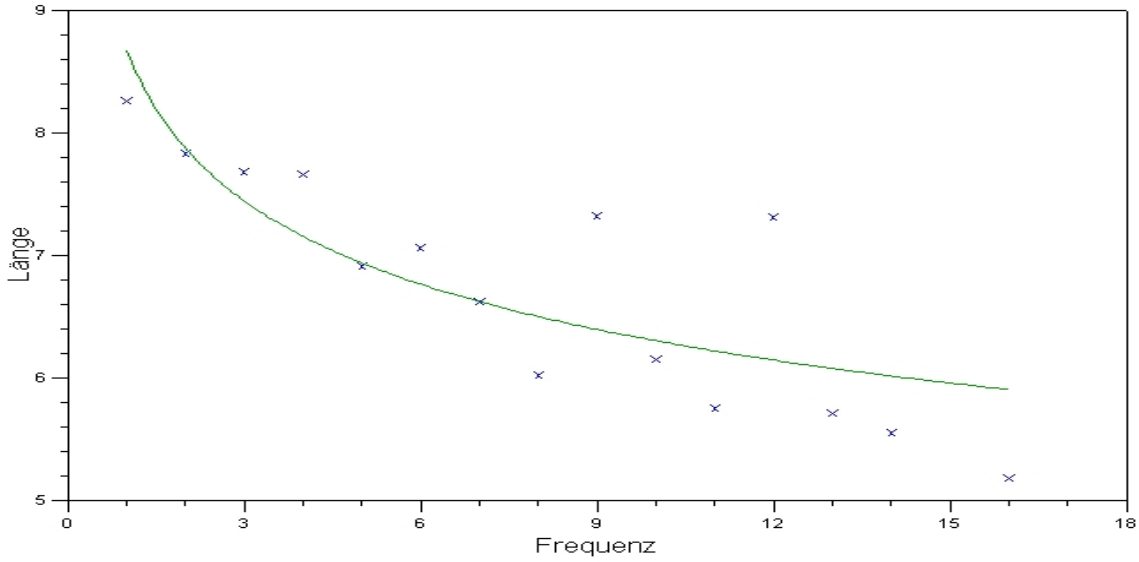

Abbildung 26: Abhängigkeit der Buchstabenlänge von der Frequenz im Korpus

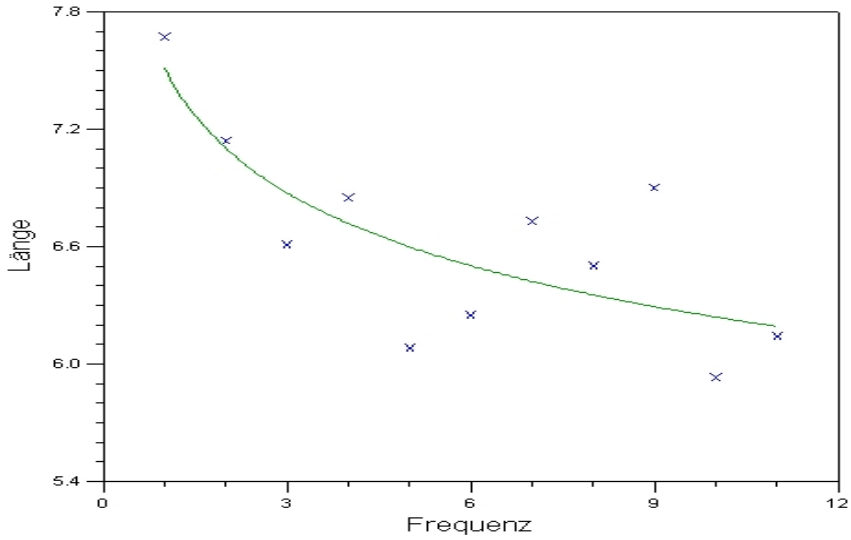

Abbildung 27: Abhängigkeit der Buchstabenlänge von der Frequenz im Text (alle Wortarten)

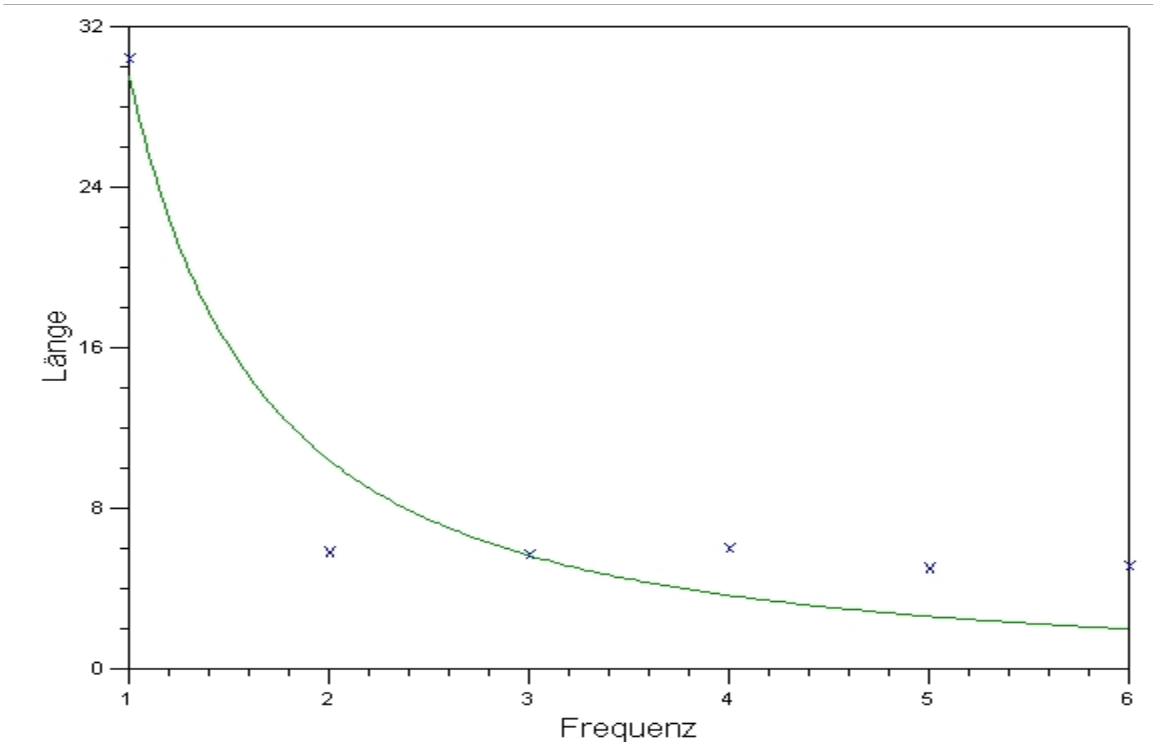

Abbildung 28: Abhängigkeit der Buchstabenlänge von der Frequenz im Text (Adjektive)

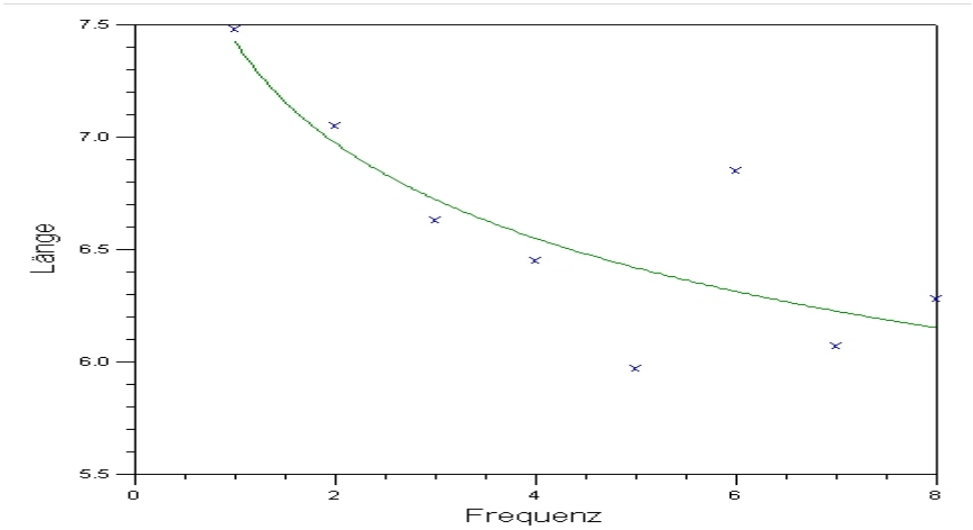

Abbildung 29: Abhängigkeit der Buchstabenlänge von der Frequenz im Text (Substantive)

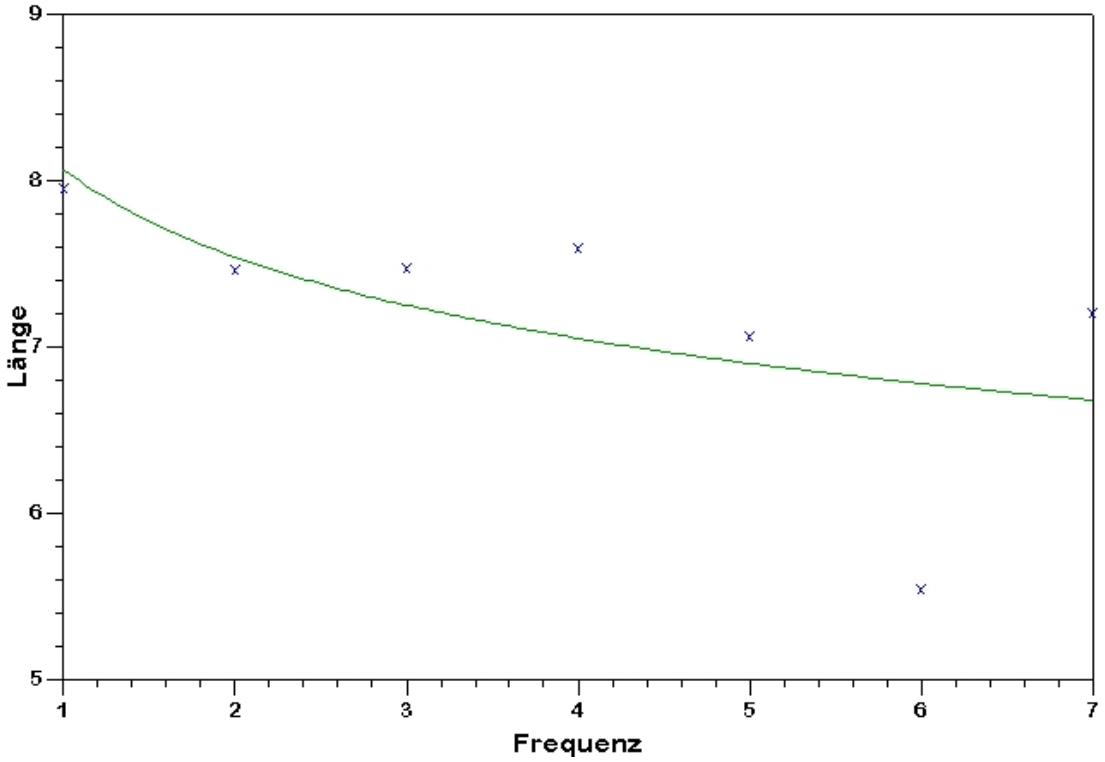

Abbildung 30: Abhängigkeit der Buchstabenlänge von der Frequenz im Text (Verben)

b) Länge-Frequenz: Untersuchung im Lexikon

Bei der Überprüfung der Abhängigkeit der Länge von der Frequenz im Lexikon wurde ebenso wie bei der Untersuchung der Beziehung zwischen Polylexie und Länge im Lexikon die Länge in Buchstaben- und in Silbenzahl gemessen. Zudem wurden nur die Frequenzwerte derjenigen Lemmata betrachtet, die auch im Frequenzwörterbuch enthalten waren, in welchem die Frequenz abgelesen wurde[23]. Die Untersuchungsergebnisse des funktionalen Zusammenhangs sind in Tabelle 6.4 zusammenfassend dargestellt. Für detaillierte Ergebnisse sei auf den Anhang Tabellen 12.3 und 12.4, verwiesen. Abbildungen 31 und 32 zeigen die empirischen Werten und die theoretischen Kurven.

Tabelle 6.4*:* Mittlere Länge in Abhängigkeit von der Frequenz im Lexikon
Anpassung der Funktion $y = ax^b$

Längenmessungen	**a**	**b**	**R^2**
Buchstabenzahl	10.0080	-0.0424	0.0887
Silbenzahl	3.9771	-0.0831	0.2994

[23] Zur Erinnerung: Die Lemmata wurden aus dem Wörterbuch „Le Petit Robert illustré 2007“ gewählt und deren Frequenzwerte aus dem Frequenzwörterbuch „ Frequency dictionary of French words“ von JUILLAND, A. et al. abgelesen.

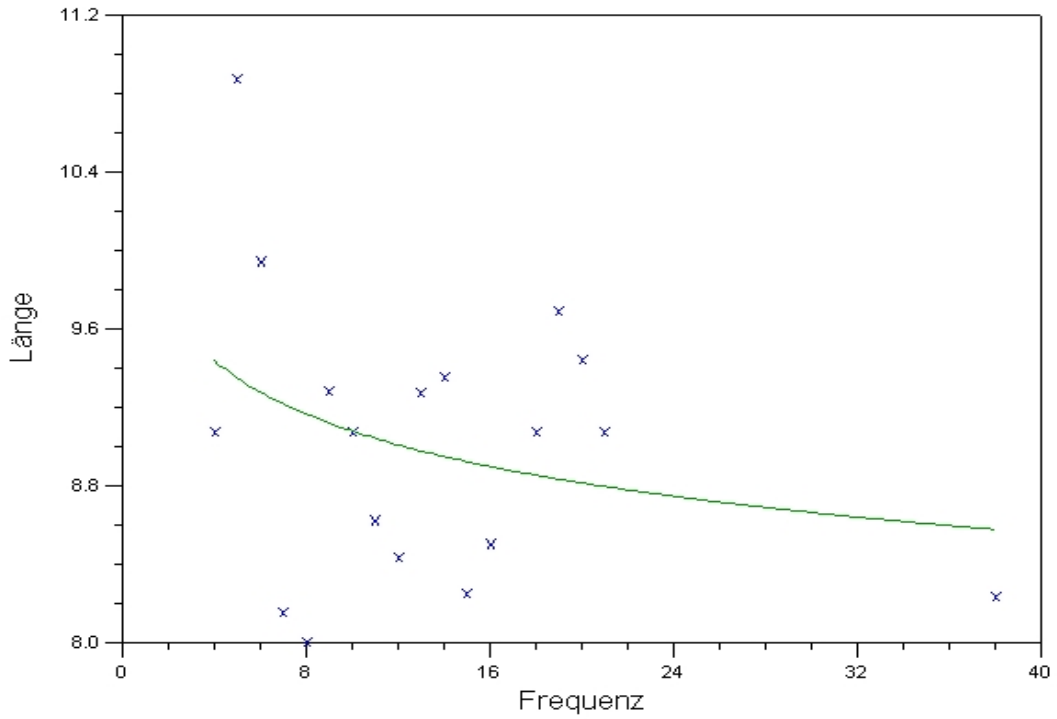

Abbildung 31: Abhängigkeit der Buchstabenlänge von der Frequenz im Lexikon

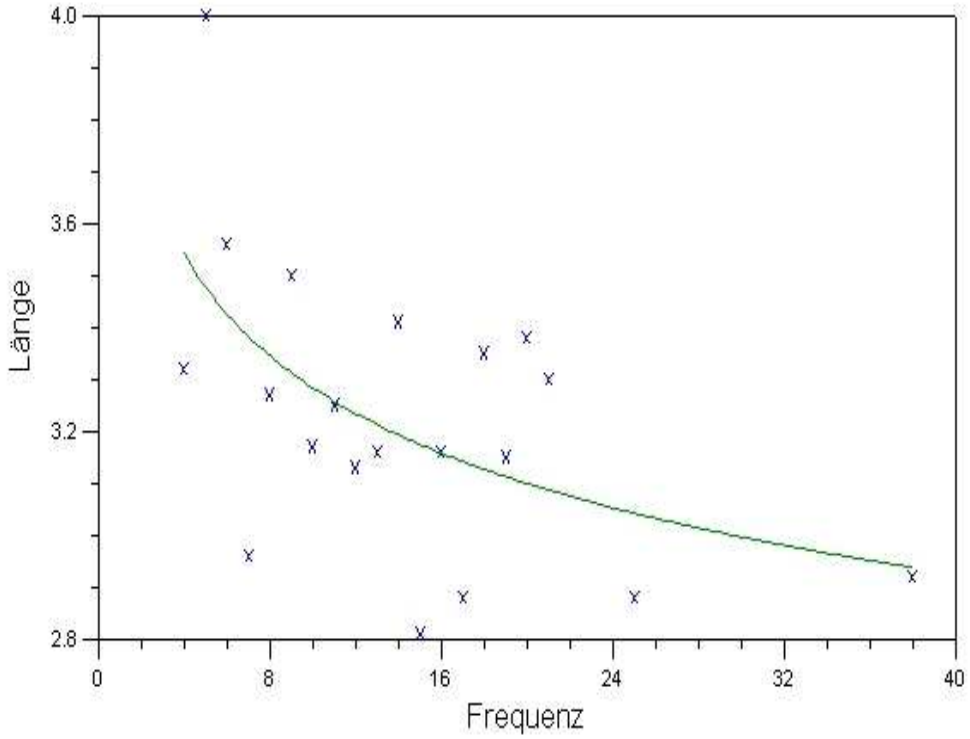

Abbildung 32: Abhängigkeit der Silbenlänge von der Frequenz im Lexikon

Die niedrigen R^2-Werte weisen auf schlechte Anpassungen hin, was auch an den Abbildungen 31 und 32 deutlich erkennbar ist. Hier ist keine glatte Kurve zu verzeichnen; die Datenpunkte bilden eine Wölbung, wie es auch KÖHLER (1986) bereits beobachtete. Die unverkennbare Streuung der empirischen Datenpunkte führen wir in Anlehnung an Köhler auf die Oszillation der Länge in der Frequenz zurück. Diese wird im Abschnitt 4.1.5. behandelt.

6.1.3. Zusammenhang zwischen Polylexie und Frequenz

Die Untersuchung dieses Zusammenhangs wurde wie bei anderen Zusammenhängen auch im Text, Korpus und Lexikon durchgeführt.

a) Polylexie-Frequenz: Untersuchung im Korpus und Text

Die Anpassung der Funktion $y = ax^b$ führte bei der Untersuchung im Korpus zu folgenden Werten für die Parameter **a** und **b**:

a = 3.6402

b = 0.2041

bei einem Determinationskoeffizienten von $R^2 = 0.6154$

Die empirischen und die theoretisch errechneten Werten sind Tabelle 13.1 im Anhang zu entnehmen.

Der R^2-Wert weist auf eine verhältnismäßig schwache Anpassung hin. Er zeigt, dass die durchschnittliche Polylexie der jeweiligen Frequenzklasse sich nicht so verhält, wie es der Zusammenhang vorhersagt. Bei Abbildung 33, welche die empirisch gewonnenen Datenpunkte zusammen mit der theoretischen Kurve $y = 3.6402x^{-0.2041}$ zeigt, ist eine große Streuung zu verzeichnen. Stark ausgeprägt ist diese Streuung besonders im Bereich der großen Frequenzwerte. Die Anpassung der Kurve im Bereich der niedrigen Frequenzen, die die Mehrzahl der Lexeme repräsentieren, ist aber recht gut gelungen, sodass die Abhängigkeit nicht als unbestätigt angesehen werden kann. Abbildung 34 zeigt die Regressionskurve für die Polylexiewerte von F = 1 bis 6.

Für die Untersuchung der Abhängigkeit im Text wurden folgende Werte für die Parameter **a** und **b** ermittelt:

a = 2.9281

b = 0.3064

Die Anpassung ergab einen Determinationskoeffizienten von $R^2 = 0.7515$

Im Vergleich zur Korpus-Untersuchung hat die theoretische Funktion hier besser gepasst. Der Wert von R^2 kann als Indiz einer relativ gut gelungenen Anpassung angesehen werden. Auf

dessen Basis kann die Abhängigkeit der Polylexie von der Frequenz im Text als bestätigt angesehen werden.

Der optische Eindruck der Abbildung 35, welche die empirisch gewonnenen Daten und die ihnen entsprechenden, aus den Parameterwerten errechneten theoretischen Werten zeigt, stützt diese Annahme. Besonders im Bereich der kleinen Frequenzwerte (F < 4), die den größten Teil der Daten ausmachen, und der großen Frequenzwerte (F > 9) liegt die theoretische Funktion $y = 2.9282x^{0.3064}$ recht nah an den Datenpunkten.

Tabelle 13.2 im Anhang enthält die empirisch ermittelten sowie die theoretisch berechneten Daten.

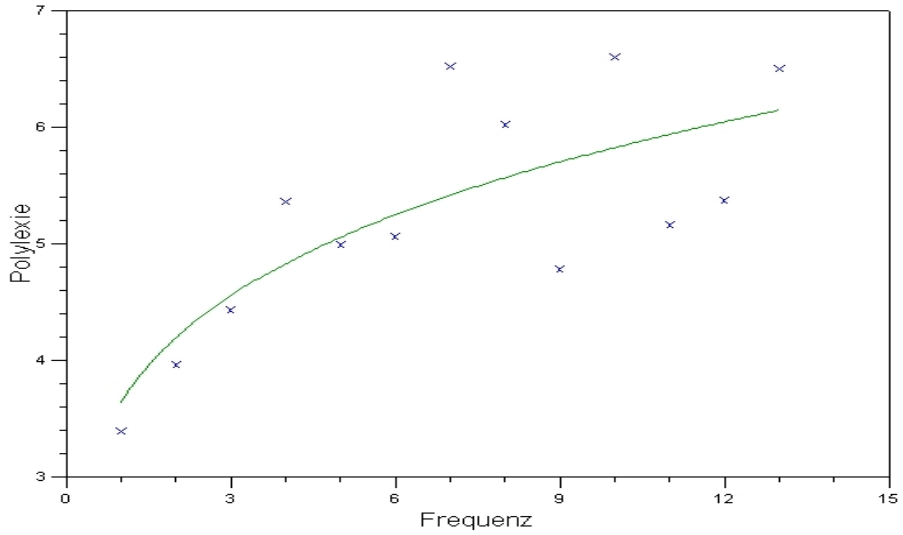

Abbildung 33: Abhängigkeit der Polylexie von der Frequenz im Korpus

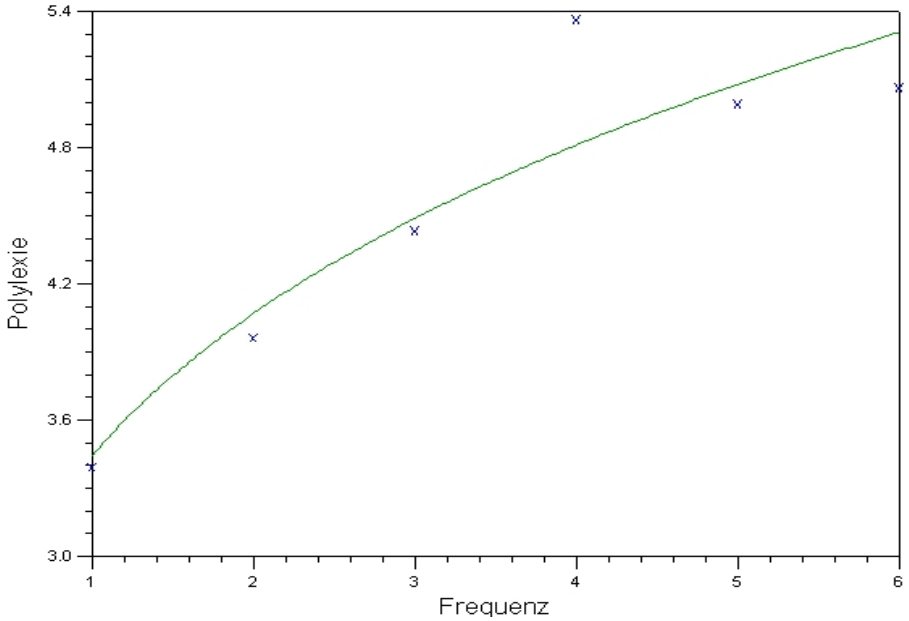

Abbildung 34: Abhängigkeit der Polylexie von der Frequenz im Korpus (Ausschnitt)

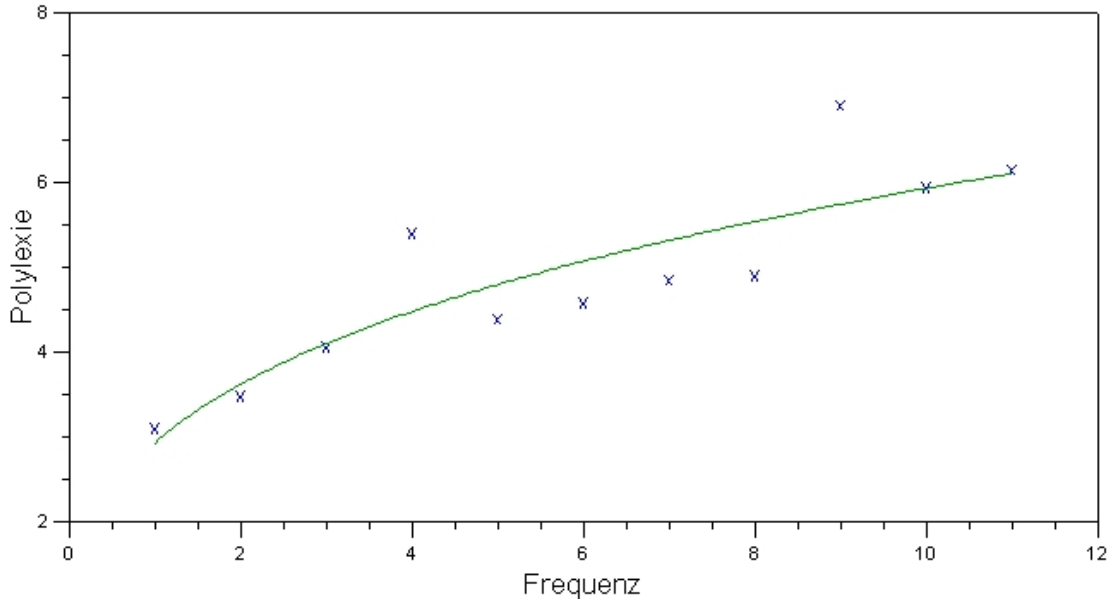

Abbildung 35: Abhängigkeit der Polylexie von der Frequenz im Text (alle Wortarten)

Eine nach Wortarten vorgenommene Untersuchung der Stichprobe führte zu den Ergebnissen in Tabelle 6.5. Für die Wortart *Verb* ist die Anpassung relativ gut gelungen. Demgegenüber sind für die Wortarten *Substantiv* und *Adjektiv* die Anpassungen schlecht. Diese verschlechterten das Gesamtergebnis.

Tabelle 6.5: Abhängigkeit der Polylexie von der Frequenz im Text (einzelne Wortarten)

Wortarten	**a**	**b**	**R^2**
Substantive	3.4277	0.1831	0.4857
Verben	2.4791	0.4205	0.7784
Adjektive	8.8283	-0.4644	0.3436

b) Polylexie-Frequenz: Untersuchung im Lexikon

Die Frequenz eines Wortes wurde hier gewonnen, indem die Anzahl seines Vorkommens im Frequenzwörterbuch von A. JUILLAND (1970) abgelesen wurde. Für die Parameter der angepassten Funktionsgleichung wurden folgende Werte erzielt:

$$a = 1.9094$$

$$b = 0.1925$$

Der Determinationskoeffizient liegt bei $R^2 = 0.3291$

Der Wert von R^2 ist hier kein Beweis für die Abhängigkeit der Polylexie von der Frequenz. Die Anpassung der theoretischen Funktion an die empirischen Daten ist misslungen. An

Tabelle 13.3 im Anhang, welche die empirischen und die theoretischen Werte enthält, kann man keinen eindeutigen Trend erkennen.

Der optische Eindruck der Abbildung 36 lässt aber einen bestehenden Zusammenhang zwischen Polylexie und Frequenz vermuten, denn es weichen nur einige Datenpunkte auffallend von der theoretischen Kurve ab. Einen Erklärungsversuch für den niedrigen R^2-Wert findet man im Abschnitt 6.1.5. der vorliegenden Arbeit.

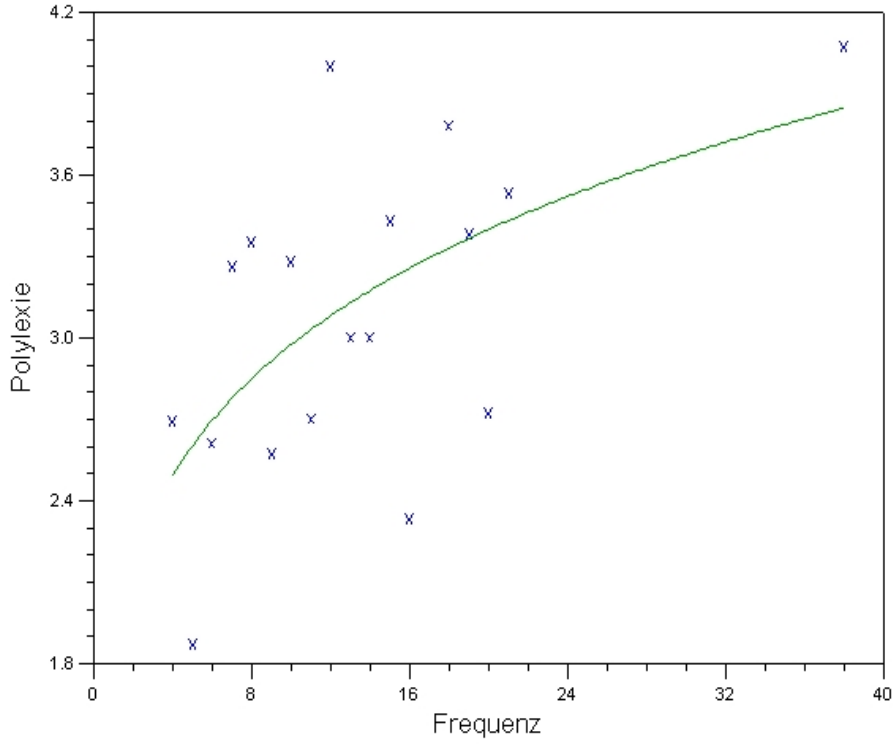

Abbildung 36: Abhängigkeit der Polylexie von der Frequenz im Lexikon

6.1.4. Überprüfung der Zusammenhänge unter Verwendung der Rohdaten

In den vorhergehenden Ausführungen wurde bei der Überprüfung der Zusammenhänge mit Mittelwerten der abhängigen Variablen gearbeitet. Die Methode der Mittelwertbildung hat den Nachteil, dass bei der Berechnung des Determinationskoeffizienten jeder Mittelwert gleich zählt, unabhängig davon, wie viele Einzelwerte er repräsentiert. Dies kann zur Verfälschung des R^2-Wertes und/oder Verschiebung der Regressionskurve führen. Zudem gehen bei der Verwendung dieser Methode Rohdaten-Informationen verloren, wie beispielsweise Varianz und Anzahl der Einzelwerte, die den Mittelwert repräsentieren.

Auf Grund der genannten Probleme wurden die Zusammenhänge zusätzlich auch noch ohne Verwendung der Mittelwerte der abhängigen Variablen überprüft. Jeder Ausprägung der

unabhängigen Variablen wurde der entsprechende Wert der abhängigen Variablen zugewiesen.

Optisch liefern die Rohdaten in allen Fällen ein anderes Bild als bei der Arbeit mit Mittelwerten. Statt einer Datenstreuung um die Regressionskurve herum ist hier eine Ansammlung der Datenpunkte unter sich zu sehen.

Die Anpassungsergebnisse sind in Tabelle 6.6. stellvertretend durch die Ergebnisse der Überprüfung der Abhängigkeit der Polylexie von der Länge im Text und Lexikon dargestellt.

Tabelle 6.6: Abhängigkeit der Polylexie von der Länge im Text und Lexikon

Länge	Untersuchung im Text (Alle Wortarten)			Untersuchung im Lexikon		
	a	b	R^2	a	b	R^2
Buchstabenlänge	16.0072	-0.7787	0.1137	6.1683	-0.3337	0.0459
Silbenlänge				4.1918	-0.3147	0.0621

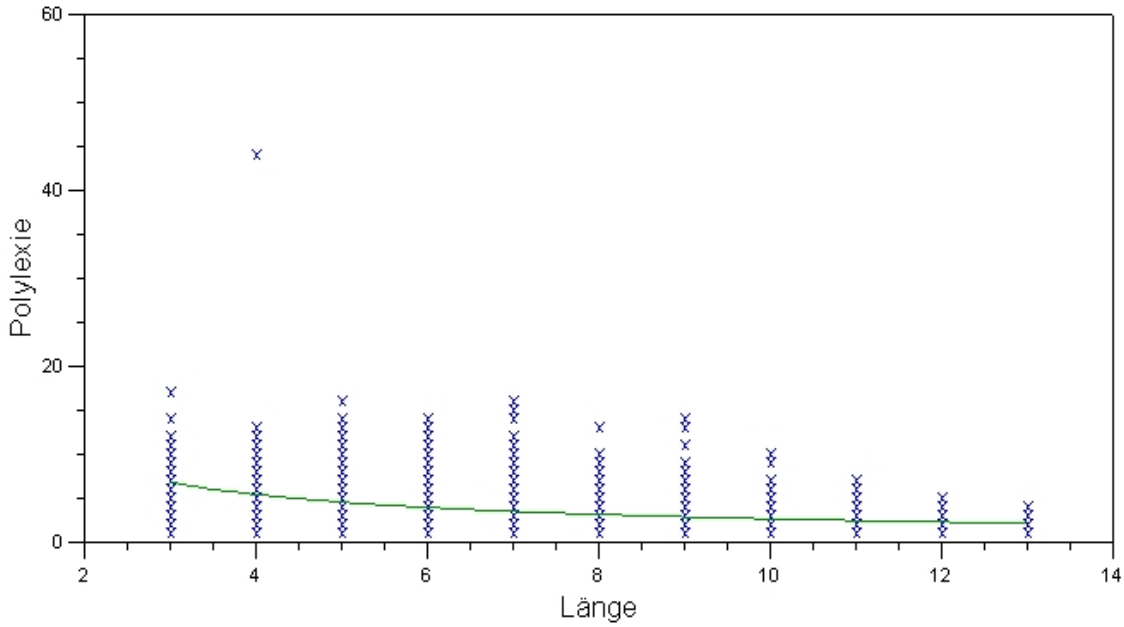

Abbildung 37: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (Verwendung der Rohdaten)

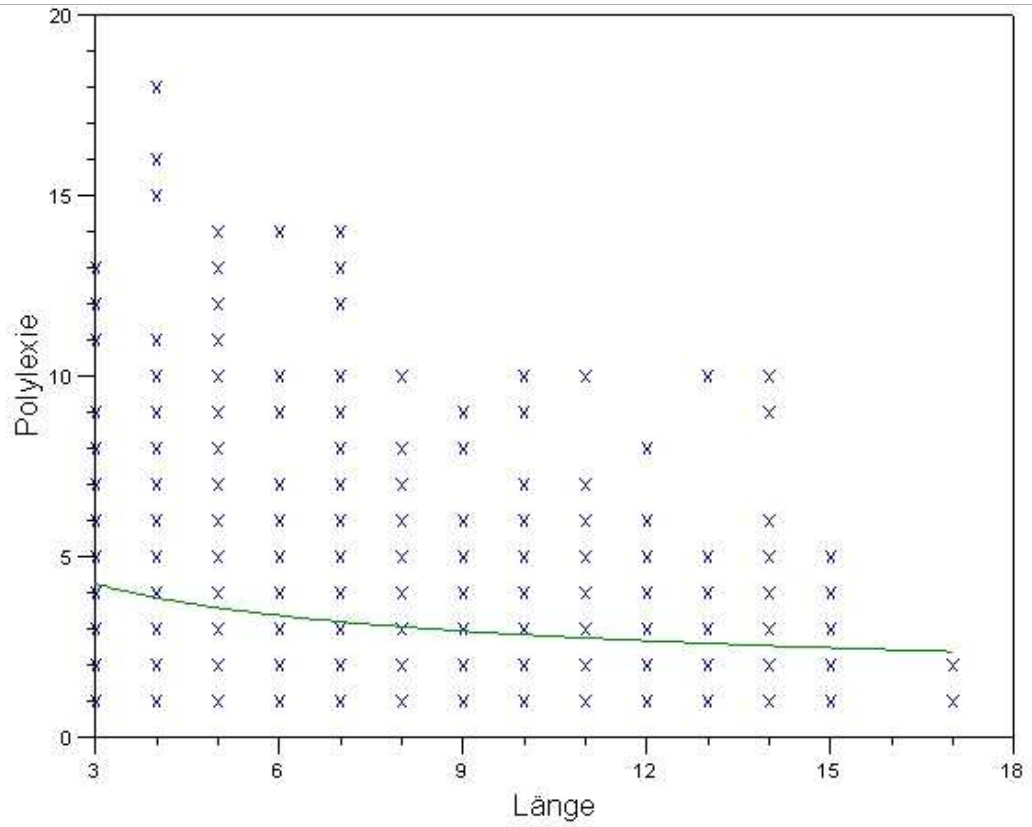

Abbildung 38: Abhängigkeit der Polylexie von der Buchstabenlänge im Lexikon (Verwendung der Rohdaten)

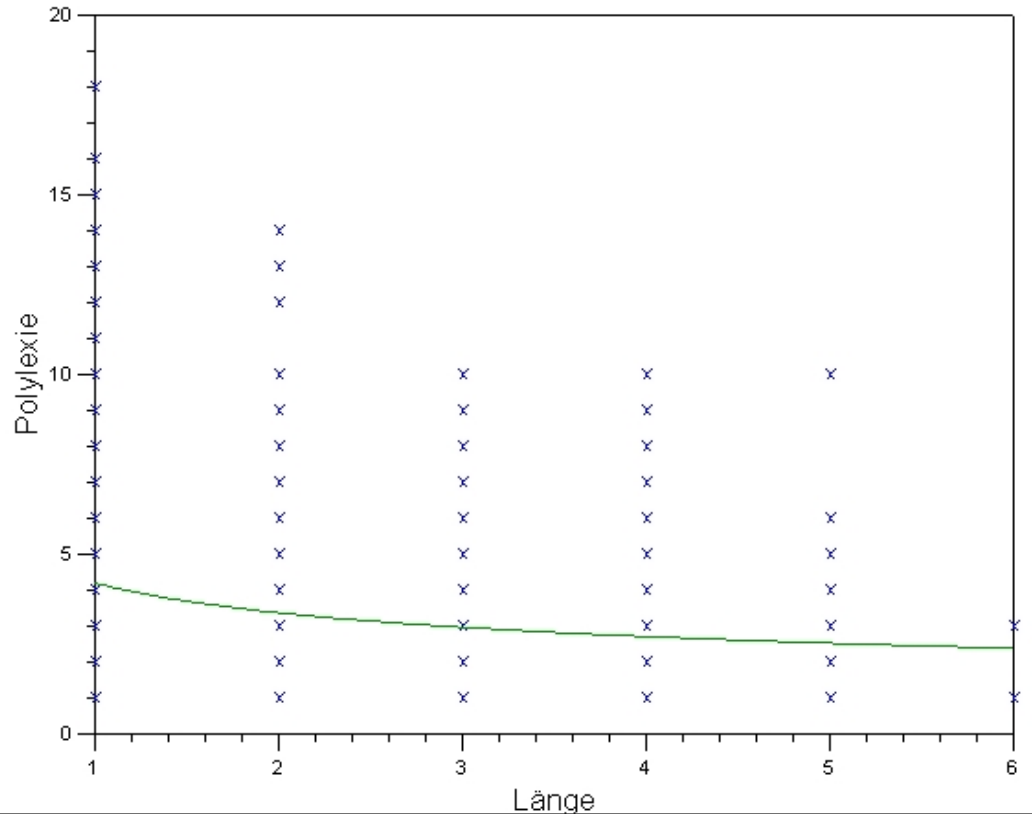

Abbildung 39: Abhängigkeit der Polylexie von der Silbenlänge im Lexikon (Verwendung der Rohdaten)

6.1.5. Oszillation der Lexik

Bei der Überprüfung der theoretischen Funktion $y = ax^b$ zur Beschreibung der Abhängigkeiten zwischen Polylexie und Länge, Länge und Frequenz und Polylexie und Frequenz an französischem Sprachmaterial wurde, ähnlich wie KÖHLER (1986), besonders für die Länge in Abhängigkeit von der Frequenz eine relativ starke Streuung der empirischen Datenpunkte festgestellt.

Um zu testen, ob die von ihm festgestellte Datenstreuung um die theoretische Kurve eine zufällige Erscheinung ist oder nicht, führt KÖHLER die Methode der gleitenden Mittelwerte ein. Dabei werden der auf der Grundlage der bereits vorhandenen $(x_i, \bar{y}_i)$- Paare gleitende Intervalle der Breite *I* gebildet und ein mittlerer Wert für die Frequenz sowie die Länge berechnet.

Die einzelnen Werte ergeben sich nach folgender Formel:

$$\bar{x}_j = \frac{\sum_{j=i}^{i+I-1} x_i}{I} \quad , \quad \bar{y}_j = \frac{\sum_{j=i}^{i+I-1} \bar{y}_i}{I} \quad ; \quad \text{I = Intervallgröße}$$

An die so erzeugten Daten bzw. $(\bar{X}_j\ ,\ \bar{Y}_j)$ -Paare führt KÖHLER eine erneute Anpassung der theoretischen Kurve durch.

Das gleiche Verfahren wurde hier angewendet. Tabelle 6.7 zeigt die Ergebnisse der Anpassung der theoretischen Funktion an die nach Anwendung der Methode der gleitenden Mittelwerte erzeugten Daten zur Abhängigkeit der Länge von der Frequenz im Korpus. Wie an den Werten für R^2 zu erkennen ist, sind die Anpassungen sehr gut gelungen. Das wird auch durch die Abbildungen 40 und 41 deutlich, welche die Daten und die theoretische Funktion mit den aus den Daten geschätzten Parametern zeigen. An diesen Grafiken ist zu erkennen, dass die Methode der gleitenden Mittelwerte die Datenpunkte durchaus glättet.

Nicht nur ist eine Annäherung der Datenpunkte an die Kurve deutlich zu erkennen; unverkennbar ist auch, dass die Datenpunkte um die theoretische Kurve zu oszillieren scheinen. Besonders stark ausgeprägt ist dies bei der Abbildung 40. Bereits KÖHLER (1986) hatte dieses Phänomen beobachtet, das auch hier bestätigt werden kann.

Mit unterschiedlicher Auflösung stellte sich das beobachtete Phänomen auch bei der Untersuchung des Zusammenhangs zwischen Länge und Frequenz im Lexikon. Tabelle 6.8 zeigt exemplarisch die Anpassungsergebnisse nach Anwendung der Methode der gleitenden Mittelwerte über die Intervalle 14 und 15. Abbildungen 42 bis 45 zeigen die geglätteten empirischen Daten.

Tabelle 6.7: Abhängigkeit der Länge von der Frequenz im Korpus bei gleitenden Mittelwerten (Intervalle 5 und 10)

Intervallgröße	a	b	R^2
5	9.2495	-0.1626	0.9588
10	10.8151	-0.2407	0.9963

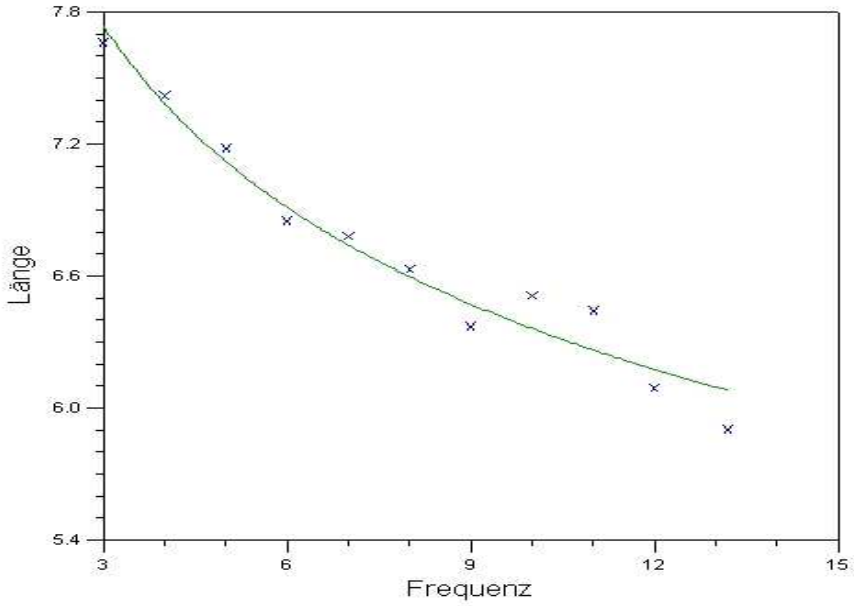

Abbildung 40: Abhängigkeit der Länge von der Frequenz im Korpus, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 5

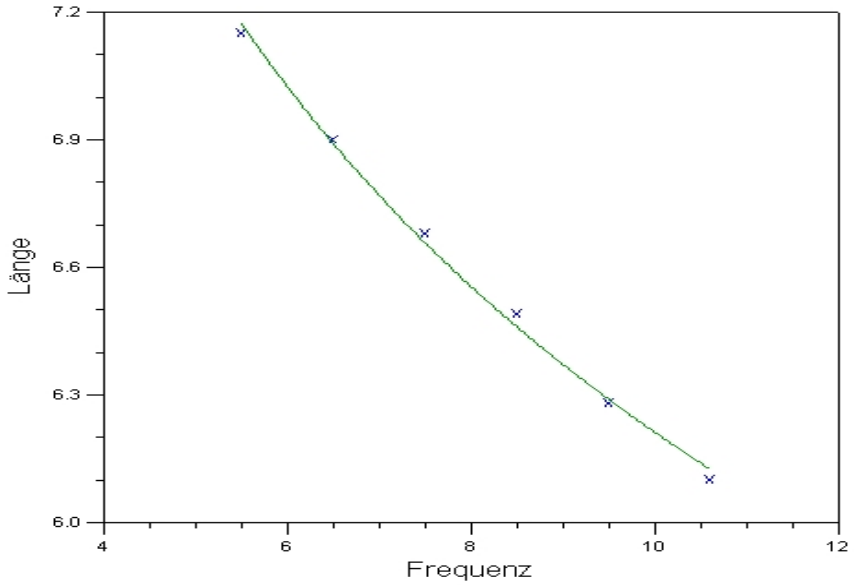

Abbildung 41: Abhängigkeit der Länge von der Frequenz im Korpus, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 10

Tabelle 6.8: Abhängigkeit der Länge von der Frequenz im Lexikon bei gleitenden Mittelwerten der Intervalle 14 und 15

Intervallgröße	Buchstabenlänge			Silbenlänge		
	a	b	R^2	a	b	R^2
14	9.9756	-0.0430	0.7295	152.4682	-14683	0.5564
15	11.0156	-0.0806	0.8542	4.1202	-0.0933	0.9168

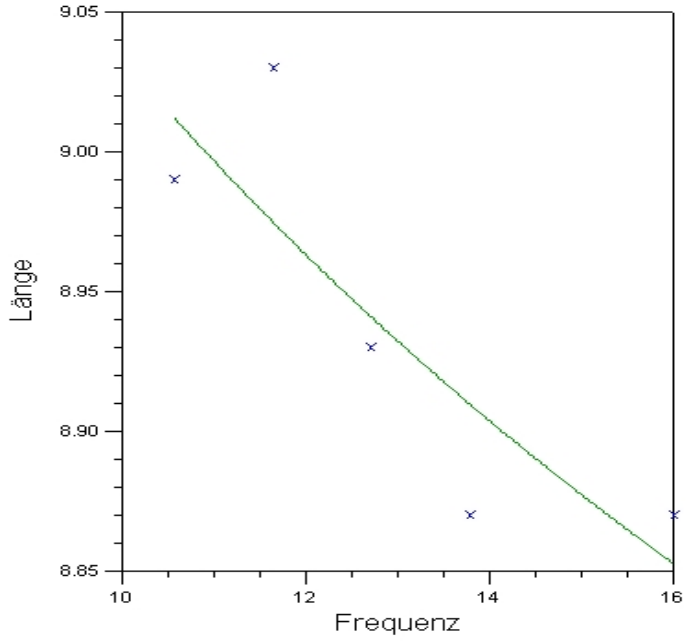

Abbildung 42: Abhängigkeit der Buchstabenlänge von der Frequenz im Lexikon, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 14

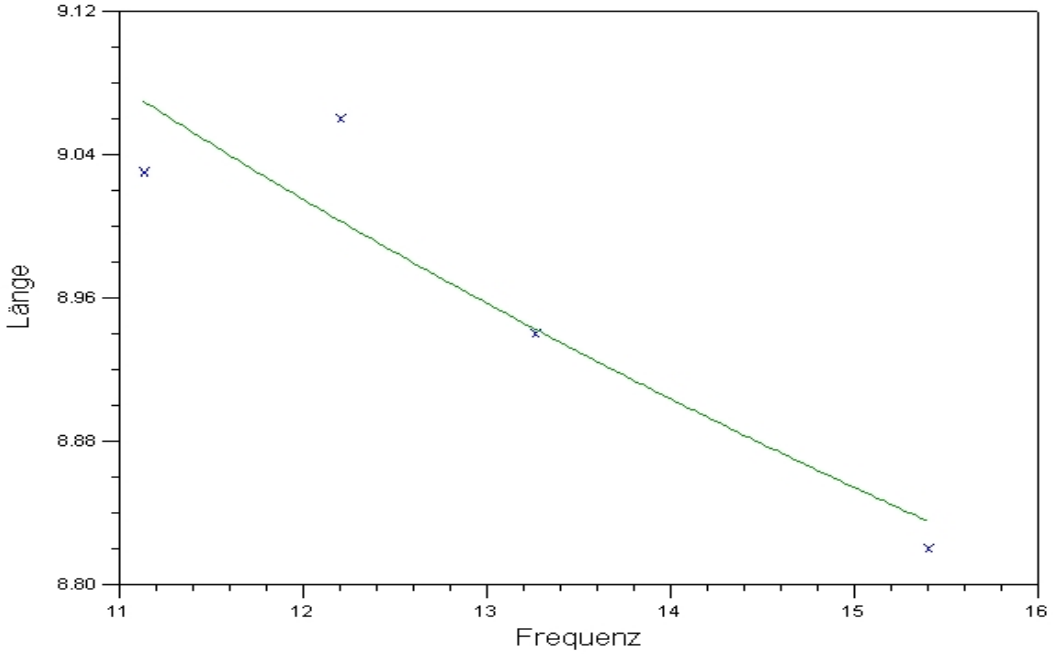

Abbildung 43: Abhängigkeit der Buchstabenlänge von der Frequenz im Lexikon, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 15

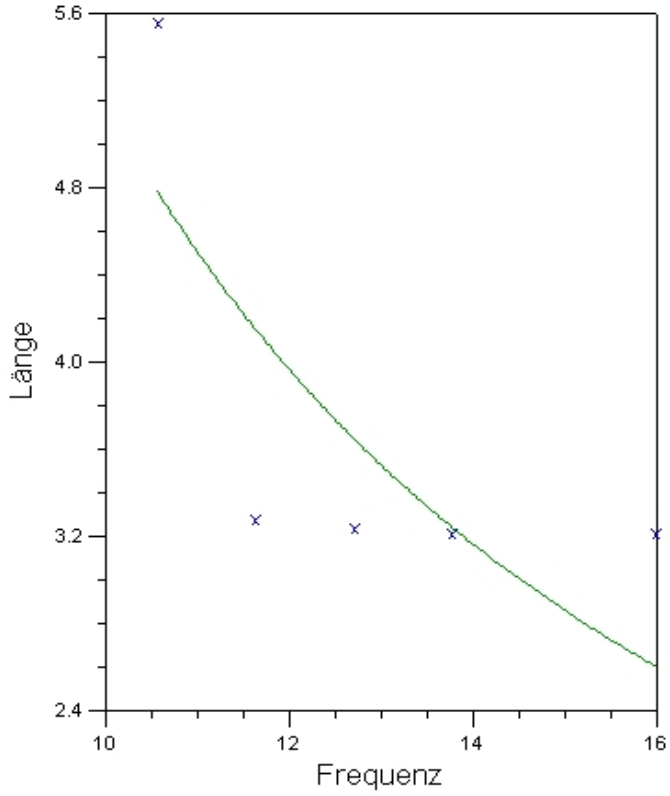

Abbildung 44: Abhängigkeit der Silbenlänge von der Frequenz im Lexikon, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 14

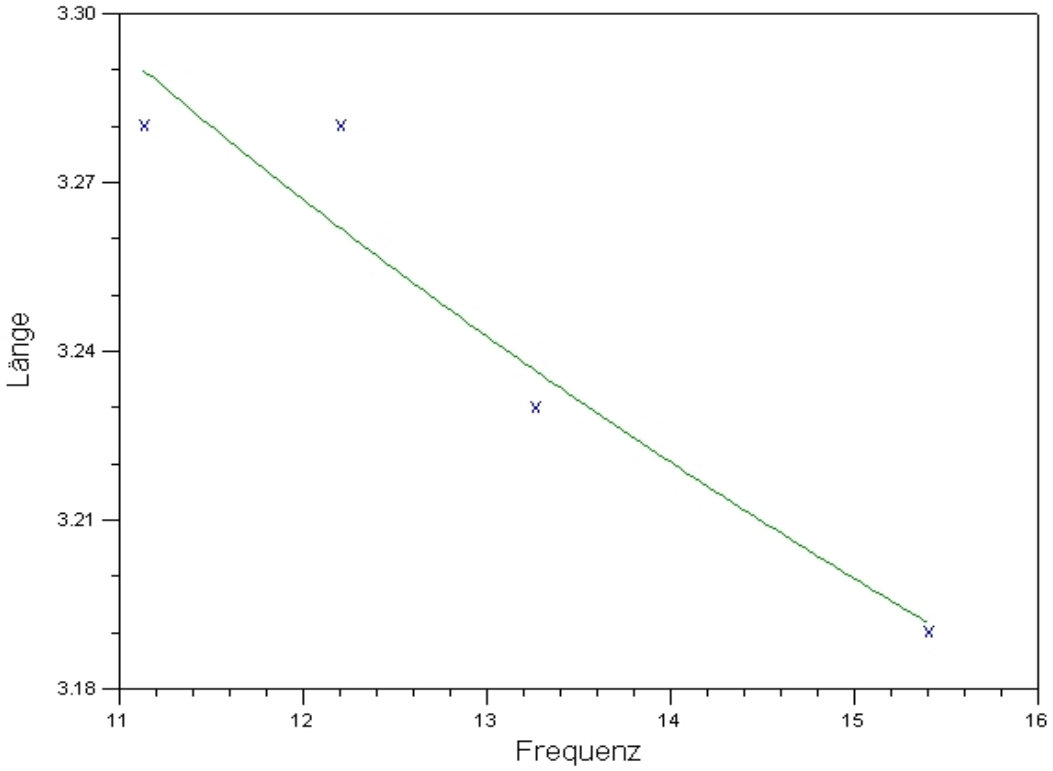

Abbildung 45: Abhängigkeit der Silbenlänge von der Frequenz im Lexikon, berechnet mit Hilfe gleitender Mittelwerte über ein Intervall von 15

Für die Untersuchung der Polylexie in Abhängigkeit von der Frequenz im Lexikon unter Verwendung der Methode der gleitenden Mittelwerte gelten entsprechende Ergebnisse, wie sie für die Abhängigkeit der Länge von der Frequenz im Korpus und Lexikon festgestellt wurden. Es soll darauf verzichtet werden, diese hier im Einzelnen darzustellen.

Für KÖHLER (1986) handelt es sich bei den aufgetretenen Schwingungen um Oszillationen, die auch schon in den Ursprungsdaten vorhanden, aber noch nicht so deutlich zu erkennen waren. Da bei der Überprüfung einiger Zusammenhänge, z.B. der Abhängigkeit der Polylexie von der Buchstabenlänge im Korpus, diese Schwingungen vor Bildung der gleitenden Mittelwerte bereits zu sehen waren, scheint es sich hier um keine Erscheinung zu handeln, die durch die Bildung der Mittelwerte hervorgerufen wurde.
Abbildung 46 zeigt die Daten noch einmal.

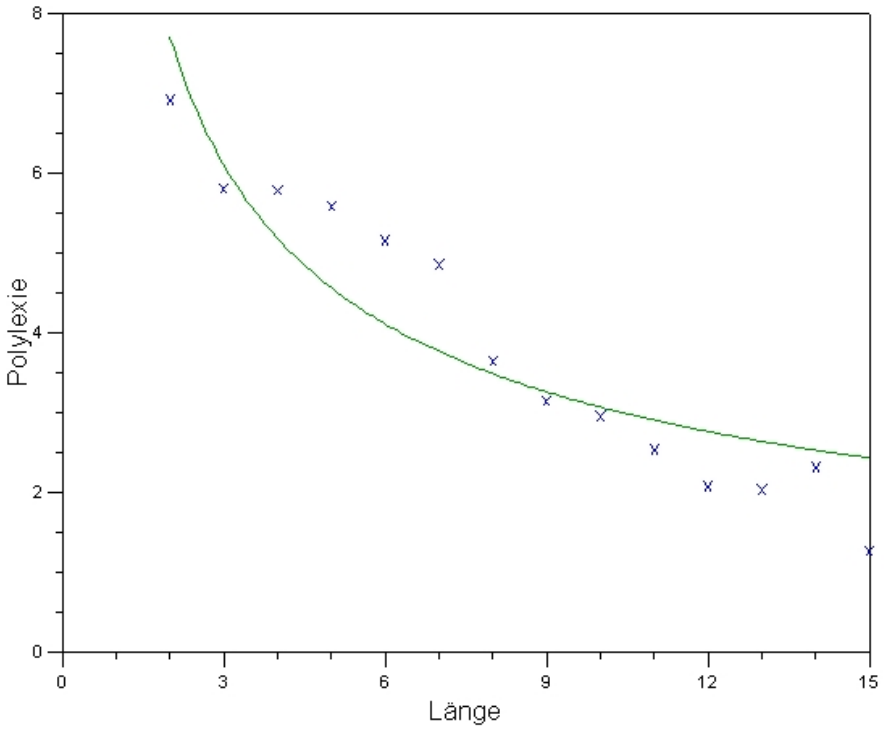

Abbildung 46: Abhängigkeit der Polylexie von der Buchstabenlänge im Korpus, berechnet ohne Methode der gleitenden Mittelwerte

In Bezug auf die bei der Überprüfung der Abhängigkeit der Länge von der Frequenz auftretenden Schwingungen vertritt HAMMERL (1990, 19) die Auffassung, dass es sich hierbei

„um eine Streuung der empirischen Zahlenwerte um die theoretischen Werte handelt, so wie sie auch bei der Untersuchung anderer Relationen zwischen quantitativen Eigenschaften sprachlicher

Einheiten auftritt, ohne diesen Abweichungen den Sonderstatus von ‚Schwingungen' einzuräumen."

Der Annahme der Oszillation setzt HAMMERL die Ergebnisse einer Periodizitäts-Untersuchung entgegen. Er wendet den Phasenverteilungstest und den Phasenhäufigkeitstest von Wallis und Moore an, die in der Lage sind, „Periodizitäten der empirischen Werte einer Größe in Bezug auf eine andere Größe nachzuweisen, wenn der Einfluss eines vorliegenden Trends eliminiert wird" (S. 19). Auf Grundlage dieser Tests konnte aber keine Periodizität nachgewiesen werden.

6.2. Zusammenhang zwischen Satz- und Teilsatzlänge

Gegenstand des vorliegenden Abschnitts ist die Überprüfung des Menzerath-Gesetzes auf der Satzebene, d.h. die Überprüfung der Hypothese, ob zwischen der Satz- und der Teilsatzlänge der folgende stochastische Zusammenhang besteht:

„Je länger ein Satz, desto kürzer die Teilsätze"

Hierzu werden als Datengrundlage die literarischen Texte 24, 35, 37, 43 und 45 herangezogen. Entsprechend den festgelegten Operationalisierungskriterien wurden aus diesen Texten 1466 Sätze ausgezählt und anschließend für jeden Satz und Teilsatz die Länge berechnet.

Bei der Wortzählung wurden folgende Konventionen aus DIECKMANN & JUDT (1996) angenommen:

- Apostrophierte Personalpronomen, Artikel und apostrophiertes *ne* (= n') wurden als Wort angesehen,
- *t* in a-t-il zählte nicht als Wort,
- Jahr- und Prozentangaben sowie Geschwindigkeits- und Zeitangaben galten jeweils als 1 Wort,
- Abkürzungen wie PT (Parti Socialiste) wurden als ein Wort betrachtet. Dasselbe galt für römische Zahlen wie XX,
- *Avez-vous* und Ähnliches wurde als zwei Wörter angesehen, aber lui-même, peut-être und ähnliche Komposita als ein Wort.

Auf die 1466 Sätze verteilen sich insgesamt 3680 Teilsätze und 33977 Wörter. Die durchschnittliche Satzlänge liegt demzufolge bei 2.5102 Teilsätzen und der Mittelwert für die Teilsatzlänge bei 9.2328 Wörtern.

Am häufigsten belegt sind Sätze mit der Länge 2, die 31.44 % der Gesamtzahl repräsentieren. Darauf folgen in geringem Abstand Sätze der Länge 1. Sie nehmen 29.33 % aller Sätze in Anspruch. Bereits weitaus weniger vertreten sind Sätze der Länge 3 und 4; sie umfassen je-

weils 19.37 % und 9.68 % aller Sätze. Minimal ist der Anteil der Sätze ab der Länge 7; Sie kommen zusammen 38-mal (2.59 %) vor. Unter der 1 %-Grenze liegt die Häufigkeit der Sätze ab der Länge 8. Die verbleibenden Satztypen – Sätze der Länge 5 und 6 – erscheinen zusammen 110-mal und umfassen 7.50 % aller Sätze.

Sätze der Länge 1 bis 6 decken fast das gesamte Untersuchungskorpus ab, da ihre Frequenz bei 97.40 % liegt.

Tabelle 6.9 zeigt die Verteilung der Satz- und Teilsatzlänge

Tabelle 6. 9: Verteilung der Satzlängen und Teilsatzlängen

Satzlänge in Teilsätzen	Anzahl der Sätze	Anzahl Teilsätze	Wörter insgesamt
1	431	431	4904
2	461	922	8912
3	284	852	7742
4	142	568	5023
5	65	325	2735
6	45	270	2158
7	21	147	1316
8	7	56	441
9	5	45	373
10	1	10	60
12	1	12	137
14	3	42	176
Summe	1466	3680	33977

Die Überprüfung des Zusammenhangs zwischen Satz- und Teilsatzlänge erfolgte über die Bildung von Mittelwerten für die Teilsätze in den einzelnen Satzlängen. Beobachtete Satzlängen mit weniger als 10 Vorkommen wurden von der Betrachtung ausgeschlossen, sodass $1 \leq x_i \leq 7$ und N = 1466 Sätze. Aus den sieben Einzelverteilungen wurden die Mittelwerte errechnet, wobei diese 7 Messungen als empirische Werte für die mittleren Teilsatzlängen in Sätzen mit x_i-Teilsätzen verwendet wurden. Tabelle 6.10 zeigt die empirisch errechneten durchschnittlichen Teilsatzlängen. An dieser Übersicht ist zu erkennen, dass prinzipiell ein Zusammenhang zwischen Satz- und Teilsatzlänge besteht; es lässt sich ersehen, dass der Mittelwert für die abhängige Variable kleiner wird, je größer das Konstrukt ist. Eine konstante Abnah-

mequote lässt sich dabei aber nicht beobachten, denn bei $x_i = 7$ steigen die Messwerte wieder an. Dies kann durch die geringe Anzahl der Messungen hervorgerufen worden sein.

Tabelle 6.10: Durchschnittliche Teilsatzlänge

Satzlänge in Teilsätze	Durchschnittliche Teilsatzlängen
1	11.3781
2	9.6659
3	9.0868
4	8.8433
5	8.4153
6	7.9925
7	8.9523

An die Daten aus Tabelle 6.10 wurde eine Anpassung der Funktionen durchgeführt, die von ALTMANN (1980) abgeleitet wurden:

(a) $y = ax^b e^{-cx}$

(b) $y = a\, e^{-cx}$

(c) $y = ax^b$

Für die Parameter dieser Funktionen und den Determinationskoeffizienten wurden die in Tabelle 6.11 aufgeführten Werte ermittelt. In Tabelle 14.1 im Anhang sind die den empirischen Ergebnissen entsprechenden, aus den Funktionsparametern theoretisch errechneten Werte aufgeführt. Abbildungen 47 bis 49 zeigen die empirischen Werte gemeinsam mit der theoretischen Kurve der jeweiligen Funktion.

Sowohl mit der Funktion (a) als auch mit der Funktion (c) wird der beobachtete Trend hinreichend erfasst. Auch die entsprechenden Abbildungen zeigen eine gute Übereinstimmung der theoretischen Kurvenverläufe mit den empirischen Werten. Ein besseres Ergebnis ermöglicht allerdings (a). Dies ist nicht überraschend, da diese Funktion nicht weniger als 3 Parameter enthält.

Mit Funktion (b) ließ sich der beobachtete Trend nicht gut erfassen. Dies zeigt sowohl das ermittelte R^2 als auch Abbildung 49. Es lässt sich hier feststellen, dass die Kurve die Datenvariabilität nicht signifikant erklärt und die beobachtete Tendenz nicht gut erfasst. Die Abweichung von einer gedachten Idealkurve scheint jedoch nicht erheblich; das Ergebnis ist befriedigend, sodass das Menzerath-Gesetz auch an dieser Stelle als bestätigt angenommen werden kann.

Tabelle 6.11: Teilsatzlänge als Funktion der Satzlänge, Anpassung der Funktionen an Daten der Gesamttexte

Funktionen	a	b	c	R^2
(a) $y = ax^b e^{-cx}$	10.7855	-0.3196	-0.0560	0.9344
(b) $y = a\, e^{-cx}$	11.0117		0.0463	0.6643
(c) $y = ax^b$	11.0814	-0.1579		0.8591

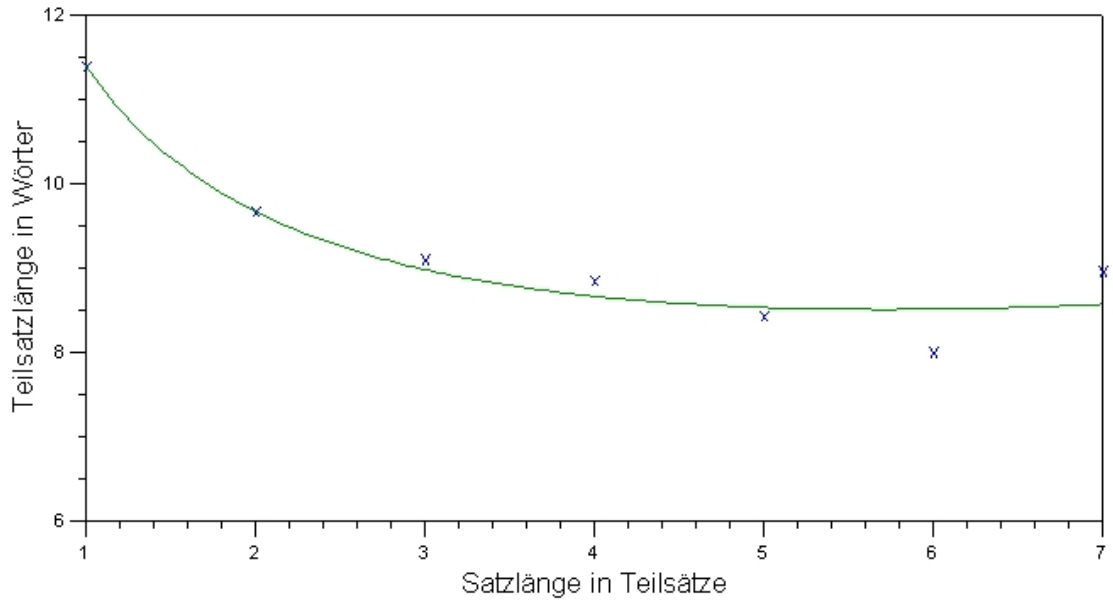

Abbildung 47: Teilsatzlänge als Funktion der Satzlänge. Anpassung der Funktion $y = ax^b e^{-cx}$

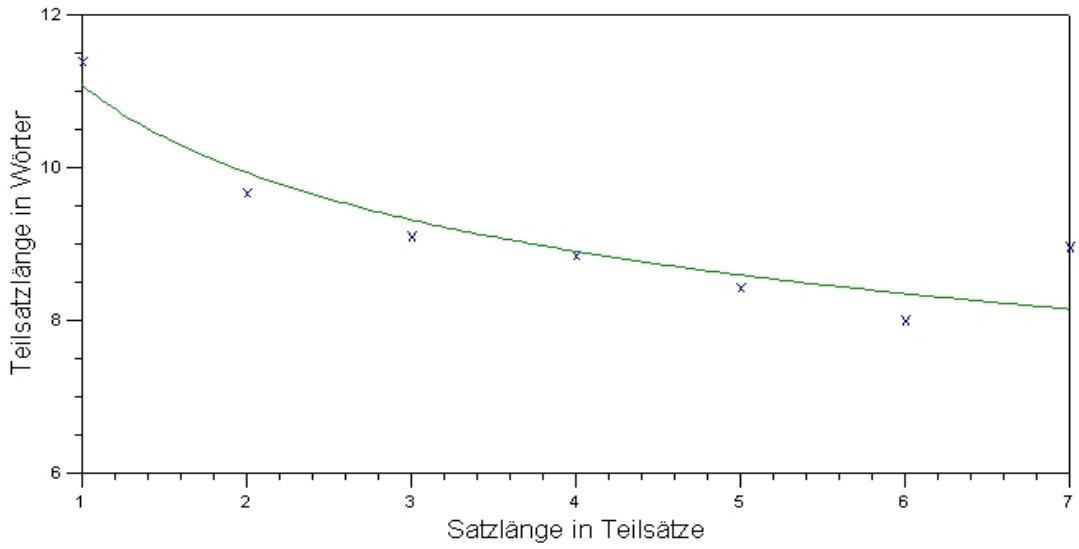

Abbildung 48: Teilsatzlänge als Funktion der Satzlänge. Anpassung der Funktion $y = ax^b$

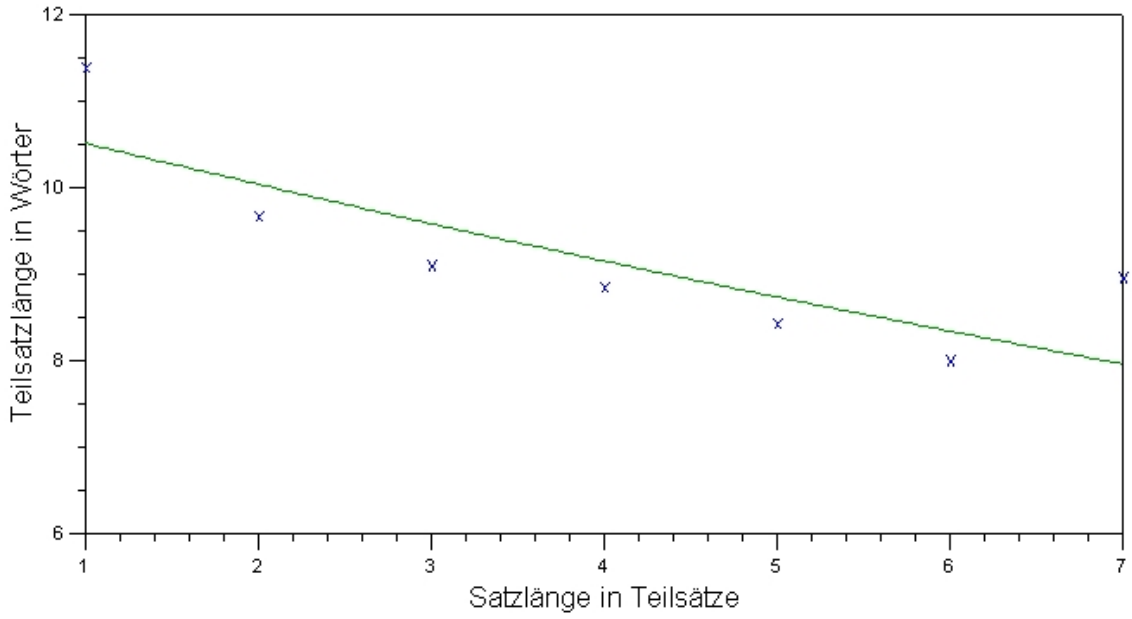

Abbildung 49: Teilsatzlänge als Funktion der Satzlänge. Anpassung der Funktion $y = a\,e^{-cx}$

Für Vergleichszwecke werden in Tabelle 6.12 für jeden untersuchten Text neben den angepassten Funktionen die abgeschätzten Parameter und die Ergebnisse des Determinationskoeffizienten aufgeführt. Die Tests wurden nur für Satzlängen mit mindestens 10 Belegen durchgeführt. Tabellen mit den beobachteten und berechneten durchschnittlichen Teilsatzlängen finden sich im Anhang, Tabelle 14.2.

Tabelle 6.12 kann man Folgendes entnehmen:

Für vier der fünf Einzelstichproben wurden signifikante bis hoch signifikante Ergebnisse erzielt; die R^2-Werte weisen auf (sehr) gute Anpassungen hin. Die besten Ergebnisse wurden im Text 35 erzielt, mit R^2 von fast 1 für die angepassten Funktionen. Abbildungen 50 bis 52 zeigen diese Resultate anschaulich. Die Kurvenverläufe entsprechen hier den Erwartungen.

Die Untersuchung im Text 24 ergab kein zufrieden stellendes Resultat. An die Daten aus dieser Stichprobe konnte keine Anpassung mit einem $R^2 \geq 0.50$ durchgeführt werden. Auch die Abbildungen 53 bis 55, welche die empirischen Werte gemeinsam mit den theoretischen Kurven zeigen, lassen keinen Trend erkennen.

Dass im Text 24 keine guten Ergebnisse erzielt werden konnten, kann damit zusammen hängen, dass die Messwerte (siehe Tabelle 14.2 im Anhang) bei $x_i = 6$ wieder ansteigen. Dieser Anstieg kann sich durch die geringe Anzahl der Messungen erklären: In der betroffenen Stichprobe kommen nämlich nur 13 Sätze vor, die jeweils aus sechs Teilsätzen bestehen.

Die aus diesem Einzeltext resultierenden Ergebnisse widersprechen nicht der Tatsache, dass zwischen Satz- und Teilsatzlänge eine Beziehung besteht. Dies ist auf Grund des stochasti-

schen Charakters der Hypothese durchaus zulässig. Der Zusammenhang kann als nachgewiesen betrachtet werden.

Tabelle 6.12: Zusammenhang zwischen Satz- und Teilsatzlänge in den einzelnen Texten

Texte	$y = ax^b e^{-cx}$	$y = a\,e^{-cx}$	$y = ax^b$
24	a = 8.6259 b = -0.1463 c = -0.03450 R^2 = 0.4903	a = 8.7891 c = 0.0166 R^2 = 0.2970	a = 8.7961 b = -0.0539 R^2 = 0.4079
35	a = 9.1672 b = -0.2146 c = 0.0314 R^2 = 1	a = 9.9867 c = 0.1341 R^2 = 0.9776	a = 8.9145 b = -0.2780 R^2 = 0.9980
37	a = 16.8130 b = -0.6225 c = -0.0942 R^2 = 0.9903	a = 20.5383 c = 0.1726 R^2 = 0.8744	a = 18.2070 b = -0.4125 R^2 = 0.9764
43	a =11.6991 b = -0.6324 c = -0.1760 R^2 = 0.9969	a = 15.0322 c = 0.1250 R^2 = 0.8071	a = 13.6840 b = -0.2760 R^2 = 0.9347
45	a =11.8531 b = -0.4693 c = -0.0739 R^2 = 0.9693	a =12.3975 c = 0.0801 R^2 = 0.7368	a =12.3530 b = -0.2612 R^2 = 0.9193

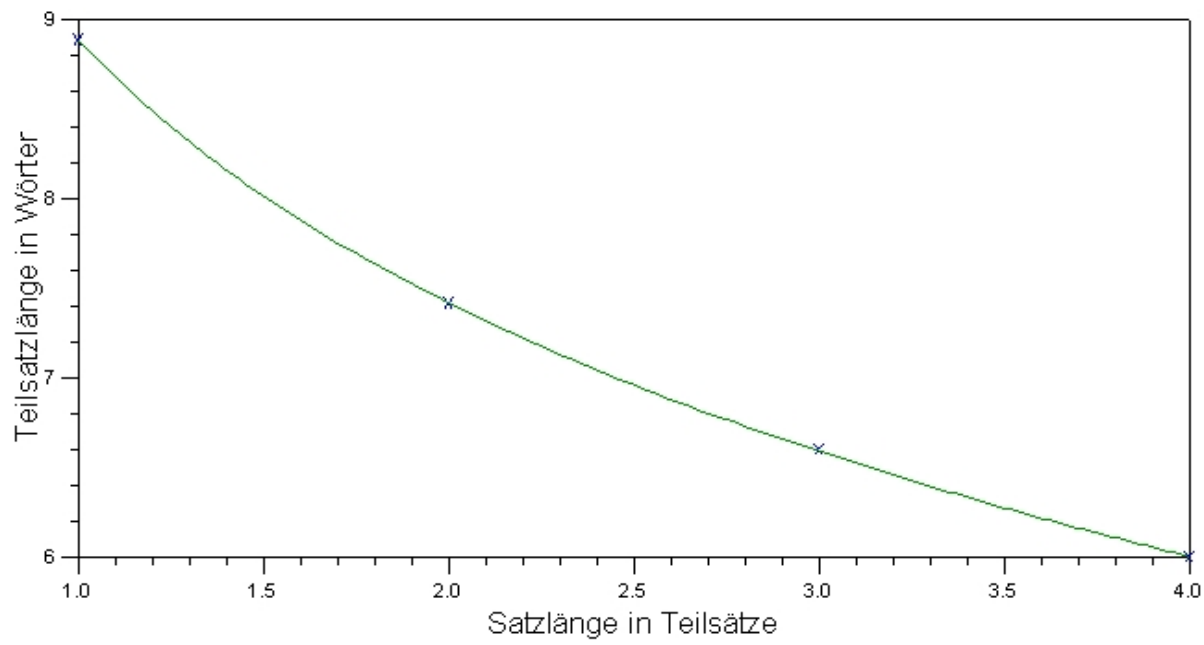

Abbildung 50: Satz- und Satzlänge im Text 35. Anpassung der Funktion $y = ax^b e^{-cx}$

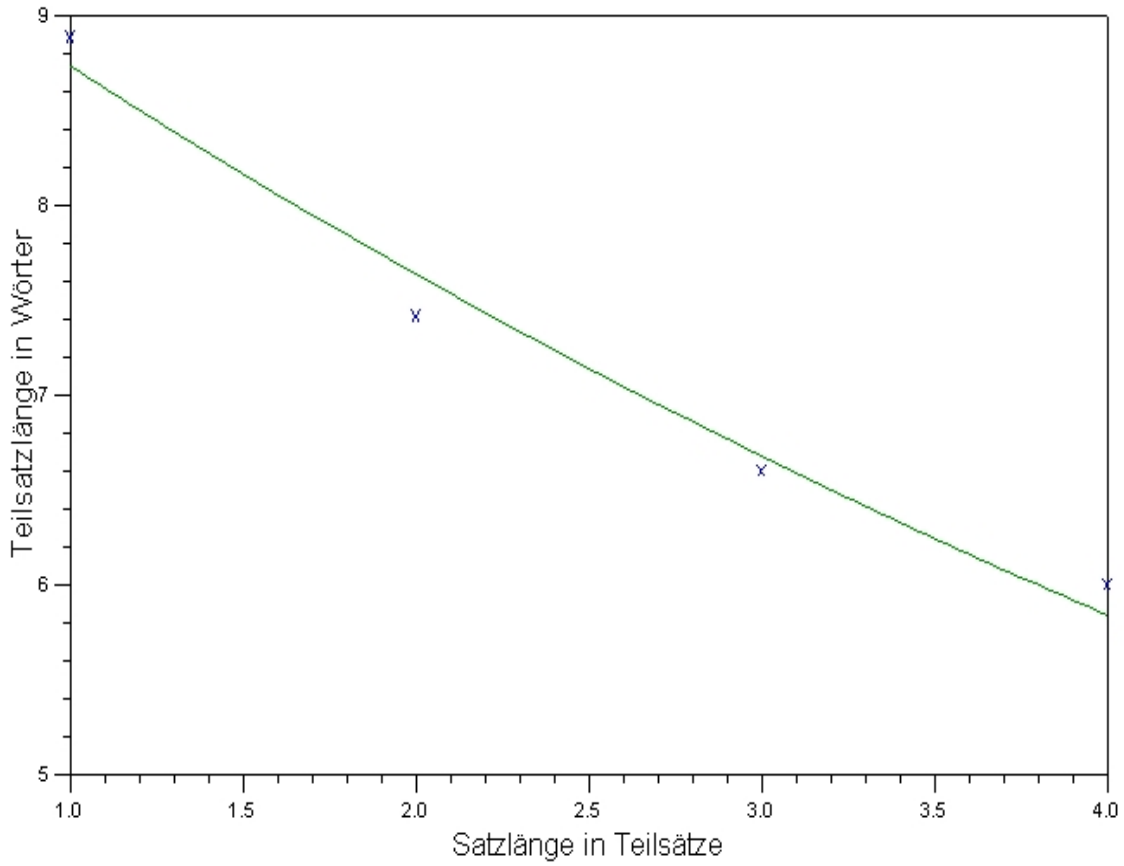

Abbildung 51: Satz- und Satzlänge im Text 35. Anpassung der Funktion $y = a\ e^{-cx}$

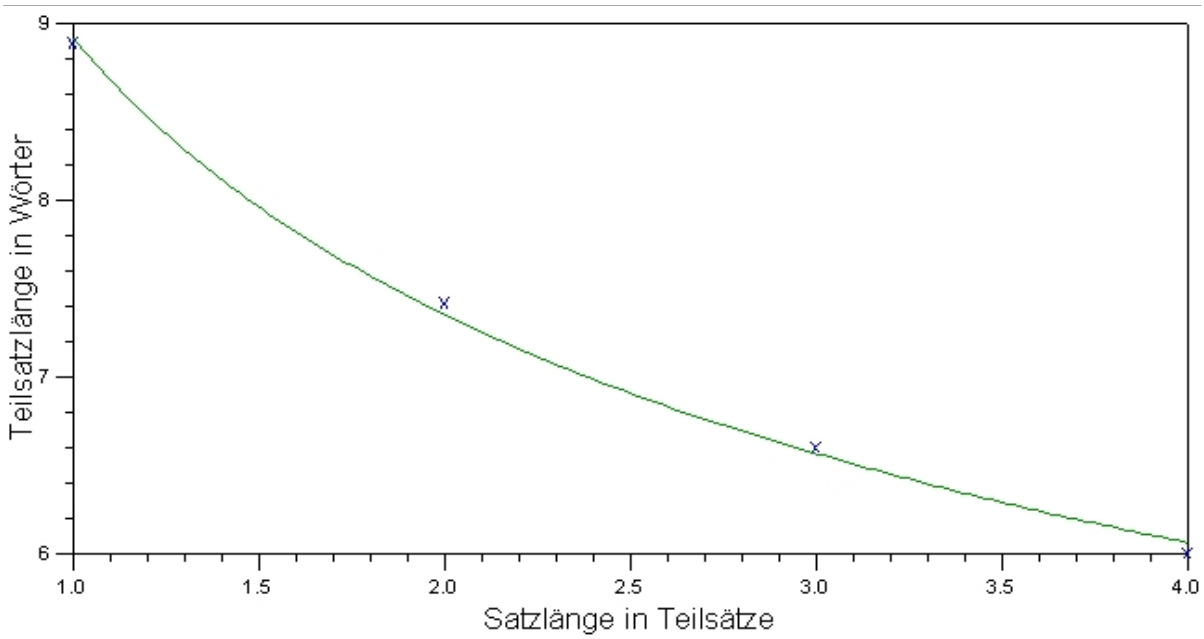

Abbildung 52: Satz- und Satzlänge im Text 35. Anpassung der Funktion $y = ax^b$

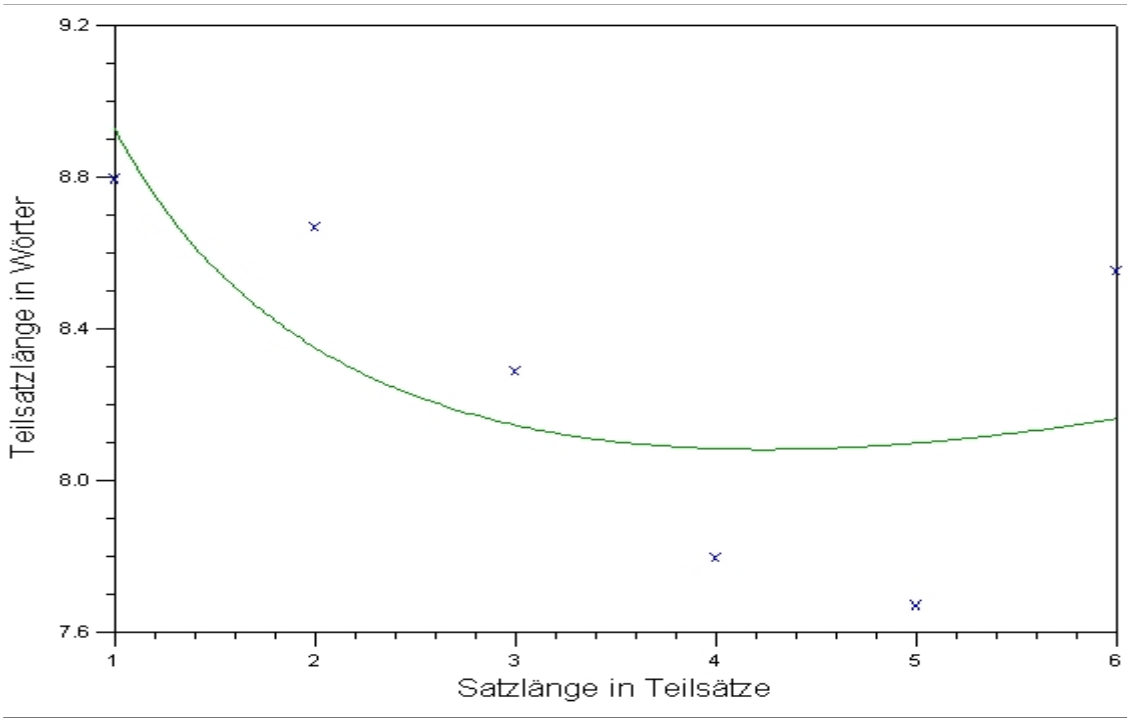

Abbildung 53: Satz- und Teilsatzlänge im Text 24. Anpassung der Funktion $y = ax^b e^{-cx}$

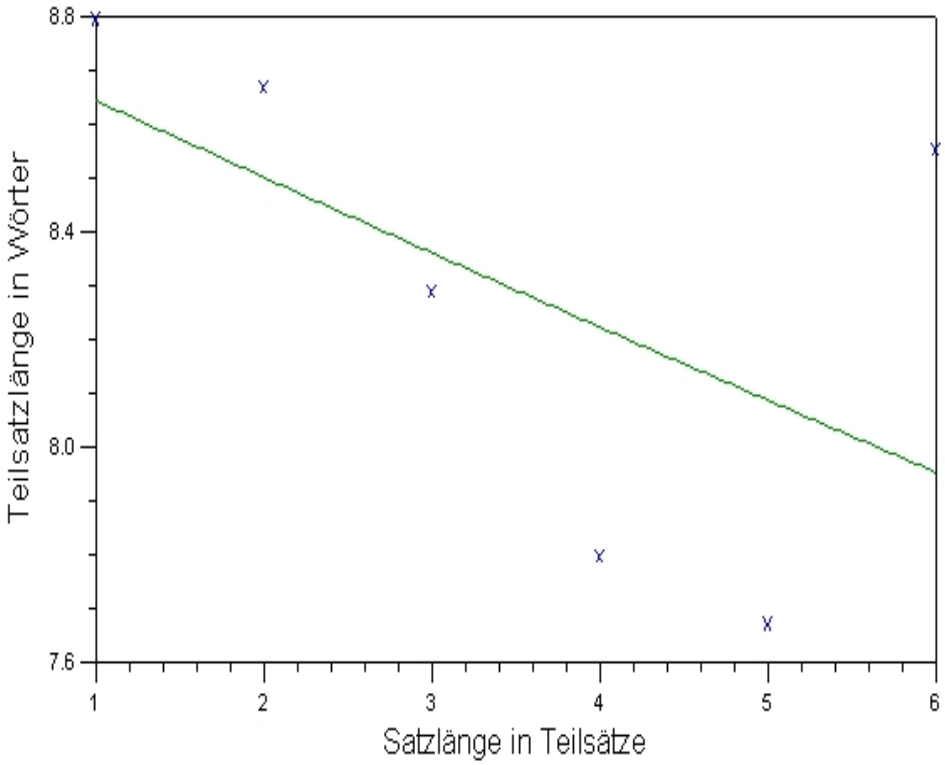

Abbildung 54: Satz- und Teilsatzlänge im Text 24. Anpassung der Funktion $y = a\ e^{-cx}$

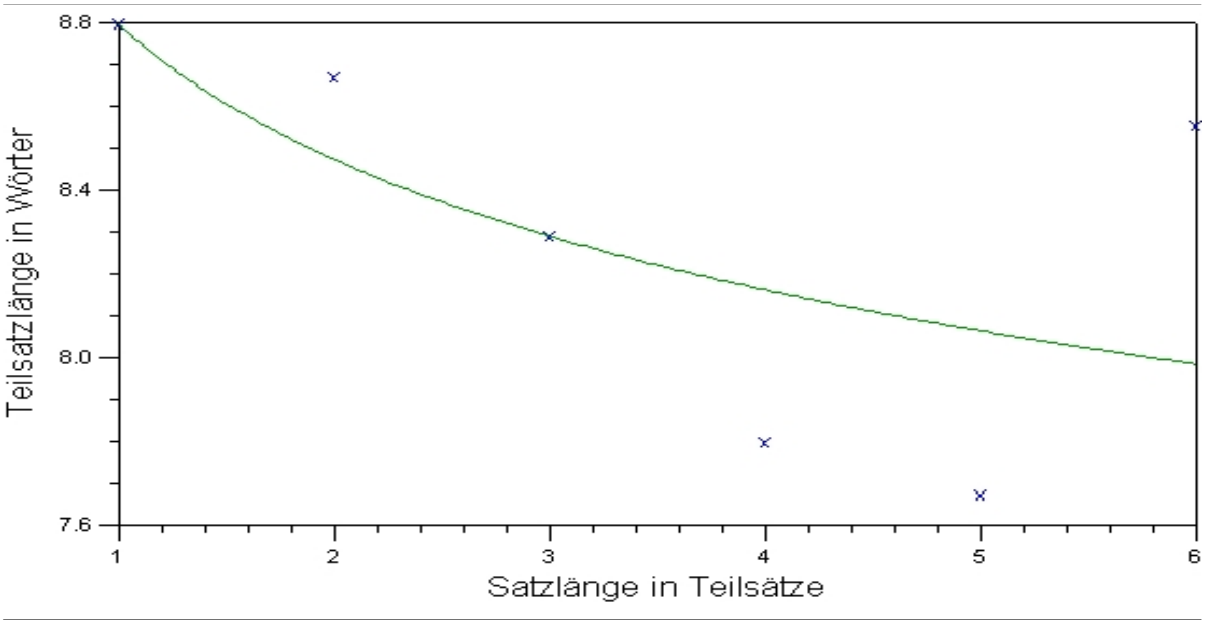

Abbildung 55: Satz- und Satzlänge im Text 24. Anpassung der Funktion $y = ax^b$

6.3. Zusammenhang zwischen Frequenz und Komplexität von Teilsätzen

Die Annahme, dass zwischen der Frequenz und der Komplexität von Teilsätzen eine Beziehung besteht, soll hier einer empirischen Überprüfung unterzogen werden. Als Datenbasis dienen die gleichen Daten, an denen die Verteilung der Frequenz und Komplexität von Teilsätzen untersucht wurde, d.h. die 1505 Teilsätze, die aus den Texten 24, 35, 37, 43 und 45 gewonnen wurden. Für jeden Teilsatz wurde die Komplexität und anschließend für alle Teilsätze mit gleicher Komplexität die durchschnittliche Frequenz berechnet. Es entstand auf diese Weise eine Tabelle mit der Komplexität als unabhängiger und der durchschnittlichen Frequenz als abhängiger Größe.

An die empirischen Daten, die aus mindestens 10 Belegen hervorgegangen waren, wurde in Anlehnung an KÖHLER (1999) das mathematische Modell

$$y = ax^b e^{cx}$$

angepasst. Für die Parameter **a, b** und **c** dieser Funktionsgleichung wurden folgende Werte ermittelt:

$a = 2.6363$

$b = 0.7340$

$c = -0.4269$

bei einem Determinationskoeffizienten von

$R^2 = 0.9711$

Tabelle 15 im Anhang enthält eine Gegenüberstellung der empirischen Daten und der aus den Funktionsparametern theoretisch errechneten Werten. Abbildung 56 zeigt die empirischen Werte der Untersuchung gemeinsam mit der theoretischen Kurve

$$y = 2.6363\, x^{-0.7340}\, e^{-0.4269\, x}$$

Der Determinationskoeffizient ist hier besonders hoch. Auf dessen Basis kann der vermutete Zusammenhang vorläufig als bestätigt angenommen werden.

Die Anpassung der Funktion $y = ax^b$ ergab für die geschätzten Parameter a = 1.8886 und b = -0.2398. Der Determinationskoeffizient liegt nur bei $R^2 = 0.5384$**,** was auf eine mäßige Anpassung hindeutet. Der empirische Kurvenverlauf der Abbildung 57 entspricht nicht dem Modell, das von einem negativen b ausgeht (monoton fallende Werte).

Im Vergleich zu der rein hyperpolischen Funktion ist bei dem erweiterten Modell das Ergebnis sowohl mathematisch besser als auch optisch näher an den Daten.

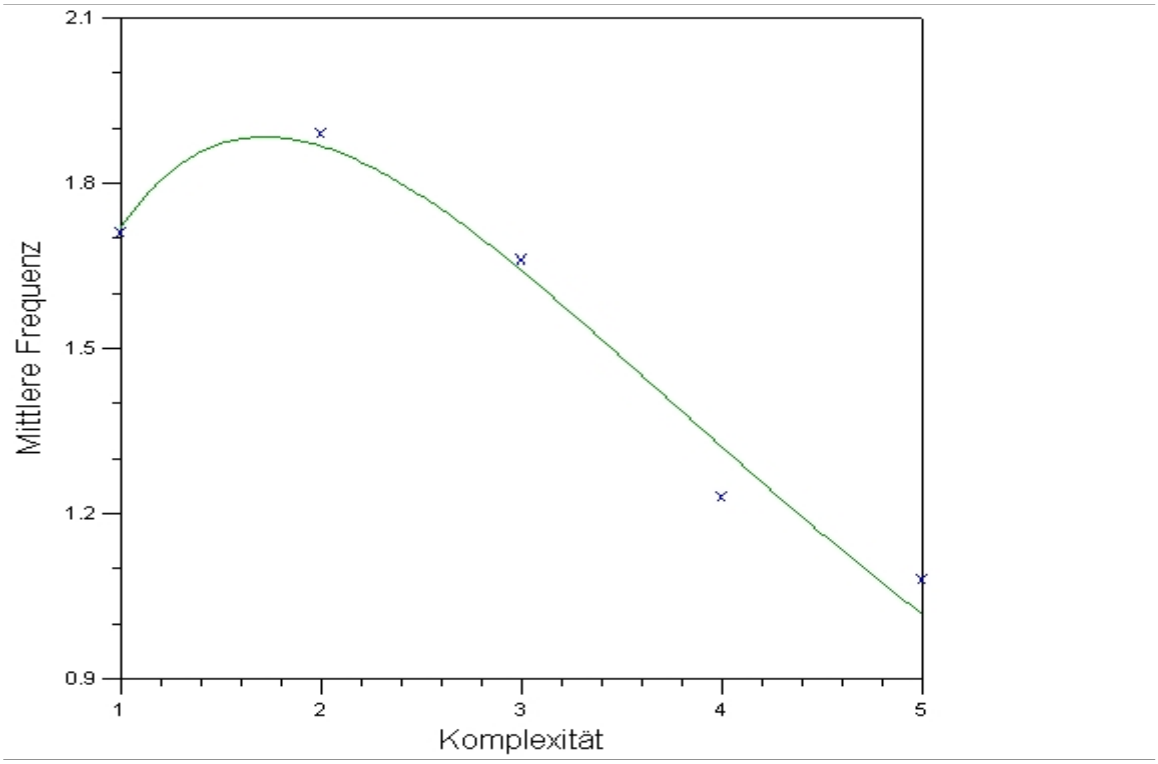

Abbildung 56: Abhängigkeit der Frequenz von Teilsätzen von deren Komplexität
Anpassung der Funktion $y = ax^b e^{cx}$

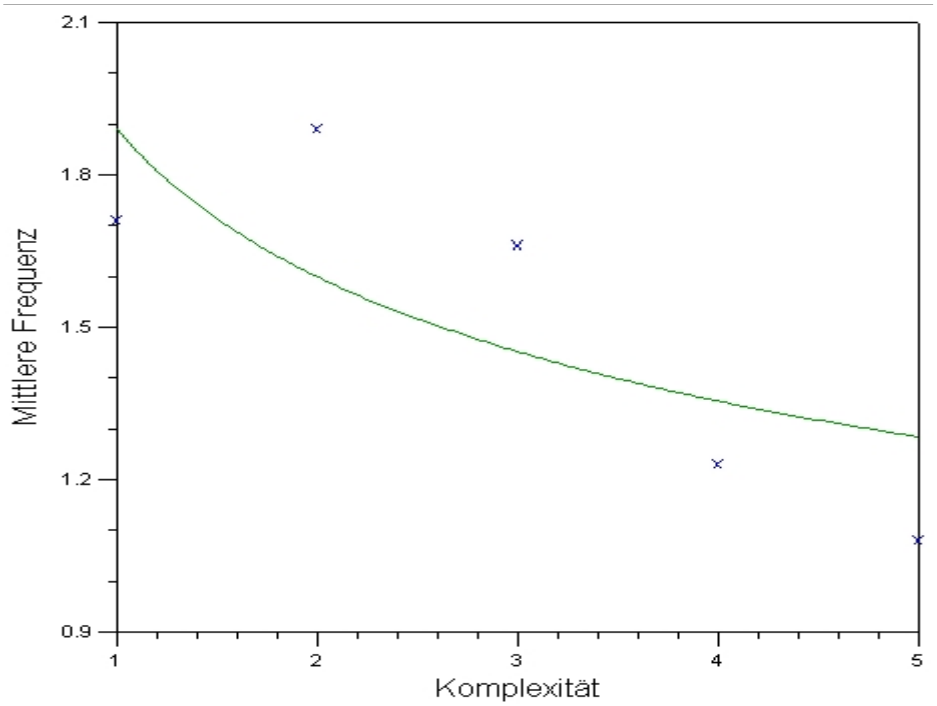

Abbildung 57: Abhängigkeit der Frequenz von Teilsätzen von deren Komplexität Anpassung der Funktion $y = ax^b$

6.4. Zusammenfassung der Testergebnisse

Funktionale Zusammenhänge zwischen Eigenschaften sprachlicher Einheiten sind empirisch überprüft worden. Die Untersuchungen waren für den Zusammenhang zwischen Satz- und Teilsatzlänge sowie zwischen Komplexität und Frequenz von Teilsätzen am erfolgreichsten. Sie erbrachten gute Anpassungen der theoretischen Funktionen an die Daten.

Die empirischen Überprüfungen der Zusammenhänge aus KÖHLER (1986) haben Folgendes erbracht:

Die Abhängigkeit der Polylexie von der Buchstabenlänge im Korpus und Text konnte mit guten bis sehr guten Ergebnissen nachgewiesen werden. Bei der Textuntersuchung führten die Einzelanalysen für die Wortarten *Substantiv*, *Adjektiv* und *Verb* ebenfalls zu sehr guten Anpassungsergebnissen. Die Untersuchung des Zusammenhangs zwischen Polylexie und Länge im Lexikon erbrachte bei der Buchstabenlänge eine mäßig gute Anpassung, bei der Silbenlänge eine gute Anpassung der theoretischen Kurve an die Daten.

Als Gesamtergebnis kann man festhalten, dass die Abhängigkeit der Polylexie von der Länge bei den zwei Längenmessungen bestätigt werden konnte.

Die Abhängigkeit der Länge von der Frequenz konnte nicht vollständig bestätigt werden. Bei der Korpus- und Textuntersuchung erbrachte die Anpassung der Daten keinen guten, jedoch einen befriedigenden Wert für den Determinationskoeffizienten R^2. Bei der Lexikonuntersu-

chung erreichte R^2 nicht einmal einen Wert von $R^2 \geq 0.30$. Der Zusammenhang wurde jedoch nicht für gescheitert erklärt, da eine Anwendung der Methode der gleitenden Mittelwerte eine sehr gute Anpassung an die theoretische Funktion erbrachte. Als Gesamtergebnis kann man festhalten: Die Länge-Frequenz-Relation von Wörtern konnte an den untersuchten Daten auf Grund der Anpassungsgüte weder abgelehnt noch eindeutig bestätigt werden. Hier sind weitere Untersuchungen notwendig.

Die Abhängigkeit der Polylexie von der Frequenz ist auch nicht vollständig bestätigt worden. Die Untersuchung im Korpus und Text lieferte akzeptable Anpassungen. Die Analyseergebnisse im Lexikon waren dagegen schlecht: Erzielt wurde ein Determinationskoeffizient von nur $R^2 = 0.3291$. Durch die Anwendung der Methode der gleitenden Mittelwerte konnten jedoch Daten erzeugt werden, an die sich die theoretische Funktion gut anpassen lässt.

Als Gesamtergebnis gilt auch hier, dass die Abhängigkeit der Polylexie von der Frequenz an den untersuchten Daten weder abgelehnt noch eindeutig bestätigt werden konnte.

7. Schlussfolgerung und Überblick

Ziel der vorliegenden Arbeit war es, einige Gesetzmäßigkeiten aus der quantitativen und synergetischen Linguistik anhand von Daten des Französischen zu überprüfen. Diese Gesetzmäßigkeiten betrafen zweidimensionale funktionale Abhängigkeiten zwischen den Eigenschaften Länge, Frequenz und Polylexie einerseits und die Häufigkeitsverteilungen dieser Größen andererseits. Auch wenn keine umfassende Überprüfung dieser Gesetzeshypothesen geleistet werden konnte, lassen die erzielten Ergebnisse die Interpretation zu, dass in allen Fällen durch theoretische Modelle das Wirken von Gesetzen nachgewiesen werden kann, die aus theoretischen Überlegungen erwachsen sind.

Neben der empirischen Überprüfung der Gesetzeshypothesen wurde auch untersucht, ob die Textsorte bzw. die Art der Stichprobe bzw. Datenbasis einen Einfluss auf die Modellierung hat. In Anbetracht der erzielten Ergebnisse kann man davon ausgehen, dass die Stichprobenart sich nicht signifikant auf die Modellierung auswirkt.

Die Modellierung der Ranghäufigkeitsverteilung von Wörtern in literarischen Texten durch die Zipf-Mandelbrot-Verteilung führte nicht in allen Fällen zu einem guten Ergebnis. Für die Worthäufigkeitsverteilung in Zeitungstexten stellte sich diese Verteilung als sehr geeignetes Modell dar. Hinsichtlich dieser Verteilung ist aber eine Interpretation der Ergebnisse unter Berücksichtigung der beiden ausgewählten Texttypen nicht leicht. Bei der Zipf-Mandelbrot-Verteilung spielt die Abgeschlossenheit des untersuchten Textes eine entscheidende Rolle.

ORLOV (1988a, b) spricht hier von dem Zipf'schen Umfang. Dass diese Verteilung für Zeitungstexte ein geeignetes Modell darstellt, wurde auf die relative Kürze der Texte zurückgeführt.

Die Untersuchungsergebnisse haben Fragen und Probleme aufgeworfen, die in der vorliegenden Arbeit nicht geklärt werden konnten, jedoch eine Anregung für weitere empirische Untersuchungen sein können. So wäre beispielsweise zu überprüfen, ob die für die Modellierung der Buchstabenhäufigkeiten verwendeten Modelle nur für die untersuchten Texte oder ganz generell für das Französische ungeeignet sind. Zur Klärung dieser Frage ist die Ausdehnung der durchgeführten Untersuchungen auf weitere Textsorten erforderlich.
Es wäre nach Erarbeitung weiterer Daten auch zu klären, ob in Bezug auf die Buchstabenhäufigkeitsverteilung in französischen Texten Einflussfaktoren wie Autor, Entstehungszeit oder Funktionalstil möglicherweise dazu zwingen, andere Verteilungen zu entwerfen. Eine Klärung dieser Frage würde ermöglichen, zuverlässige Aussagen über die Abweichungen der empirischen Werte von den Modellvoraussagen zu machen.

Eine weitere noch zu klärende Frage ist die bei der Überprüfung der Zusammenhänge zwischen Länge und Frequenz und Polylexie und Frequenz aufgetretenen Abweichungen der empirischen Daten von den theoretischen Kurven. Zu klären ist, ob es sich dabei um eine zufällige Streuung der Datenpunkte um die theoretische Funktion oder um Oszillationen handelt, oder ob die Streuung durch einen Faktor hervorgerufen wurde, der in den bisherigen Betrachtungen noch keine Berücksichtigung fand. Das Vorhandensein von Oszillationen wurde zwar angenommen, konnte jedoch nicht bewiesen werden.

Es wäre auch wünschenswert, Gesetzeshypothesen an typologisch möglichst sich unterscheidenden Sprachen zu überprüfen, d.h. nicht nur an Sprachen wie Deutsch, Englisch, Russisch, Französisch, etc. Verhältnismäßig gut überprüft ist z.B. das Gesetz, dem die Wortlängenverteilung folgt. Dennoch findet beispielsweise keine einzige Sprache der Sprachgruppen Afrikas Berücksichtigung.

Ein weiterer Aspekt zukünftiger Arbeit sollte in der Weitererprobung des von KÖHLER (1986) entwickelten synergetischen Modells der Lexik sowie in dessen Weiterentwicklung bestehen, sodass viele der bisher erfolgreich überprüften Gesetzeshypothesen darin integriert werden können.

Anhang

A. Untersuchungsmaterial

Tabelle 1: Zeitungstexte

Nr.	Quellen	Texte	Kapitel
1	http://www.lemonde.fr	Les autres victimes d'une négation	Gesamttext
2	http://www.lemonde.fr	La Tchétchénie selon Anna Politkovskaïa	Gesamttext
3	http://www.lemonde.fr	L'innommable Raphaël Confiant **?**	Gesamttext
4	http://www.lemonde.fr	L'air du protectionnisme électoral	Gesamttext
5	http://www.lemonde.fr	François Bayrou, candidat contre la "guerre des deux camps"	Gesamttext
6	http://www.lemonde.fr	Le Dollar ? Et le Yen, le yuan, les autres?	Gesamttext
7	http://www.lemonde.fr	EADS tente de se relancer avec l'A350 XWB	Gesamttext
8	http://www.lemonde.fr	Les Touments de la chaîne Imedi animent la campagne présidentielle en Georgie	Gesamttext
9	http://www.lemonde.fr	La "réponse pénale", nouvel indicateur phare	Gesamttext
10	http://www.lemonde.fr	Une année de l'histoire d'internet	Gesamttext
11	http://www.lemonde.fr	Continuité en vue à Madagascar	Gesamttext
12	http://www.lemonde.fr	La Norvège ne connaît plus ni chômage ni inflation	Gesamttext
13	http://www.lemonde.fr	Le camp pro-occidental perd une nouvelle bataille en Ukraine	Gesamttext
14	http://www.lemonde.fr	Laurent Wauquiez, une pointure chez Sarkozy	Gesamttext
15	http://www.lemonde.fr	Rencontre historique annulée entre les leaders nord-irlandais	Gesamttext
16	http://www.lemonde.fr	Les Européens attendent L'Allemagne dans le domaine de l'énergie	Gesamttext
17	http://www.lemonde.fr	Paris échoue à fédérer ses partenaires contre l'euro fort	Gesamttext
18	http://www.lemonde.fr	Valentin Paniagua, ancien président du Pérou	Gesamttext

Tabelle 2:: Literarische Texte

Nr.	Autoren	Texte	Kapitel
19	Banville, Théodore de	Le sang de la coupe	Vous en qui je salue une nouvelle aurore ... (Gesamtgedicht)
		Ondes funambulesques	Le saut du tremplin (Gesamtgedicht)
		Les cariatides	Bien souvent je revois sous mes paupières closes (Gesamtgedicht)
		Les Stalactites	Sculpteur, cherche avec soin ... (Gesamtgedicht)
20	Barbier, Auguste	Lambes et poèmes	L'idole (Gesamtgedicht)
		Lambes et Poèmes	Le Spleen (Gesamtgedicht)
		La cuve	Gesamtgedicht
21	Ackermann, Louise	Poésies philosophiques	L'amour et la mort (Gesamtgedicht)
22	Gauthier, Théophile	Aria Marcella : Souvenir de Pompéi	Gesamttext
23	Gauthier, Théophile	La morte amoureuse	Gesamttext
24	Maupassant, Guy de	Le Horla	Gesamttext
25	Maupassant, Guy de	Le lit	Gesamttext
26	Stendhal	Vittoria Accoramboni: La Duchesse de Bracciano	Gesamttext
27	Chasles, Philarète	L'œil sans paupières	Gesamttext
28	Huysmans, Joris Karl	La Bièvre	Gesamttext
29	Eekhoud, Georges	La dernière lettre du matelot	Gesamttext
30	Gaboriau, Emile	Maudite maison	Gesamttext
31	Raboux, Charles	Le ministère public	Gesamttext
32	Allais, Alphonse	Amour d'escale	Gesamttext
33	Leroux, Gaston	L'homme qui a vu le diable	Gesamttext
34	Maupassant, Guy de	Boule de Suif	Text bearbeitet bis: "(...) en mettant la main dans la poche de leur culotte."
35	Denon, Dominique	Point de lendemain	Text bearbeitet bis: „ (…) Dès qu'il viendra du monde (et sans doute il en viendra) …"
36	Flaubert, Gustave	"Trois contes"	Un cœur simple (Kapp. 1 - 5)
37	Stendhal	La Chartreuse de Parme	Erstes Buch (Kap. 1 - 2)
38	Theodore, Moreux	La vie sur Mars	Teil 1 (Kap. 1)

39	Goncourt, Edmond de/ Goncourt, Jules de	Germine Lacerteux	Kap. 1
40	Dumas, Alexandre	Les Trois Mousquetaires	Kap. 1 - 4
41	Balzac, Honoré de	Les Chouans	Teil 1 (Kap. 1)
42	Balzac, Honoré de	La fille aux yeux d'or	Kap. 1
43	Verne, Jules	De la terre à lune	Kap. 1 - 3
44	Constant, Benjamin	Adolphe	Kap. 1 - 3
45	Nerval, Gérard de	Aurelia	Teil 1 (Kap. 1 – 10)
46	Maupassant, Guy de	Le conte sur l'eau	Gesamttext
47	Flaubert, Gustave	Madame Bovary	Gesamttext

B. Empirische und theoretische Daten zu den Häufigkeitsverteilungen

Tabelle 3.1: Verteilung der Bedeutungen von Nominalsuffixen in *„Madame Bovary"* von G. Flaubert (Gesamtberechnung)

Rang	n_x	1-verschobene NB	0-gestutzte NB
1	131	128.2295	120.9931
2	82	87.9158	88.7991
3	69	68.0907	69.2413
4	58	54.7536	55.5779
5	57	44.8401	45.3749
6	28	37.12	37.4609
7	27	30.949	31.1724
8	26	25.935	26.0927
9	21	21.8155	21.9404
10	20	18.4043	18.516
11	12	15.5628	15.6723
12	12	13.1852	13.298
13	11	11.1886	11.3068
14	10	9.5071	9.6309
15	9	8.0875	8.216
16	8	6.8868	7.0184
17	6	5.8694	6.0024
18	6	5.0062	5.1389
19	5	4.2729	4.4038
20	5	3.6491	3.777
21	5	3.1182	3.2419
22	4	2.6658	2.7845
23	4	2.2801	2.3932
24	3	1.951	2.058
25	2	1.67	1.7708
26	1	1.43	1.5244
27	1	8.6159	9.5943

Tabelle 3.2: Verteilung der Bedeutungen einzelner Nominalsuffixe in „Madame Bovary“ von G. Flaubert

Suffixe	Vorkommen im Text	1-verschobene NB	0-Gestutzte NB
-ance/-ence	82	82.5882	83.4535
	58	47.2979	45.6964
	6	17.2627	17.6619
	5	5.1053	5.4041
	3	1.7459	1.784
-ement	131	123.6615	125.6363
	28	34.9305	34.8771
	12	13.6838	12.824
	8	5.859	5.2929
	4	4.8652	4.3696
-erie	26	23.8899	26.0877
	21	22.8865	20.7524
	12	13.1037	12.227
	6	5.819	5.9427
	4	3.3009	3.9902
-esse(s)	57	51.2798	53.3579
	11	18.684	17.0761
	9	7.2477	6.5847
	5	4.7885	4.9812
-(u)ité	20	19.918	19.9861
	5	5.1115	5.0357
	1	1.4114	1.398
	1	0.5591	0.5802
-(at)ure	69	67.3587	67.9356
	27	30.1291	29.1686
	10	8.2932	8.5518
	2	2.2191	2.3441

Tabelle 4.1: Ranghäufigkeitsverteilung von Wörtern im Zeitungstext 5 und ihre Modellierung durch die Zipf-Mandelbrot-Verteilung

Rang	f_x	NP_x
1	36	31.6473
2	25	24.5072
3	19	20.1519
4	18	17.1974
5	15	15.0515
6	14	13.4168
7	14	12.1267
8	13	11.0807
9	12	10.2141
10	12	9.4834
11	11	8.8583
12	10	8.317
13	8	7.8433
14	8	7.425

15	7	7.0526
16	7	6.7189
17	6	6.418
18	6	6.1451
19	6	5.8965
20	6	5.6689
21	6	5.4597
22	5	5.2667
23	5	5.088
24	5	4.9222
25	5	4.7677
26	5	4.6235
27	4	4.4885
28	4	4.3619
29	4	4.2428
30	4	4.1307
31	3	4.0248
32	3	3.9248
33	3	3.83
34	3	3.7401
35	3	3.6547
36	3	3.5734
37	3	3.496
38	3	3.4222
39	3	3.3518
40	3	3.2844
41	3	3.2199
42	3	3.1582
43	3	3.099
44	3	3.0421
45	3	2.9875
46	3	2.935
47	3	2.8845
48	2	2.8359
49	2	2.789
50	2	2.7438
51	2	2.7001
52	2	2.658
53	2	2.6173
54	2	2.5779
55	2	2.5398
56	2	2.5029
57	2	2.4671
58	2	2.4325
59	2	2.3989
60	2	2.3663
61	2	2.3347
62	2	2.304
63	2	2.2741
64	2	2.2451
65	2	2.2169

66	2	2.1894
67	2	2.1627
68	2	2.1367
69	2	2.1114
70	2	2.0867
71	2	2.0626
72	2	2.0392
73	2	2.0163
74	2	1.9939
75	2	1.9722
76	2	1.9509
77	2	1.9301
78	2	1.9098
79	2	1.89
80	2	1.8706
81	2	1.8516
82	2	1.8331
83	2	1.8149
84	2	1.7972
85	2	1.7798
86	1	1.7628
87	1	1.7461
88	1	1.7298
89	1	1.7138
90	1	1.6981
91	1	1.6828
92	1	1.6677
93	1	1.653
94	1	1.6385
95	1	1.6243
96	1	1.6103
97	1	1.5967
98	1	1.5832
99	1	1.57
100	1	1.5571
101	1	1.5444
102	1	1.5319
103	1	1.5196
104	1	1.5076
105	1	1.4957
106	1	1.4841
107	1	1.4727
108	1	1.4614
109	1	1.4503
110	1	1.4394
111	1	1.4287
112	1	1.4182
113	1	1.4078
114	1	1.3976
115	1	1.3876
116	1	1.3777

117	1	1.368
118	1	1.3584
119	1	1.349
120	1	1.3397
121	1	1.3305
122	1	1.3215
123	1	1.3126
124	1	1.3039
125	1	1.2952
126	1	1.2867
127	1	1.2783
128	1	1.2701
129	1	1.2619
130	1	1.2539
131	1	1.246
132	1	1.2382
133	1	1.2305
134	1	1.2229
135	1	1.2154
136	1	1.208
137	1	1.2007
138	1	1.1934
139	1	1.1863
140	1	1.1793
141	1	1.1724
142	1	1.1655
143	1	1.1588
144	1	1.1521
145	1	1.1455
146	1	1.139
147	1	1.1326
148	1	1.1263
149	1	1.12
150	1	1.1138
151	1	1.1077
152	1	1.1017
153	1	1.0957
154	1	1.0898
155	1	1.084
156	1	1.0782
157	1	1.0725
158	1	1.0669
159	1	1.0613
160	1	1.0558
161	1	1.0504
162	1	1.045
163	1	1.0397
164	1	1.0344
165	1	1.0292
166	1	1.024
167	1	1.019

168	1	1.0139
169	1	1.0089
170	1	1.004
171	1	0.9991
172	1	0.9943
173	1	0.9895
174	1	0.9848
175	1	0.9802
176	1	0.9755
177	1	0.971
178	1	0.9664
179	1	0.9619
180	1	0.9575
181	1	0.9531
182	1	0.9488
183	1	0.9444
184	1	0.9402
185	1	0.936
186	1	0.9318
187	1	0.9276
188	1	0.9235
189	1	0.9195
190	1	0.9155
191	1	0.9115
192	1	0.9075
193	1	0.9036
194	1	0.8997
195	1	0.8959
196	1	0.8921
197	1	0.8883
198	1	0.8846
199	1	0.8809
200	1	0.8773
201	1	0.8736
202	1	0.87
203	1	0.8665
204	1	0.8629
205	1	0.8594
206	1	0.856
207	1	0.8525
208	1	0.8491
209	1	0.8457
210	1	0.8424
211	1	0.8391
212	1	0.8358
213	1	0.8325
214	1	0.8293
215	1	0.8261
216	1	0.8229
217	1	0.8197
218	1	0.8166

219	1	0.8135
220	1	0.8104
221	1	0.8074
222	1	0.8044
223	1	0.8013
224	1	0.7984
225	1	0.7954
226	1	0.7925
227	1	0.7896
228	1	0.7867
229	1	0.7838
230	1	0.781
231	1	0.7782
232	1	0.7754
233	1	0.7726
234	1	0.7699
235	1	0.7672
236	1	0.7644
237	1	0.7618
238	1	0.7591
239	1	0.7565
240	1	0.7538
241	1	0.7512
242	1	0.7486
243	1	0.7461
244	1	0.7435
245	1	0.741
246	1	0.7385
247	1	0.736
248	1	0.7335
249	1	0.7311
250	1	0.7286
251	1	0.7262
252	1	0.7238
253	1	0.7214
254	1	0.7191
255	1	0.7167
256	1	0.7144
257	1	0.7121
258	1	0.7098
259	1	0.7075
260	1	0.7052
261	1	0.703
262	1	0.7007
263	1	0.6985
264	1	0.6963
265	1	0.6941
266	1	0.692
267	1	0.6898
268	1	0.6877
269	1	0.6855

270	1	0.6834
271	1	0.6813
272	1	0.6792
273	1	0.6771
274	1	0.6751
275	1	0.673
276	1	0.671
277	1	0.669
278	1	0.667
279	1	0.665
280	1	0.663
281	1	0.661
282	1	0.6591
283	1	0.6572
284	1	0.6552
285	1	0.6533
286	1	0.6514
287	1	0.6495

Tabelle 4.2: Ranghäufigkeitsverteilung von Wörtern im literarischen Text 21 und ihre Modellierung durch die Zipf-Mandelbrot-Verteilung

Rang	$\mathbf{f_x}$	$\mathbf{NP_x}$
1	77	66.1244
2	50	47.4115
3	31	37.21
4	29	30.7506
5	26	26.2767
6	23	22.9864
7	23	20.4599
8	21	18.4562
9	21	16.8262
10	18	15.4731
11	16	14.331
12	13	13.3534
13	12	12.5068
14	11	11.766
15	11	11.1122
16	11	10.5307
17	10	10.0099
18	10	9.5407
19	10	9.1156
20	9	8.7286
21	9	8.3748
22	8	8.05
23	8	7.7507
24	8	7.4739
25	7	7.2172
26	7	6.9785
27	6	6.7558

28	6	6.5477
29	6	6.3526
30	6	6.1694
31	6	5.997
32	5	5.8345
33	5	5.681
34	5	5.5358
35	4	5.3983
36	4	5.2677
37	4	5.1437
38	4	5.0256
39	4	4.9132
40	4	4.8059
41	4	4.7034
42	4	4.6055
43	4	4.5118
44	4	4.422
45	4	4.3359
46	4	4.2533
47	4	4.1739
48	3	4.0976
49	3	4.0242
50	3	3.9535
51	3	3.8853
52	3	3.8196
53	3	3.7562
54	3	3.695
55	3	3.6359
56	3	3.5787
57	3	3.5234
58	3	3.4699
59	3	3.418
60	3	3.3678
61	3	3.3191
62	3	3.2719
63	3	3.226
64	3	3.1815
65	3	3.1383
66	3	3.0963
67	3	3.0555
68	3	3.0158
69	3	2.9771
70	3	2.9395
71	3	2.9029
72	3	2.8672
73	3	2.8325
74	3	2.7986
75	3	2.7656
76	2	2.7333
77	2	2.7019
78	2	2.6712

79	2	2.6413
80	2	2.612
81	2	2.5834
82	2	2.5555
83	2	2.5282
84	2	2.5015
85	2	2.4754
86	2	2.4499
87	2	2.4249
88	2	2.4005
89	2	2.3765
90	2	2.3531
91	2	2.3301
92	2	2.3076
93	2	2.2856
94	2	2.264
95	2	2.2429
96	2	2.2221
97	2	2.2018
98	2	2.1818
99	2	2.1622
100	2	2.143
101	2	2.1241
102	2	2.1056
103	2	2.0875
104	2	2.0696
105	2	2.0521
106	2	2.0349
107	2	2.018
108	2	2.0013
109	2	1.985
110	2	1.9689
111	2	1.9532
112	2	1.9376
113	2	1.9224
114	2	1.9074
115	2	1.8926
116	2	1.8781
117	2	1.8638
118	2	1.8497
119	2	1.8359
120	2	1.8223
121	2	1.8089
122	2	1.7957
123	2	1.7827
124	2	1.7699
125	2	1.7572
126	2	1.7448
127	2	1.7326
128	2	1.7205
129	2	1.7086

130	2	1.6969
131	2	1.6854
132	2	1.674
133	2	1.6628
134	2	1.6517
135	2	1.6408
136	2	1.6301
137	2	1.6195
138	2	1.609
139	2	1.5987
140	2	1.5885
141	2	1.5784
142	2	1.5685
143	2	1.5588
144	2	1.5491
145	1	1.5396
146	1	1.5302
147	1	1.5209
148	1	1.5117
149	1	1.5027
150	1	1.4937
151	1	1.4849
152	1	1.4762
153	1	1.4676
154	1	1.4591
155	1	1.4507
156	1	1.4424
157	1	1.4342
158	1	1.4261
159	1	1.4181
160	1	1.4102
161	1	1.4023
162	1	1.3946
163	1	1.387
164	1	1.3794
165	1	1.372
166	1	1.3646
167	1	1.3573
168	1	1.3501
169	1	1.3429
170	1	1.3359
171	1	1.3289
172	1	1.322
173	1	1.3152
174	1	1.3084
175	1	1.3017
176	1	1.2951
177	1	1.2886
178	1	1.2821
179	1	1.2757
180	1	1.2694

181	1	1.2631
182	1	1.2569
183	1	1.2508
184	1	1.2447
185	1	1.2387
186	1	1.2328
187	1	1.2269
188	1	1.221
189	1	1.2153
190	1	1.2096
191	1	1.2039
192	1	1.1983
193	1	1.1927
194	1	1.1873
195	1	1.1818
196	1	1.1764
197	1	1.1711
198	1	1.1658
199	1	1.1606
200	1	1.1554
201	1	1.1502
202	1	1.1451
203	1	1.1401
204	1	1.1351
205	1	1.1302
206	1	1.1253
207	1	1.1204
208	1	1.1156
209	1	1.1108
210	1	1.1061
211	1	1.1014
212	1	1.0967
213	1	1.0921
214	1	1.0876
215	1	1.083
216	1	1.0786
217	1	1.0741
218	1	1.0697
219	1	1.0653
220	1	1.061
221	1	1.0567
222	1	1.0525
223	1	1.0482
224	1	1.0441
225	1	1.0399
226	1	1.0358
227	1	1.0317
228	1	1.0277
229	1	1.0236
230	1	1.0197
231	1	1.0157

232	1	1.0118
233	1	1.0079
234	1	1.0041
235	1	1.0002
236	1	0.9964
237	1	0.9927
238	1	0.9889
239	1	0.9852
240	1	0.9816
241	1	0.9779
242	1	0.9743
243	1	0.9707
244	1	0.9672
245	1	0.9636
246	1	0.9601
247	1	0.9566
248	1	0.9532
249	1	0.9498
250	1	0.9464
251	1	0.943
252	1	0.9396
253	1	0.9363
254	1	0.933
255	1	0.9297
256	1	0.9265
257	1	0.9233
258	1	0.9201
259	1	0.9169
260	1	0.9137
261	1	0.9106
262	1	0.9075
263	1	0.9044
264	1	0.9013
265	1	0.8983
266	1	0.8953
267	1	0.8923
268	1	0.8893
269	1	0.8863
270	1	0.8834
271	1	0.8805
272	1	0.8776
273	1	0.8747
274	1	0.8718
275	1	0.869
276	1	0.8662
277	1	0.8634
278	1	0.8606
279	1	0.8578
280	1	0.8551
281	1	0.8524
282	1	0.8497

283	1	0.847
284	1	0.8443
285	1	0.8416
286	1	0.839
287	1	0.8364
288	1	0.8338
289	1	0.8312
290	1	0.8286
291	1	0.8261
292	1	0.8236
293	1	0.821
294	1	0.8185
295	1	0.8161
296	1	0.8136
297	1	0.8111
298	1	0.8087
299	1	0.8063
300	1	0.8039
301	1	0.8015
302	1	0.7991
303	1	0.7967
304	1	0.7944
305	1	0.792
306	1	0.7897
307	1	0.7874
308	1	0.7851
309	1	0.7829
310	1	0.7806
311	1	0.7783
312	1	0.7761
313	1	0.7739
314	1	0.7717
315	1	0.7695
316	1	0.7673
317	1	0.7651
318	1	0.763
319	1	0.7608
320	1	0.7587
321	1	0.7566
322	1	0.7545
323	1	0.7524
324	1	0.7503
325	1	0.7482
326	1	0.7462
327	1	0.7441
328	1	0.7421
329	1	0.7401
330	1	0.7381
331	1	0.7361
332	1	0.7341
333	1	0.7321

334	1	0.7302
335	1	0.7282
336	1	0.7263
337	1	0.7243
338	1	0.7224
339	1	0.7205
340	1	0.7186
341	1	0.7167
342	1	0.7148
343	1	0.713
344	1	0.7111
345	1	0.7093
346	1	0.7074
347	1	0.7056
348	1	0.7038
349	1	0.702
350	1	0.7002
351	1	0.6984
352	1	0.6966
353	1	0.6948
354	1	0.6931
355	1	0.6913
356	1	0.6896
357	1	0.6879
358	1	0.6861
359	1	0.6844
360	1	0.6827
361	1	0.681
362	1	0.6793
363	1	0.6777
364	1	0.676
365	1	0.6743
366	1	0.6727
367	1	0.671
368	1	0.6694
369	1	0.6678

Tabelle 5.1: Empirische Verteilung der Buchstaben in den Einzeltexten (I = 26)

Buchstaben	**Texte**												
	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
a	364	594	220	4184	4218	3003	317	3428	2501	1030	2328	645	1540
b	74	107	50	516	523	388	58	275	290	212	274	93	152
c	166	239	106	1683	1499	1181	175	1336	875	436	902	238	589
d	231	277	147	2068	1904	1303	232	1623	1079	731	1166	351	709
e	1039	1435	717	9144	9881	7047	1120	7201	5303	2792	5287	1488	3286
f	74	112	48	566	577	471	55	540	367	133	338	70	194
g	41	80	19	485	468	340	46	366	286	173	324	57	148
h	60	78	30	436	418	384	49	278	230	139	253	67	136

i	362	478	271	3761	3895	3011	371	3309	2289	840	2119	733	1370
g	39	50	21	154	583	529	61	123	163	30	249	28	57
k	1	0	0	2	0	0	0	0	63	1	13	0	0
l	366	461	224	2984	2938	2183	321	2405	1989	883	1589	469	1147
m	168	291	118	1442	1883	1499	220	1169	846	308	898	288	516
n	372	569	318	3599	3634	2737	368	2965	2155	938	2053	655	1327
o	350	518	273	2656	2760	2327	324	2195	1478	669	1606	516	954
p	136	197	101	1495	1435	1088	137	1310	814	372	845	254	599
q	60	64	54	575	596	505	80	504	286	117	335	105	278
r	401	637	304	3502	3463	2627	348	2989	1927	966	2065	672	1276
s	607	801	387	4388	4218	3424	540	3140	2563	1501	2530	701	1474
t	329	595	305	3783	3890	2718	369	3233	2117	911	2033	606	1341
u	412	621	307	3394	3537	2895	412	2661	1976	917	1918	490	1365
v	116	167	103	863	982	726	100	680	459	197	448	163	327
w	2	0	0	3	0	0	0	0	21	0	6	0	0
x	41	42	13	253	204	143	34	194	114	69	165	27	117
y	20	12	4	135	141	96	14	71	90	31	50	34	51
z	14	14	16	46	47	66	14	26	46	10	23	16	9

Tabelle 5.1 (Fortsetzung)

Buchstaben	**Texte**													
	32	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
a	389	2901	1325	2300	4728	5252	1216	621	5507	4040	3518	2396	2632	3909
b	53	287	188	243	608	611	160	65	470	512	410	429	231	339
c	188	995	465	908	1698	1786	508	222	1986	1775	1477	1069	1061	1360
d	190	1231	628	1055	1961	2139	642	269	2358	2131	1736	1188	1145	1757
e	937	6259	2793	5370	9176	9702	2984	1205	9353	9242	8948	5251	6676	8410
f	63	388	224	296	651	736	195	88	596	541	589	312	372	503
g	45	311	206	200	572	605	140	50	661	649	486	321	251	394
h	45	376	138	236	477	496	97	68	635	536	427	274	182	289
i	382	2759	1113	2306	3988	4452	1187	522	4484	3677	3527	2225	2705	3612
g	15	413	29	312	156	283	67	70	313	113	235	106	411	441
k	21	43	1	0	3	3	6	0	1	0	1	19	0	0
l	353	2048	895	1541	3694	3256	907	417	3188	2894	2954	1743	1790	2441
m	195	1262	468	1107	1447	1646	479	257	1969	1521	1553	859	1342	1714
n	371	2492	1197	2251	3649	3807	1282	449	4400	3765	3541	2482	2621	3270
o	290	2190	888	1797	2670	2876	1040	430	3499	2772	2599	1595	1966	2323
p	148	957	440	827	1432	1584	530	190	1673	1438	1365	852	1057	1217
q	74	481	147	435	422	713	203	75	841	586	569	314	516	569
r	277	2266	1095	1950	3437	3793	1106	460	4140	3734	3185	2017	2394	3027
s	429	2717	1387	2486	4220	4519	1604	547	4618	4597	4434	2634	2972	4235
t	391	2587	1255	2205	3948	4047	1209	546	4751	3755	3361	2223	2545	3376
u	323	2504	1019	2112	3367	3488	1048	525	4160	3293	3190	1862	2295	2999
v	73	730	270	607	849	923	250	146	1179	707	729	404	693	819
w	1	0	0	0	0	2	2	0	0	4	1	13	0	1
x	26	131	60	86	194	234	72	30	206	279	230	164	121	204
y	18	121	38	77	118	132	40	24	125	143	165	100	70	130
z	7	114	5	87	44	62	21	15	159	52	65	39	49	18

Tabelle 5.2: Empirische Verteilung der Buchstaben in den Einzeltexten (I = 38)

Buchstaben	Texte												
	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
a	356	579	214	4118	4130	2945	307	3389	2474	999	2283	634	1520
b	74	107	50	516	523	388	58	275	290	212	274	93	152
c	166	238	106	1659	1470	1165	173	1297	859	429	886	232	573
d	231	277	147	2068	1904	1303	232	1623	1079	731	1166	351	709
e	917	1287	608	7799	8479	6188	990	6085	4511	2431	4569	1290	2766
f	74	112	48	566	577	471	55	540	367	133	338	70	194
g	41	80	19	485	468	340	46	366	286	173	324	57	148
h	60	78	30	436	418	384	49	278	230	139	253	67	136
i	359	476	263	3723	3873	2985	369	3284	2279	826	2094	726	1360
j	39	50	21	154	583	529	61	123	163	30	249	28	57
k	1	0	0	2	0	0	0	0	63	1	13	0	0
l	366	461	224	2984	2938	2183	321	2405	1989	883	1589	469	1147
m	168	291	118	1442	1883	1499	220	1169	846	308	898	288	516
n	372	569	318	3599	3634	2737	368	2965	2155	938	2053	655	1327
o	346	507	269	2616	2717	2280	321	2173	1466	656	1587	510	944
p	136	197	101	1495	1435	1088	137	1310	814	372	845	254	599
q	60	64	54	575	596	505	80	504	286	117	335	105	278
r	401	637	304	3502	3463	2627	348	2989	1927	966	2065	672	1276
s	607	801	387	4388	4218	3424	540	3140	2563	1501	2530	701	1474
t	329	595	305	3783	3890	2718	369	3233	2117	911	2033	606	1341
u	394	610	303	3347	3485	2838	408	2617	1942	898	1887	487	1340
v	116	167	103	863	982	726	100	680	459	197	448	163	327
w	2	0	0	3	0	0	0	0	21	0	6	0	0
x	41	42	13	253	204	143	34	194	114	69	165	27	117
y	20	12	4	135	141	96	14	71	90	31	50	34	51
z	14	14	16	46	47	66	14	26	46	10	23	16	9
à	1	7	2	9	25	24	8	15	12	11	18	5	9
â	7	8	4	57	63	34	2	24	15	20	27	6	11
é	88	105	82	1043	1051	653	93	852	633	273	522	149	413
è	18	24	15	206	195	103	19	182	96	68	113	35	69
ê	0	19	12	95	156	102	18	77	62	20	78	14	38
ë	0	0	0	1	0	1	0	5	1	0	5	0	0
î	0	2	8	25	21	23	2	19	8	12	20	4	10
ï	0	0	0	13	1	3	0	6	2	2	5	3	0
ô	0	11	4	40	43	47	3	22	12	13	19	6	10
ù	0	7	2	18	24	15	3	10	15	11	13	3	12
û	0	4	2	29	28	42	1	34	19	8	18	0	13
ç	0	1	0	24	29	16	2	39	16	7	16	6	16

Tabelle 5.2 (Fortsetzung)

Buchstaben	Texte													
	32	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
a	379	2829	1314	2237	4675	4832	1114	604	5106	3951	3457	2364	2600	3678
b	53	287	188	243	608	611	160	65	470	512	410	429	231	339
c	184	975	445	890	1460	1733	498	209	1939	1737	1454	1051	1044	1334
d	190	1231	628	1055	1961	2139	642	269	2358	2131	1736	1188	1145	1757
e	813	5482	2402	4665	7949	8334	2510	1059	8932	7973	7810	4464	5579	7253
f	63	388	224	296	651	736	195	88	596	541	589	312	372	503
g	45	311	206	200	572	605	140	50	661	649	486	321	251	394
h	45	376	138	236	477	496	97	68	635	536	427	274	182	289
i	369	2740	1103	2276	3948	4412	1179	514	4443	3654	3498	2211	2685	3574
j	15	413	29	312	156	283	67	70	313	113	235	106	411	441
k	21	43	1	0	3	3	6	0	1	0	1	19	0	0
l	353	2048	895	1541	3694	3256	907	417	3188	2894	2954	1743	1790	2441
m	195	1262	468	1107	1447	1646	479	257	1969	1521	1553	859	1342	1714
n	371	2492	1197	2251	3649	3807	1282	449	4400	3765	3541	2482	2621	3270
o	288	2149	873	1775	2643	2833	1029	425	3418	2750	2578	1589	1948	2304
p	148	957	440	827	1432	1584	530	190	1673	1438	1365	852	1057	1217
q	74	481	147	435	422	713	203	75	841	586	569	314	516	569
r	277	2266	1095	1950	3437	3793	1106	460	4140	3734	3185	2017	2394	3027
s	429	2717	1387	2486	4220	4519	1604	547	4618	4597	4434	2634	2972	4235
t	391	2587	1255	2205	3948	4047	1209	546	4751	3755	3361	2223	2545	3376
u	321	2471	1008	2088	3321	3444	1037	516	4094	3231	3135	1838	2272	2939
v	73	730	270	607	849	923	250	146	1179	707	729	404	693	819
w	1	0	0	0	0	2	2	0	0	4	1	13	0	1
x	26	131	60	86	194	234	72	30	206	279	230	164	121	204
y	18	121	38	77	118	132	40	24	125	143	165	100	70	130
z	7	114	5	87	44	62	21	15	159	52	65	39	49	18
à	3	44	4	21	13	355	99	13	336	54	31	132	7	192
â	7	28	7	42	40	65	3	4	65	35	30	10	25	39
é	93	570	311	544	889	1097	353	105	1158	945	860	653	830	922
è	17	129	42	85	219	176	92	23	157	205	178	92	145	115
ê	14	78	37	76	117	95	29	17	145	113	99	42	92	119
ë	0	0	1	0	2	0	0	1	1	6	1	0	0	1
î	13	16	7	29	38	31	8	8	35	17	23	10	18	24
ï	0	3	3	1	2	9	0	0	6	6	6	4	2	14
ô	2	41	15	22	27	43	11	5	81	22	21	6	18	19
ù	0	14	8	11	22	17	9	6	27	47	36	5	11	51
û	2	19	3	13	24	27	2	3	39	15	19	19	22	9
ç	4	20	20	18	58	53	9	13	47	38	23	18	17	26

Tabelle 6.1 Empirische Verteilung der Phoneme (I = 35)

Korpus AUV 1, 1				Korpus AUV 1, 2			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**
1	1292.244	19	388.692	1	808.016	19	257.959
2	1173.2824	20	362.861	2	794.954	20	244.018
3	1079.447	21	354.778	3	737.983	21	210.977
4	1000.373	22	346.871	4	709.992	22	197.037
5	992.466	23	263.931	5	617.017	23	166.082
6	905.485	24	228.787	6	583.976	24	163.008
7	899.510	25	201.902	7	519.980	25	162.020
8	890.548	26	177.477	8	509.003	26	120.966
9	850.484	27	163.946	9	509.003	27	103.952
10	808.487	28	153.930	10	456.972	28	92.975
11	721.506	29	121.598	11	442.043	29	90.99933
12	651.569	30	100.863	12	429.969	30	87.047
13	618.710	31	82.939	13	336.993	31	70.033
14	590.594	32	78.371	14	319.979	32	34.028
15	572.671	33	68.530	15	307.026	33	33.040
16	470.7538	34	57.987	16	300.989	34	26.015
17	461.7921	35	43.93	17	269.046	35	5.049
18	404.6831			18	258.947		

Tabelle 6.1 (Fortsetzung)

Korpus AUV 1, 3				Korpus AUV 1, 4			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**
1	680.965	19	176.007	1	235.011	19	68.991
2	647.006	20	149.963	2	200.99	20	64.011
3	544.023	21	149.963	3	193.997	21	56.014
4	485.978	22	142.984	4	183.011	22	49.012
5	477.977	23	125.026	5	165.99	23	46.990
6	456.019	24	109.962	6	163.998	24	46.990
7	449.040	25	86.982	7	146.010	25	32.986
8	425.975	26	85.024	8	135.991	26	30.994
9	413.975	27	84.003	9	135.991	27	30.994
10	377.973	28	73.024	10	135.991	28	23.993
11	361.036	29	64.002	11	126.001	29	21.005

12	334.992	30	58.981	12	125.005	30	20.009
13	285.033	31	49.959	13	105.992	31	19.013
14	256.010	32	46.980	14	98.990	32	19.013
15	218.988	33	41.959	15	89.996	33	14.999
16	213.966	34	32.001	16	74.001	34	9.989
17	205.029	35	8.000	17	72.009	35	4.013
18	190.986			18	69.987		

Tabelle 6.1 (Fortsetzung)

Korpus ORL 14			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**
1	2032.042	19	454.925
2	1586.913	20	398.059
3	1408.010	21	395.077
4	1324.948	22	352.055
5	1177.992	23	313.080
6	1101.958	24	313.080
7	1018.044	25	312.015
8	949.038	26	280.920
9	890.043	27	248.973
10	888.978	28	198.923
11	822.954	29	156.966
12	746.068	30	132.047
13	674.081	31	107.980
14	647.033	32	104.999
15	566.100	33	96.053
16	530.959	34	86.043
17	510.087	35	8.093
18	464.083		

Tabelle 6.2 Empirische Verteilung der Phoneme (I = 36)

Korpus Szklarczyk				Korpus Wioland			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**
1	21664.423	19	5853.698	1	6301.632	19	1561.810
2	20487.944	20	5566.752	2	5889.811	20	1530.729
3	16040.281	21	5308.501	3	4576.647	21	1367.555
4	15581.167	22	5193.722	4	4467.865	22	1219.921
5	15351.611	23	4304.19	5	4312.461	23	1204.381
6	15064.665	24	3758.992	6	4188.137	24	1072.287
7	14691.635	25	3443.352	7	4102.665	25	994.585
8	14490.773	26	3414.657	8	3947.261	26	901.343
9	14146.437	27	3156.406	9	3294.564	27	839.181
10	11736.091	28	3127.711	10	3038.148	28	800.330
11	11449.145	29	2094.705	11	3014.837	29	473.982
12	10731.780	30	2037.316	12	2913.825	30	435.131
13	10559.612	31	1779.065	13	2634.097	31	419.590
14	9210.966	32	1549.508	14	2494.234	32	396.280
15	8292.739	33	1463.424	15	2400.991	33	341.888
16	8292.739	34	1319.951	16	2331.06	34	287.497
17	8206.655	35	1233.867	17	2035.792	35	108.782
18	5968.476	36	344.335	18	1763.835	36	38.851

Tabelle 6.2 (Fortsetzung)

Korpus Haton/Lamotte				Korpus De Kock 1			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**	**Rang**	**Häufigkeit**
1	4187.762	19	1080.712	1	95443.558	19	27317.276
2	3752.475	20	1065.702	2	84789.264	20	23826.117
3	3397.240	21	720.475	3	84552.811	21	21225.134
4	3357.214	22	615.405	4	84024.269	22	19973.324
5	2951.947	23	590.389	5	82911.549	23	18874.513
6	2911.920	24	535.353	6	82605.551	24	16482.165
7	2711.788	25	495.326	7	56832.174	25	16287.439
8	2466.626	26	470.310	8	55941.998	26	11224.563

9	2406.587	27	375.247	9	53090.653	27	11029.837
10	2256.488	28	300.198	10	49154.406	28	9833.663
11	1971.300	29	280.184	11	47026.329	29	9277.303
12	1951.287	30	270.178	12	41421.002	30	8985.214
13	1866.230	31	245.161	13	41407.093	31	8067.22
14	1490.983	32	215.141	14	39626.741	32	7093.59
15	1315.867	33	100.066	15	33395.509	33	5563.6
16	1230.811	34	0	16	33145.147	34	4826.423
17	1230.811	35	0	17	30781.617	35	4130.973
18	1140.752	36	0	18	30780.617	36	528.542

Tabelle 6.2 (Fortsetzung)

Korpus De Kock 2			
Rang	**Häufigkeit**	**Rang**	**Häufigkeit**
1	694.966	19	146.978
2	666.483	20	126.975
3	472.939	21	122.975
4	461.977	22	121.935
5	442.935	23	120.975
6	429.973	24	117.934
7	427.973	25	116.974
8	415.971	26	110.973
9	352.924	27	97.932
10	340.922	28	95.931
11	334.921	29	66.968
12	290.996	30	60.967
13	265.953	31	46.965
14	218.987	32	39.924
15	188.983	33	27.923
16	179.942	34	23.922
17	173.941	35	12.961
18	151.938	36	7.920

Tabelle 7: Empirische Verteilung der Satzlänge in den einzelnen Texten

	Text 23	**Text 24**	**Text 25**	**Text 28**	**Text 29**	**Text 30**	**Text 31**
x	f_x	f_x	f_x	f_x	f_x	f_x	f_x
1	85	209	18	17	59	66	14
2	100	139	14	11	55	41	13
3	111	67	12	14	50	19	26
4	71	38	5	7	19	8	16
5	48	17	3	3	10	3	5
6	23	13	1	2	6	1	8
7	13	3	1	4	2	-	9
8	6	4		2	1	-	1
9	1	-		-	2	1	9
10		1		1	-		1
11		-		2	1		1
12		-		-			-
13		-		1			-
14		3		-			-
15				1			-
16							-
17							1

Tabelle 7 (Fortsetzung)

	Text 32	**Text 33**	**Text 35**	**Text 37**	**Text 43**	**Text 45**
x	f_x	f_x	f_x	f_x	f_x	f_x
1	22	285	26	43	64	89
2	17	160	52	60	68	142
3	13	66	29	56	42	90
4	2	47	10	29	21	44
5	1	20	6	16	9	17
6	-	5	5	4	6	17
7	1	4	1	4	1	12
8		1	-	1	-	2
9			1		1	3
10					-	
11					-	
12					1	

Tabelle 8.1: Anpassung der negativen hypergeometrischen Verteilung an das Vorkommen von Adjektiven in 100-Wort-Textblöcken

	Text 27		**Text 28**		**Text 29**		**Text 30**	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	-	-	-	-	1	0.7044		
2	-	-	1	0.4962	1	1.7306	1	0.0801
3	2	1.6891	1	1.0051	3	2.8397	0	0.4136
4	2	3.2853	0	1.4759	3	3.8718	0	1.1289
5	5	4.5704	0	1.8872	3	4.7238	2	2.1324
6	9	5.488	4	2.2268	6	5.3365	5	3.0743

7	2	6.0401	3	2.487	7	5.6851	4	3.5271
8	6	6.256	3	2.663	8	5.7712	2	3.2598
9	9	6.18	1	2.7519	9	5.6163	1	2.3991
10	7	5.8636	4	2.7522	2	5.2565	1	1.3494
11	3	5.3617	3	2.6641	7	4.737	1	0.5264
12	4	4.7292	4	2.4887	2	4.1077	1	0.1089
13	3	4.0189	1	2.229	2	3.4198		
14	6	3.28	1	1.8897	2	2.7219		
15	2	2.5561	2	1.4784	1	2.0571		
16	2	1.8849	1	1.0072	3	1.461		
17	0	1.2961	1	0.4976	0	0.9596		
18	0	0.8115	-	-	1	0.5677		
19	0	0.4432	-	-	0	0.289		
20	1	0.1933	-	-	0	0.1153		
21	1	0.0527	-	-	1	0.028		

Tabelle 8.1 (Fortsetzung)

	Text 31		**Text 32**		**Text 33**		**Text 35**	
x	**f_x**	**NP_x**	**f_x**	**NP_x**	**f_x**	**NP_x**	**f_x**	**NP_x**
0	-	-	-	-	2	0.7512	1	0.8431
1	-	-	1	0.711	3	2.7491	3	3.0412
2	1	4.3991	0	0.96	3	5.7528	4	6.1397
3	10	6.2047	1	1.0765	10	8.9415	10	9.0423
4	8	6.7941	1	1.1192	11	11.3671	15	10.7192
5	5	6.5939	1	1.1123	11	12.373	14	10.7197
6	9	5.8506	2	1.0694	13	11.803	4	9.261
7	2	4.7476	2	0.9991	15	9.9716	4	6.9852
8	1	3.4441	2	0.9081	3	7.4725	4	4.6052
9	4	2.095	0	0.8014	5	4.9348	5	2.6345
10	1	0.8709	0	0.6836	3	2.8267	0	1.2857
11			0	0.5588	1	1.3638	1	0.5189
12			0	0.4311	0	0.5256	0	0.1639
13			0	0.3049	0	0.1455	0	0.0363
14			1	0.2647	1	0.022	1	0.0043

Tabelle 8.1 (Fortsetzung)

	Text 37		**Text 40**		**Text 46**	
x	**f_x**	**NP_x**	**f_x**	**NP_x**	**f_x**	**NP_x**
1	-	-	1	1.1719	-	-
2	-	-	5	3.7832	1	0.0952
3	1	0.6731	6	6.8179	0	0.4856
4	1	2.0109	9	8.9439	1	1.2859
5	1	3.6497	14	9.4284	2	2.3363
6	6	5.152	5	8.3763	2	3.2465
7	8	6.1767	4	6.4287	4	3.6462
8	7	6.5528	5	4.3141	3	3.414
9	7	6.2817	3	2.54	2	2.7119
10	5	5.4966	0	1.3073	5	1.8433

11	6	4.4032	0	0.5818	0	1.0738
12	0	3.2199	1	0.2195	0	0.5337
13	2	2.1302	0	0.0678	0	0.2236
14	3	1.2536	0	0.0162	0	0.0774
15	0	0.6374	0	0.0027	0	0.0214
16	0	0.2659	1	0.0002	0	0.0044
17	0	0.0821	-	-	0	0.0006
18	1	0.0143	-	-	1	0

Tabelle 8.2: Anpassung der negativen hypergeometrischen und negativen Binomialverteilung an das Vorkommen von Adjektiven in 100-Textblöcken (Gesamttexte)

		Negative hypergeometrische Verteilung	**Negetive Binomialverteilung**
x	**f_x**	**NP_x**	**NP_x**
0	3	4.4692	2.7599
1	9	13.3895	10.7026
2	17	24.6996	23.2092
3	44	35.9272	37.1068
4	51	45.0618	48.7552
5	58	50.8952	55.7266
6	65	53.0333	57.3446
7	55	51.7352	54.342
8	44	47.6866	48.1793
9	46	41.7753	40.4279
10	28	34.9059	32.388
11	22	27.8706	24.9404
12	12	21.2758	18.5595
13	8	15.5178	13.4044
14	15	10.7941	9.4294
15	5	7.1391	6.4797
16	7	4.4698	4.3604
17	1	2.633	2.8794
18	3	1.4471	1.8693
19	0	0.7336	1.1949
20	1	0.3376	0.753
21	2	0.2026	1.1878

Tabelle 9.1: Anpassung der Hyperpascal-Verteilung an die Komplexität von Teilsätzen

Hyperpascal-Verteilung (keine Zusammenfassung der Kompkexitätswerte zu Klassen)			**Hyperpascal-Verteilung (Komplexitätswerte zu Klassen zusammengefasst)**		
x	**f_x**	**NP_x**	**x**	**f_x**	**NP_x**
1	130	158.2503	1	872	871.8942
2	742	759.0823	2	602	602.1847
3	492	384.3196	3	30	29.5745
4	110	140.7046	4	1	1.2904

5	27	44.6496	5	0	0.0562
6	3	13.0614			
7	1	4.9322			

Tabelle 9.2. Anpassung der Dacey-Poisson-Verteilung an die Komplexität von Teilsätzen

Dacey-Poisson-Verteilung		
x	**f_x**	**NP_x**
1	130	128.8201
2	742	765.9123
3	492	444.8286
4	110	133.7226
5	27	27.0418
6	3	4.1164
7	1	0.5582

Tabelle 10.1. Empirische Ranghäufigkeitsverteilung von Teilsätzen und ihre Modellierung durch die Zipf-Mandelbrot-Verteilung (Auszug)

Rang	**f_x**	**NP_x**	**Rang**	**f_x**	**NP_x**	**Rang**	**f_x**	**NP_x**
1	52	20.3354	101	3	2.8644	201	2	1.8232
2	21	18.2893	102	3	2.8462	202	2	1.8172
3	20	16.6865	103	3	2.8283	203	2	1.8113
4	20	15.3913	104	3	2.8106	204	2	1.8054
5	18	14.3192	105	3	2.7932	205	2	1.7995
6	16	13.4146	106	3	2.776	206	2	1.7937
7	14	12.6392	107	3	2.7592	207	2	1.788
8	10	11.9659	108	3	2.7425	208	2	1.7823
9	10	11.3748	109	3	2.7261	209	2	1.7766
10	9	10.851	110	3	2.71	210	2	1.771
11	9	10.3831	111	3	2.6941	211	2	1.7654
12	9	9.962	112	2	2.6784	212	2	1.7599
13	8	9.5808	113	2	2.663	213	2	1.7544
14	7	9.2337	114	2	2.6477	214	2	1.7489
15	7	8.9161	115	2	2.6327	215	2	1.7435
16	7	8.6243	116	2	2.6179	216	2	1.7381
17	6	8.3549	117	2	2.6033	217	2	1.7328
18	6	8.1054	118	2	2.5889	218	2	1.7275

19	6	7.8736	119	2	2.5746	219	2	1.7223
20	6	7.6575	120	2	2.5606	220	2	1.7171
21	6	7.4555	121	2	2.5468	221	2	1.7119
22	6	7.2661	122	2	2.5331	222	2	1.7068
23	6	7.0883	123	2	2.5197	223	2	1.7017
24	6	6.9208	124	2	2.5064	224	2	1.6966
25	6	6.7627	125	2	2.4932	225	2	1.6916
26	6	6.6133	126	2	2.4803	226	2	1.6866
27	5	6.4718	127	2	2.4675	227	2	1.6817
28	5	6.3376	128	2	2.4549	228	2	1.6768
29	5	6.21	129	2	2.4424	229	2	1.6719
30	5	6.0886	130	2	2.4301	230	2	1.6671
31	5	5.9729	131	2	2.4179	231	2	1.6623
32	5	5.8625	132	2	2.4059	232	2	1.6575
33	5	5.7569	133	2	2.394	233	2	1.6527
34	5	5.656	134	2	2.3823	234	2	1.648
35	5	5.5593	135	2	2.3707	235	2	1.6434
36	5	5.4667	136	2	2.3593	236	2	1.6387
37	5	5.3777	137	2	2.348	237	2	1.6341
38	5	5.2922	138	2	2.3368	238	2	1.6296
39	5	5.21	139	2	2.3257	239	2	1.625
40	4	5.1309	140	2	2.3148	240	2	1.6205
41	4	5.0547	141	2	2.304	241	2	1.616
42	4	4.9813	142	2	2.2934	242	2	1.6116
43	4	4.9104	143	2	2.2828	243	2	1.6071
44	4	4.842	144	2	2.2724	244	2	1.6028
45	4	4.7758	145	2	2.2621	245	2	1.5984
46	4	4.7119	146	2	2.2519	246	2	1.5941
47	4	4.65	147	2	2.2418	247	2	1.5898
48	4	4.5902	148	2	2.2318	248	2	1.5855
49	4	4.5321	149	2	2.222	249	1	1.5812
50	4	4.4759	150	2	2.2122	250	1	1.577
51	4	4.4213	151	2	2.2025	251	1	1.5728
52	4	4.3684	152	2	2.193	252	1	1.5687

53	4	4.317	153	2	2.1836	253	1	1.5645
54	4	4.2671	154	2	2.1742	254	1	1.5604
55	4	4.2186	155	2	2.165	255	1	1.5563
56	4	4.1714	156	2	2.1558	256	1	1.5523
57	4	4,1255	157	2	2.1467	257	1	1.5483
58	3	4.0808	158	2	2.1378	258	1	1.5443
59	3	4.0373	159	2	2.1289	259	1	1.5403
60	3	3.9949	160	2	2.1201	260	1	1.5363
61	3	3.9537	161	2	2.1114	261	1	1.5324
62	3	3.9134	162	2	2.1028	262	1	1.5285
63	3	3.8742	163	2	2.0943	263	1	1.5246
64	3	3.8359	164	2	2.0859	264	1	1.5208
65	3	3.7985	165	2	2.0775	265	1	1.5169
66	3	3.762	166	2	2.0693	266	1	1.5131
67	3	3.7264	167	2	2.0611	267	1	1.5093
68	3	3.6916	168	2	2.053	268	1	1.5056
69	3	3.6576	169	2	2.0449	269	1	1.5018
70	3	3.6243	170	2	2.037	270	1	1.4981
71	3	3.5918	171	2	2.0291	271	1	1.4944
72	3	3.56	172	2	2.0213	272	1	1.4908
73	3	3.5289	173	2	2.0136	273	1	1.4871
74	3	3.4985	174	2	2.0059	274	1	1.4835
75	3	3.4687	175	2	1.9984	275	1	1.4799
76	3	3.4395	176	2	1.9908	276	1	1.4763
77	3	3.4109	177	2	1.9834	277	1	1.4727
78	3	3.3829	178	2	1.976	278	1	1.4692
79	3	3.3554	179	2	1.9687	279	1	1.4657
80	3	3.3285	180	2	1.9615	280	1	1.4622
81	3	3.3022	181	2	1.9543	281	1	1.4587
82	3	3.2763	182	2	1.9472	282	1	1.4553
83	3	3.2509	183	2	1.9401	283	1	1.4518
84	3	3.226	184	2	1.9332	284	1	1.4484
85	3	3.2016	185	2	1.9262	284	1	1.445
86	3	3.1776	186	2	1.9194	286	1	1.4416

87	3	3.1541	187	2	1.9126	287	1	1.4383
88	3	3.131	188	2	1.9058	288	1	1.4349
89	3	3.1083	189	2	1.8991	289	1	1.4316
90	3	3.086	190	2	1.8925	290	1	1.4283
91	3	3.0641	191	2	1.8859	291	1	1.425
92	3	3.0426	192	2	1.8794	292	1	1.4218
93	3	3.0215	193	2	1.873	293	1	1.4185
94	3	3.0007	194	2	1.8666	294	1	1.4153
95	3	2.9802	195	2	1.8602	295	1	1.4121
96	3	2.9601	196	2	1.8539	296	1	1.4089
97	3	2.9404	197	2	1.8477	297	1	1.4057
98	3	2.9209	198	2	1.8415	298	1	1.4026
99	3	2.9018	199	2	1.8353	299	1	1.3994
100	3	2,883	200	2	1,8293	300	1	1.3963

Tabelle 10.2.: Frequenzspektrum von Teilsätzen. Anpassung der Waring-Verteilung

x	**f_x**	**NP_x**	**x**	**f_x**	**NP_x**	**x**	**f_x**	**NP_x**
1	633	628.6638	19	0	0.2849	37	0	0.0377
2	137	135.0091	20	2	0.2442	38	0	0.0348
3	53	50.2111	21	1	0.2109	39	0	0.0321
4	18	23.9331	22	0	0.1833	40	0	0.0297
5	14	13.1979	23	0	0.1603	41	0	0.0275
6	10	8.0182	24	0	0.1409	42	0	0.0256
7	3	5.2212	25	0	0.1246	43	0	0.0238
8	1	3.582	26	0	0.1106	44	0	0.0222
9	3	2.5598	27	0	0.0986	45	0	0.0207
10	2	1.8901	28	0	0.0883	46	0	0.0193
11	0	1.4337	29	0	0.0794	47	0	0.0181
12	0	1.1122	30	0	0.0716	48	0	0.017
13	0	0.8795	31	0	0.0648	49	0	0.0159
14	1	0.7069	32	0	0.0588	50	0	0.015
15	0	0.5764	33	0	0.0535	51	0	0.0141
16	1	0.4759	34	0	0.0489	52	1	0.3345
17	0	0.3973	35	0	0.0447	-	-	-
18	1	0.3349	36	0	0,041	-	-	-

C. Empirische und theoretische Daten zu den funktionalen Zusammenhängen

Tabelle 11.1: Abhängigkeit der Polylexie von der Buchstabenlänge im Textkorpus
Anpassung der Funktion $y = ax^b$

Wortlänge	Polylexie, empirisch	Polylexie, theoretisch	Wortlänge	Polylexie, empirisch	Polylexie, theoretisch
2	6.91	7.69	9	3.14	3.25
3	5.80	6.10	10	2.95	3.06
4	5.78	5.17	11	2.53	2.90
5	5.58	4.55	12	2.07	2.76
6	5.15	4.10	13	2.03	2.63
7	4.85	3.76	14	2.31	2.52
8	3.64	3.48	15	1.26	2.43

Tabelle 11.2: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (alle Wortarten)
Anpassung der Funktion $y = ax^b$

Wortlänge	Empirische Polylexie	Theoretische Polylexie	Wortlänge	Empirische Polylexie	Theoreti-sche Poly-lexie
3	6.41	6.61	9	2.85	2.85
4	5.34	5.30	10	2.55	2.62
5	4.70	4.47	11	2.46	2.44
6	4.18	3.88	12	2.19	2.28
7	3.57	3.45	13	1.94	2.15
8	2.95	3.11			

Tabelle 11.3: Abhängigkeit der Polylexie von der Buchstabenlänge im Text (einzelne Wortarten). Anpassung der Funktion $y = ax^b$

	Substantive		Adjektive		Verben	
Wortlänge	Empirische Polylexie	Theoretische Polylexie	Empirische Polylexie	Theoretische Polylexie	Empirische Polylexie	Theoretische Polylexie
3	6.57	6.64	5.92	6.26		
4	5.21	5.34	5.64	4.85	5.1	5.49
5	5	4.52	3.70	3.98	4.91	4.74

6	3.94	3.94	3.41	3.38	4.93	4.21
7	3.5	3.51	2.96	2.95	3.94	3.80
8	2.9	3.17	2.47	2.62	3.22	3.49
9	3	2.90	1.94	2.36	3.18	3.23
10	2.64	2.68	2.47	2.15	2.48	3.01
11	2.25	2.49	2	1.97	3.04	2.83
12	2.44	2.33				
13	2.3	2.20				

Tabelle 11.4: Abhängigkeit der Polylexie von der Buchstabenlänge im Lexikon
Anpassung der Funktion $y = ax^b$

Wortlänge	Empirische Polylexie	Theoretische Polylexie	Wortlänge	Empirische Polylexie	Theoretische Polylexie
3	3.89	4.26	10	3.20	2.91
4	3.83	3.86	11	3.06	2.72
5	3.77	3.57	12	2.34	2.64
6	3.26	3.35	13	2.23	2.56
7	4.06	3.18	14	2.28	2.50
8	2.96	3.03	15	2.20	2.44
9	2.87	2.91			

Tabelle 11.5: Abhängigkeit der Polylexie von der Silbenlänge im Lexikon
Anpassung der Funktion $y = ax^b$

Wortlänge	Mittlere Polylexie	
	Empirisch	Theoretisch
1	4.08	4.1748
2	3.41	3.3618
3	3.2	2.9618
4	2.92	2.7071
5	2.11	2.5248

Tabelle 12.1: Abhängigkeit der Buchstabenlänge von der Frequenz im Textkorpus
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Länge	Theoretische Länge	Frequenz	Empirische Länge	Theoretische Länge
1	8.26	8.66	9	7.32	6.39
2	7.83	7.87	10	6.15	6.30
3	7.68	7.44	11	5.75	6.2
4	7.66	7.15	12	7.31	6.14
5	6.91	6.93	13	5.71	6.07
6	7.06	6.76	14	5.55	6.01
7	6.62	6.62	16	5.18	5.90
8	6.02	6.50			

Tabelle 12.2: Abhängigkeit der Buchstabenlänge von der Frequenz im Text (alle Wortarten)
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Länge	Theoretische Länge	Frequenz	Empirische Länge	Theoretische Länge
1	7.67	7.50	7	6.73	6.42
2	7.14	7.10	8	6.5	6.35
3	6.61	6.87	9	6.90	6.29
4	6.85	6.71	10	5.93	6.23
5	6.08	6.59	11	6.14	6.19
6	6.25	6.50			

Tabelle 12.3: Abhängigkeit der Buchstabenlänge von der Frequenz im Lexikon
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Länge	Theoretische Länge	Frequenz	Empirische Länge	Theoretische Länge
4	9.07	9.43	13	9.27	8.97
5	10.87	9.34	14	9.35	8.94
6	9.94	9.27	15	8.25	8.92
7	8.15	9.21	16	8.5	8.89

8	8	9.16	18	9.07	8.85
9	9.28	9.11	19	9.69	8.83
10	9.07	9.07	20	9.44	8.81
11	8.62	9.04	21	9.07	8.79
12	8.43	9	38	8.23	8.57

Tabelle 12.4: Abhängigkeit der Silbenlänge von der Frequenz im Lexikon
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Länge	Theoretische Länge	Frequenz	Empirische Länge	Theoretische Länge
4	3.32	3.54	14	3.41	3.19
5	4	3.47	15	2.81	3.17
6	3.56	3.42	16	3.16	3.15
7	2.96	3.38	17	2.88	3.14
8	3.27	3.34	18	3.35	3.12
9	3.5	3.31	19	3.15	3.11
10	3.17	3.28	20	3.38	3.10
11	3.25	3.25	21	3.30	3.08
12	3.13	3.23	25	2.88	3.04
13	3.16	3.21	38	2.92	2.93

Tabelle 13.1: Abhängigkeit der Polylexie von der Frequenz im Textkorpus
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Polylexie	Theoretische Polylexie	Frequenz	Empirische Polylexie	Theoretische Polylexie
1	3.39	3.64	8	6.02	5.56
2	3.96	4.19	9	4.78	5.70
3	4.43	4.55	10	6.60	5.82
4	5.36	4.83	11	5.16	5.93
5	4.99	5.05	12	5.37	6.04
6	5.06	5.24	13	6.50	6.14
7	6.52	5.41			

Tabelle 13.2: Abhängigkeit der Polylexie von der Frequenz im Text(alle Wortarten)
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Polylexie	Theoretische Polylexie	Frequenz	Empirische Polylexie	Theoretische Polylexie
1	3.09	2.92	7	4.84	5.31
2	3.47	3.62	8	4.89	5.53
3	4.05	4.10	9	6.90	5.74
4	5.39	4.47	10	5.93	5.93
5	4.38	4.79	11	6.14	6.10
6	4.57	5.07			

Tabelle 13.3: Abhängigkeit der Polylexie von der Frequenz im Lexikon
Anpassung der Funktion $y = ax^b$

Frequenz	Empirische Polylexie	Theoretische Polylexie	Frequenz	Empirische Polylexie	Theoretische Polylexie
4	2.69	2.49	13	3	3.18
5	1.87	2.60	14	3	3.17
6	2.61	2.69	15	3.43	3.21
7	3.26	2.77	16	2.33	3.25
8	3.35	2.84	18	3.78	3.33
9	2.57	2.91	19	3.38	3.36
10	3.28	2.97	20	2.72	3.39
11	2.70	3.02	21	3.53	3.43
12	4	3.08	38	4.07	3.84

Tabelle 14.1: Teilsatzlänge als Funktion der Satzlänge. Messwerte in den Gesamttexten

Satzlänge	Empirische Teilsatz-länge	Theoretisch berechnete Teilsatzlänge		
		$y = ax^b e^{-cx}$	$y = ax^b$	$y = a\,e^{-cx}$
1	11.3781	11.4071	11.0814	10.5133
2	9.6659	9.6667	9.9323	10.0374
3	9.0868	8.9810	9.3161	9.5830
4	8.8433	8.6640	8.9023	9.1493
5	8.4153	8.5324	8.5940	8.7351

6	7.9925	8.5132	8.3501	8.3397
7	8.9523	8.5709	8.1492	7.9622

Tabelle 14.2: Teilsatzlänge als Funktion der Satzlänge. Messwerte in den einzelnen Texten

Texte	**Satzlänge**	**Empirische Teil-satzlänge**	**Theoretisch berechnete Teilsatzlänge**		
			$y = ax^b e^{-cx}$	$y = ax^b$	$y = a\,e^{-cx}$
Text 24	1	8.7942	8.9287	8.7961	8.6439
	2	8.6690	8.3505	8.4729	8.5012
	3	8.2885	8.1456	.2894	8.3607
	4	7.7960	8.0838	8.1617	8.2226
	5	7.6705	8.0988	8.0639	8.0868
	6	8.5512	8.1623	7.9849	7.9532
Text 35	1	8.8846	8.8836	8.9145	8.7333
	2	7.4134	7.4186	7.3517	7.6373
	3	6.5977	6.5898	6.5678	6.6788
	4	6	6.0035	6.0629	5.8405
Text 37	1	18.4418	18.4749	18.207	17.2817
	2	13.4333	13.1856	13.6784	14.5414
	3	10.75	11.2565	11.5713	12.2357
	4	10.7327	10.3408	10.2762	10.2955
	5	9.7875	9.8890	9.3723	8.663
Text 43	1	13.9687	13.9513	13.684	13.2647
	2	10.6323	10.7325	11.3008	11.705
	3	10.0476	9.9037	10.104	10.3287
	4	9.7857	9.8457	9.3325	9.1141
Text 45	1	12.8988	12.7623	12.3531	11.4427
	2	9.4119	9.9254	10.3072	10.5613
	3	9	8.8350	9.2713	9.7479
	4	8.6988	8.3113	8.6001	8.9971
	5	8.1882	8.0590	8.1131	8.3041
	6	7.6470	7.9656	7.7358	7.6645
	7	8	7.9780	7.4305	7.0742

Tabelle 15: Abhängigkeit der Frequenz von Teilsätzen von der Komplexität
Anpassung der Funktion $y = ax^b e^{cx}$

Komplexität	Mittlere Frequenz	
	empirisch	**theoretisch**
1	1.71	1.72
2	1.89	1.86
3	1.66	1.64
4	1.23	1.32
5	1.08	1.01

Literaturverzeichnis

AICHELE, D. (2005): Quantitative Linguistik in Deutschland und Österreich. In: KÖHLER, R. et al. (Hrsg.)(2005), S. 16-23.

ALTMANN, G. (1972): Status und Ziele der quantitativen Sprachwissenschaft. In: JÄGER, S. (Hrsg.): Linguistik und Statistik. Braunschweig: Vieweg, S. 1-10.

ALTMANN, G. (1978): Towards a theory of language. In: ALTMANN, G. (Hrsg.): Glottometrika 1, S. 1-25.

ALTMANN, G. (1980): Prolegomena to Menzerath's Law. In: GROTJAHN, R. (Hrsg.): Glottometrika 2, S. 1-10.

ALTMANN, G. (1981): Zur Funktionalanalyse in der Linguistik. In: ESSER, J. & HÜBLER, A. (Hrsg.): Forms and Functions. Tübingen: Narr, S. 25-32.

ALTMANN, G. (1985a): Semantische Diversifikation. In: *Folia Linguistica* 19, S. 177-200.

ALTMANN, G. (1985b): Sprachtheorie und mathematische Modelle. Christian-Albrechts-Universität Kiel, SAIS Arbeitsberichte. Heft 8, S. 1-13.

ALTMANN, G. (1988a). Wiederholungen in Texten. Bochum: Brockmeyer.

ALTMANN, G. (1988b): Verteilung der Satzlängen. In: SCHULZ, K.-P. (Hrsg.): Glottometrika 9, S. 147-169.

ALTMANN, G. (1991): Modelling diversifikation phenomena in language. In: ROTHE, U. (Hrsg.)(1991), S. 33-46.

ALTMANN, G. (1992): Das Problem der Datenhomogenität. In: RIEGER, B. (Hrsg.): Glottometrika 13, S. 287-298.

ALTMANN, G. (1996): Diversification processes of the word. In: SCHMIDT, P. (Hrsg.): Glottometrika 15, S. 102-111.

ALTMANN, G. (2002): Morphologie. In: ALTMANN, G. et al. (Hrsg.): Einführung in die quantitative Lexikologie. Göttingen: Peust & Gutschmidt, S. 56-62.

ALTMANN, G. & LEHFELDT, W. (1973): Allgemeine Sprachtypologie. München: Fink.

ALTMANN, G. & LEHFELDT, W. (1980): Einführung in die quantitative Phonologie. Bochum: Brockmeyer.

ALTMANN, G., BEÖTHY, E. & BEST, K.-H. (1982): Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationswissenschaft 35.5, S. 537-543.

ALTMANN, G. & BURDINSKI, V. (1982): Towards a Law of Word Repetition in Text Blocks. In: LEHFELDT, W. & STRAUSS, U. (Hrsg.): Glottometrika 4, S. 147-167.

ALTMANN, G., KIND, B. (1983): Ein semantisches Gesetz. In: KÖHLER, R. & BOY, J. (Hrsg.): Glottometrika 5, S. 1-13

ALTMANN, G. & BEST, K.-H. & KIND, B. (1987): Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In: FICKERMANN, I. (Hrsg.): Glottometrika 8, S. 130-139.

ALTMANN, G. & HAMMERL, R. (1989): Diskrete Wahrscheinlichkeitsverteilungen. Bochum: Brockmeyer.

ALTMANN, G. & SCHWIBBE, M.H. (1989): Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim, Zürich, New York: Olms.

ALTMANN, G. & KÖHLER, R. (1996): "Language Forces" and Synergetic Modelling of Language Phenomena. In: SCHMIDT, P. (Hrsg.): Glottometrika 15. Trier: WVT wissenschaftlicher Verlag, S. 62-76.

ANTIC, G. &KELIK, E. &GRZYBEK, P. (2006): Zero-syllable Words in Determining Word Length. In: GRZYBEK, P. (Hrsg.): Contributions to the Science of Language. Word Length Studies and Related Issues Dordrecht, NL: Springer, S. 117–156

ARAPOV, M. V. & CHERC, M. M. (1983): Mathematische Methoden in der historischen Linguistik. [Übers. aus dem Russischen] Bochum: Brockmeyer

BAAYEN, R. H. (2001). Word frequency distributions. Dordrecht u.a.: Kluwer.

BÉCHADE, H. (1992): Phonétique et morphologie du français moderne et contemporain, PUF.

BEÖTHY, E. & ALTMANN, G. (1984): Semantic Diversification of Hungarian verbal prefixes. III. „föl-", „el-", „be-". In: ROTHE, U. (Hrsg.): Glottometrika 7, S. 45-56.

BEÖTHY, E. & ALTMANN, G (1991): The diversification of meaning of Hungarian verbal prefixes. In: ROTHE, U. (Hrsg.)(1991), S. 60-66.

BERGENHOLTZ, H. & SCHAEDER, B. (1977): Die Wortarten des Deutschen: Versuch einer syntaktisch orientierten Klassifikation. Stuttgart: Klett.

BERNET, CH. (1988): Faits lexicaux. Richesse du vocabulaire. In: THOIRON, P. et al. (Hrsg.): Études sur la richesse et la structure lexicale.Paris u.a.: Champion u.a., S. 1-11.

BEST, K.H. (1990): Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhoch-deutschen. In: HŘEBÍČEK, L. (Hrsg.): Glottometrika 11, S. 107-110.

BEST, K.H. (1991): Zur Diversifikation einer Partikel des Deutschen. In: ROTHE, U. (Hrsg.)(1991), S. 94-104.

BEST, K.-H. (1996): Zur Bedeutung von Wortlänge, am Beispiel althochdeutscher Texte. In: Papiere zur Linguistik 55, S. 141-152.

BEST, K.-H. (1997a): Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: BEST, K.-H. (Hrsg.): Glottometrika 16, S. 276-285.

BEST, K.-H. (19997b): Warum nur: Wortlänge? Nicht nur ein Vorwort. In: BEST, K.-H. (Hrsg.): Glottometrika 16, S. V-XII.

BEST, K.-H. (1997c): Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In: BEST, K.-H. (Hrsg.): Glottometrika 16, S. 1-15.

BEST, K.-H. (1998): Zur Interaktion der Wortarten in Texten. In: Papiere zur Linguistik 58, S. 83-95.

BEST, K.-H. (1999): Quantitative Linguistik: Entwicklung, Stand und Perspektive. In: Göttinger Beiträge zur Sprachwissenschaft, Heft 2. Göttingen: Peust & Gutschmidt, S. 7-23.

BEST, K.-H. (2000): Verteilungen der Wortarten in Anzeigen. In: Göttinger Beiträge zur Sprachwissenschaft 4, S. 37-51.

BEST, K.H (Hrsg.)(2001): Häufigkeitsverteilungen in Texten. Göttingen: Preust & Gutschmidt, S.167-201

BEST, K.H. (2001a): Quantitative Linguistik: Eine Annäherung. 2., überarbeitete und erweiterte Auflage. Göttingen.

BEST, K.-H. (2001b): „Wie viele Wörter enthalten Sätze im Deutschen? Ein Beitrag zu den Scheman-Altmann-Gesetzen“. In: BEST, K.-H. (Hrsg.)(2001), S.167-201.

BEST, K.-H. (2001c): Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. In: ALTMANN, G. (Hrsg.): Glottometrics 1, S. 1-26.

BEST, K.-H. (2002): Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. In: Göttinger Beiträge zur Sprachwissenschaft 7, S. 7-32.

BEST, K.H. (2005a): Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. In: Glottometrics 11, S. 9-31.

BEST, K.H. (2005b): Sprachliche Einheiten in Textblöcken. In: Glottometrics 9, S. 1-12.

BEST, K.H. (2005c): Satzlänge. In: KÖHLER, R. et al. (Hrsg.) (2005), S. 298-304.

BILLMEIER, G. (1969): Worthäufigkeitsverteilungen vom Zipfschen Typ, überprüft an deutschen Textmaterial. Hamburg: Buske.

BRAINERD, B. (1972): Article use as indicator of style among English-language authors. In: JÄGER, S. (Hrsg.): Linguistik und Statistik. Braunschweig: Vieweg, S. 11-32.

BRUNET, E. (1978): Le vocabulaire de Giraudoux. Structure et évolution. Genève: Slatkine.

BRUNET, E. (1988): La structure lexicale dans l'œuvre de Hugo. In: THOIRON, P. et al. (Hrsg.): Études sur la richesse et la structure lexicale. Paris u.a.: Champion u.a., S. 23-42

BUNGE, M. (1967): Scientific Research I. Berlin u.a.: Springer.

CHITASHVILI, R.J. & BAAYEN, R.H. (1993): Word frequency distributions of texts and corpora as large number of rare event distributions. In: HŘEBÍČEK, L. & ALTMANN, G. (Hrsg.): Quantitative text analysis. Trier: WVT, S. 54-135.

COCHRAN, G. (1932): The preparation of French reading material for beginners. In: French Review (Baltimore) 6, S. 458-471.

CRAMER, I. (2005): Das Menzerathsche Gesetz. In: KÖHLER, R. et al. (Hrsg.)(2005), S. 659-688.

DIECKMANN, S., & JUDT, B. (1996): Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: SCHMIDT, P. (Hrsg.): Glottometrika 15, S. 158-165.

DOTTRENS, R. & MASSARENTI, D. (1948): Vocabulaire fondamental du français. Contribution à un enseignement rationnel de l'orthographe d'usage. In: Cahiers de pédagogie expérimentale et de psychologie de l'enfant.

DUGAST, D. (1978): Sur quoi se fonde la notion d'étendue théorique du vocabulaire. In: Le français moderne, S. 25-32.

DUGAST, D. (1979a): Vocabulaire et discours. Fragments de lexicologie quantitative. Essai de lexicométrie organisationnelle. Genève: Slatkine.

DUGAST, D. (1979b): Vocabulaire et stylistique. Théâtre et dialogue. Genève: Slatkine.

DUNN-LARDEAU, B. (1986): Maîtrise du français écrit et applications pédagogiques de l'ordinateur. In: MULLER, CH. (Hrsg.): Méthodes quantitatives et informatiques dans l’étude des textes. Colloque international de CNRS à l'université de Nice, 5-8 Juin 1985. Paris u.a.: Champion u.a., S. 353-360.

EATON H. (1961): An English, French, German, Spanish Word Frequency Dictionary. A correlation of the first 6000 words in four single-language frequency list. 2. Aufl. New York: Dover Publications.

FELDT, S., JANSSEN, M. & KULEISA, S. (1997): Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: BEST, K.-H. (Hrsg.): Glottometrika 16, S. 145-151.

FIALA, P. (1986): Inventaires distributionnels et opérateurs textuels dans le "Rivage des Syrtes" de Julien Gracq. In: MULLER, CH. (Hrsg.): Méthodes quantitatives et informatiques dans l'étude des textes. Paris u.a.: Champion u.a., S. 381-390.

FICKERMANN, I. , MARKNER-JÄGER, B. & ROTHE, U. (1984): Wortlänge und Bedeutungskomplexität. In: BOY, J. & KÖHLER, R. (Hrsg.): Glottometrika 6, S. 115-126.

FLÄMIG, W. (1981): Wortklassen und Wortstrukturen. Grundzüge einer deutschen Grammatik. Von einem Autorenkollektiv unter Leitung von Karl Erich Heidolph, Walter Flämig, Wolfgang Motsch. Berlin: Akademie-Verlag, 458-636; 682-701.

FÖRSTEMANN, E. (1846): Über die numerische Lautverhältnisse im Deutschen. Germania. In: Neues Jahrbuch der Berlinischen Gesellschaft für deutsche Sprache und Alterthumskunde, Bd.7, S. 83-90.

FÖRSTEMANN, E. (1852): Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen. In: Zeitschrift für vergleichende Sprachforschung 1, S. 163-179.

GERLACH, R. (1982), Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie. In: LEHFELDT, W. & STRAUSS, U. (Hrsg.): Glottometrika 4, 95-102.

GIESEKING, K. (2002). Untersuchung zur Synergetik der englischen Lexik. In: KÖHLER, R. (Hrsg.): Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.(URL:http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/01_koehler_hg.pdf), S. 387-433.

GIRAUD, P. (1955): A Propos des caractères statistiques du vocabulaire et de l'équation de Zipf. In: Bulletin de la Société de linguistique de Paris (51), S. 236-239.

GUIRAUD, P. (1960): Problèmes et méthodes de la statistique linguistique. Paris: PUF.

GUIRAUD, P. (1963): Structures des répartitions et répartitions fréquentielles des éléments de la statistique du vocabulaire écrit. In: MOLES, A. & VALLANCIEN, B. (Hrsg.): Communications et Langage. Paris: Gauthier-Villars, S. 37-48.

GOUGENHEIM, G. (1959): La statistique du vocabulaire et son application à l'enseignement des langues. In: Revue de l'enseignement supérieure, S. 137-144.

GRAMONT, M. (1937): Le vers français, ses moyens d'expression, son harmonie. 3. Aufl. Paris u.a.: Champion u.a.

GREENBERG, J.H. (1960): A quantitative Approch to the morphological typology of languages. In: Journal of American Linguistics 26, S. 178-194.

GROTJAHN, R. (1992): Evaluating the adequacy of regression models: Some potential pitfalls. In: RIEGER, B. (Hrsg.), Glottometrika 13, S. 121-172.

GROTJAHN, R. & ALTMANN, G. (1993): Modelling the distribution of Word Length: Some Methodological Problems. In: KÖHLER, R. & BURGHARD, R. (Hrsg.): Contributions to quantitative linguistics. Dordrecht u.a.: Kluwer, S. 141-153.

GRZYBEK, P. (2007a): What a Difference an ‚E' Makes: Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen. In: DEUTSCHMANN, P. unter Mitarbeit von Peter Grzybek, Ludwig Karničar, Heinrich Pfandl (Hrsg.): Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag. Wien: Praesens, S. 105-128.

GRZYBEK, P. (2007b): On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. In: Glottometrics 15, S. 82-91.

GRZYBEK, P. & ALTMANN, G. (2002): Oscillation in the frequency-length relation. In: Glottometrics 5, S. 97-107.

GRZYBEK, P. & KELIH, E. (2003): „Graphemhäufigkeiten am Beispiel des Russischen. Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen“. In: Anzeiger für Slawische Philologie (31), S. 131-162.

GRZYBEK, P. & KELIH, E (2004): Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). In: Glottometrics 8, S. 23-41.

GRZYBEK, P.; KELIH, E. (2005a): „Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph.“ In: ALTMANN, G., LEVICKIJ, V. & PEREBEJNIS, V. (Hrsg.): Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics. Cernovici, S. 159-179.

GRZYBEK, P.; KELIH, E. (2005b): „Towards a General Model of Grapheme Frequencies in Slavic Languages.“ In: GARABÍK, R. (Hrsg.): Computer Treatment of Slavic and East European Languages. Bratislava, S. 73-87.

GRZYBEK, P. & KELIH, E. & ALTMANN, G. (2004): „Graphemhäufigkeiten am Beispiel des Russischen. Teil II: Modelle der Häufigkeitsverteilung“. In: Anzeiger für Sklavische Philologie (32), S. 25-54.

GRZYBEK, P., KELIH, E. & ALTMANN, G. (2005): „Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das ‘ë_’. In: Anzeiger für Slawische Philologie (33), S. 117-140.

GRZYBEK, P., KELIH, E. & ALTMANN, G. (2006): „Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen.“ In: KOZMOVÁ, R. (Hrsg.): Sprache und Sprachen im mitteleuropäischen Raum. Trnava, S. 661-664.

GUITER, H. & ARAPOV, M.V. (Hrsg.) (1982): Studies on Zipf's Laws. Bochum: Brockmeyer.

HAMMERL, R. (1988): Neue Perspektive der sprachlichen Synergetik: Begriffsstrukturen – Kognitive Gesetze. In: HAMMERL, R. (Hrsg.): Glottometrika 10, S. 129-140.

HAMMERL, R. (1990a): Untersuchung zur Verteilung der Wortarten im Text. In: HŘEBÍČEK, L. (Hrsg.): Glottometrika 11, S. 142-156.

HAMMERL, R. (1990b). Länge-Frequenz, Länge-Rangnummer: Überprüfung von zwei lexikalischen Modellen. In: HMMERL, R. (Hrsg.), Glottometrika 12, S. 1-24.

HAMMERL, R. (1991). Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells. Trier: WVT.

HAYGOOD, J.D. (1937): Le vocabulaire fondamental du Français, étude pratique sur l'enseignement des langues vivantes. Paris.

HEINICKE, N. (2008): Wortlängenverteilungen in französischen Briefen eines Autors. In: Glottometrics 16, S. 38-45.

HEMPEL, C.G. (1965): The logic of functional analysis. In: HEMPEL, C.G. (Hrsg.): Aspects of scientific explanation. New York: The Free Press, S. 297-330.

HENNERN, A. (1991): Zur semantischen Diversifikation von "in" im Englischen. In: ROTHE, U. (Hrsg.)(1991), S. 116-126.

HEUPS, G. (1983): Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge. Am Beispiel deutscher Texte verschiedener Textklassen. In: KÖHLER, R.& BOY, J. (Hrsg.): Glottometrika 5, S. 113-133.

HOFFMANN, L. (1976): Fachwortschatz Bauwesen. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig: Enzyklopädie.

IMBS, P. (1971): Dictionnaires des fréquences. Vocabulaire littéraire des XIXe et XXe siècles. Prais: Klincksieck.

JING, Z. (2001): Satzlängenhäufigkeiten in chinesischen Texten. In: BEST, K.H. (Hrsg.)(2001), S. 202-210.

KAEDING, F.W.(1897/98): Häufigkeitswörterbuch der deutschen Sprache. Berlin: Selbstverlag.

KELIH, E. (2007): Grapheme und Laute des Russischen: Zwei Ebenen - ein Häufigkeitsmodell? Re-Analyse einer Untersuchung von A.M. Peškovskij. In: GRZYBEK, P. & KÖHLER, R. (Hrsg.): Exacts methods in the study of language and text. Berlin: de Gruyter, S. 269-280.

KNÜPPEL, A. (2001): Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten. In: BEST, K-H.(Hrsg.)(2001), S. 248-280.

KÖHLER, R. (1982): Das Menzerathsche Gesetz auf Satzebene. In: LEHFELDT, W. & STRAUSS, U. (Hrsg.): Glottometrika 4, S. 103-113.

KÖHLER, R. (1984): Zur Interpretation des Menzerathschen Gesetzes. In: BOY, J. & KÖHLER, R. (Hrsg.): Glottometrika 6, S. 177-183.

KÖHLER, R. (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum: Brockmeyer.

KÖHLER, R. (1987): Selbstregulation der Lexik. In: BLUHME, H. (Hrsg.): Beiträge zur quantitativen Linguistik. Gedächtniskolloquium für Eberhard Zwirner, Antwerpen, 9.-12. April 1986. Tübingen: Narr, S. 156-166.

KÖHLER, R. (1990)**:** Linguistische Analyseebenen. Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. In: HREBÍCEK, L. (Hg.): Glottometrika11, S. 1-18.

KÖHLER, R. (1990a): Elemente der synergetischen Linguistik. In: HAMMERL, R. (Hrsg.): Glottometrika 12, S. 179-187)

KÖHLER, R. (1990b): Synergetik und sprachliche Dynamik. In: KOCH, W.A. (Hrsg.): Natürlichkeit der Sprache und der Kultur. Bochum: Brockmeyer, S. 96-112.

KÖHLER, R. (1991): Diversification of Coding Methods in Grammar. In: ROTHE, U. (Hrsg.) (1991), S. 47-55.

KÖHLER, R. (1995): Bibliography of Quantitative Linguistics. With the Assistance of Christiane Hoffmann. Amsterdam: Benjamins.

KÖHLER, R. (1999): Syntactic structures. Properties and interrelations. In: Journal of Quantitative Linguistics 6, S. 46-57.

KÖHLER, R. (2001): The distribution of some syntactic construction types in text blocks. In: UHLÍŘOVÁ, L. et al. (Eds.): Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček. Trier: WVT, S. 136-148.

KÖHLER, R. (2005a): Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: KÖHLER, R. et al. (Hrsg.)(2005), S. 1-16.

KÖHLER, R. (2005b): Quantitative Untersuchungen zur Valenz deutscher Verben. In: Glottometrics 9, S. 13-20.

KÖHLER, R. & ALTMANN, G. (2000): Probality distributions of syntactic units and properties. In: Journal of Quantitative Linguistics 7&3, S. 189-200.

KÖHLER, R. et al. (Hrsg.)(2005): Quantitative Linguistik: ein internationales Handbuch. Berlin: de Gruyter.

KROTT, A. (1999). The influence of morpheme polysemy on morpheme frequency in Proceedings of the Third International Conference on Quantitative Linguistics, August 26-29, 1997, Helsinki, Finnland. In: Journal of Quantitative Linguistics 6, 1, S. 58-65.

KROTT, A. (2002): Ein funktionalanalytisches Modell der Wortbildung. In**:** KÖHLER, R. (Hrsg.): Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik (URL: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/01_koehler_hg.pdf), S. 75-126.

KRYLOV, J. K. (1982): Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: GUITER, H. & ARAPOV, M.V. (Hrsg.): Studies on Zipf's Law. Bochum: Brockmeyer, S. 234-262.

KUCHNER, K.G. (1932): A study of the verb vocabulary of basis first year French grammars and second year French reading texts of secondary schools.

LEOPOLD, E. (1988): Stochastische Modellierung lexikalischer Evolutionsprozesse. Hamburg: Kovač.

MANDELBROT, B. (1953): An informational theory of statistical structure of language. In: JACKSON, W. (Hrsg.): Communication Theory. London, S. 486-502.

MENARD, N. (1983): Mesure de la richesse lexicale. Théorie et vérifications expérimentales. Etudes stylométriques et sociolinguistiques. Genève u.a.: Slatkine.

MENZERATH, P. (1954): Architektonik des deutschen Wortschatzes. Bonn.

MICHÉA, R. (1949): Le vocabulaire de base aux examens. In: Les Langues Modernes (43), S. 135-140.

MILLER, G.A., NEWMAN, E.B. & FRIEDMAN, E.A. (1958): "Length-frequency statistics for written English." In: Information and Control I, S. 370-389.

MONSONÉGO, S. (1966): Étude stylo-statistique du vocabulaire des vers et de la prose de la chantefable "Aucassin et Nicolette". Paris: Klincksieck.

MULLER, CH. (1965): Du nouveau sur les distributions lexicales: la formule de waring-Herdan. In : Cahiers de Lexikologie 1, Nr. 6, S. 35-53.

MULLER, CH. (1968): Initiation à la statistique linguistique. Paris : Librairie Larousse.

MULLER, CH. (1973): “Éléments de statistique linguistique”. In: Linguistica, Matematica e Calcolatori. Atti del convegno e della prima scuola internazionale Pisa 1970. Firenze :Zampolli, S. 349-378.

MULLER, CH. (1974): Initiation aux méthodes de la statistique linguistique. 2. Aufl. Paris: Hachette.

MULLER, CH. (1977a): Observation, prévision et modèles statistiques. In : DAVID, J. & MARTIN, R. (Hrsg.) : Études de statistique linguistique. Paris : Klincksieck, S. 9-19.

MULLER, CH. (1977b): Principes et méthodes de statistique lexicale. Paris: Hachette.

MULLER, CH. (1979): Langue française et linguistique quantitative. Genève : Slatkine repr.

MULLER, CH. (1985): Langue française, linguistique quantitative, informatique. Genève, Paris: Slatkine u.a.

NASVYTIS, A. (1953): Die Gesetzmäßigkeiten kombinatorischer Technik. Berlin u.a.: Springer.

NEMCOVÁ, E. (1991): Semantic diversification of Slovak verbal prefixes. In: ROTHE (Hrsg.)(1991), S. 67-74.

NIEHAUS, B. (1997): Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: BEST, K.-H. (Hrsg.): Glottometrika 16, S. 213-275.

ORLOV, Ju. K. (1982a): Dynamik der Häufigkeitsstrukturen. In: ORLOV, Ju. K. et al. (Hrsg.): Sprache, Text, Kunst. Bochum: Brockmeyer, S. 82-117.

ORLOV, Ju. K. (1982b): Ein Modell der Häufigkeitsstrukturen des Vokabulars. In: ORLOV, Ju. K. et al. (Hrsg.): Sprache, Text, Kunst. Bochum: Brockmeyer., S. 118-192.

PIEPER, U. (1979) : Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse. Tübingen: Narr.

PIOTROWSKI, R.G. et al. (1985) : Mathematische Linguistik. Bochum: Brockmeyer.

PRÜM, C. (1999): G.K. Zipf's Conception of Language as an Early Prototype of Synergetic Linguistics. In: Journal of Quantitative Linguistics 6, S. 78-84.

ROTHE, U. (1983) : Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzearthschen Gesetz an drei romanischen Sprachen. In: KÖHLER, R. & BOY, J. (Hrsg.): Glottometrika 5, S. 101-133.

ROTHE, U. (1985): Die Semantik des textuellen *et*. Bochum: Dissertation.

ROTHE**,** U**.** (1988). Polylexy and Compounding. In: SCHULZ, K.-P. (Hrsg.): Glottometrika 9, S. 121-134.

ROTHE, U. (1989): Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina. In: HŘEBÍČEK, L. (Hrsg.): Glottometrika 11, S. 111-121.

ROTHE, U. (Hrsg.) (1991): Diversification processes in language: grammar. Hagen.

ROTHE, U. (1991): Diversification processes in grammar. An introduction. In: ROTHE, U. (Hrsg.)(1991), S. 3-32.

ROUKK, M. (2001a): Satzlängen im Russischen. In. BEST, K-H. (Hrsg.)(2001), S. 211-218.

ROUKK, M. (2001b): Satzlängen in Texten von A. Tschechow. In: Göttinger Beiträge zur Sprachwissenschaft, Heft 5, S. 113-120.

FLESCH, R. (1948): *A New Readability Yardstick.* In: *Journal of Applied Psychology* 32(3), S. 221-233.

SAMBOR, J. (1984). Menzerath's Law and the Polysemy of Words. In: BOY, J. & KÖHLER, R. (Hrsg.): Glottometrika 6, S. 94-114.

SCHIERHOLZ, S. J. (1991). Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive. Tübingen: Niemeyer.

SCHLEICHER, A. (1852): Die Formlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt. Bonn u.a.: König.

SCHWEERS, A. & ZHU, J. (1991): Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: ROTHE, U. (Hrsg.) (1991), S. 157-165.

SCHWIBBE, M.H. (1984): Text- und Wortstatistische Untersuchungen zur Validität der Menzerathschen Regel. In: BOY, J. & KÖHLER, R. (Hrsg.): Glottometrika 6, S. 152-176.

SHERMAN, L. A. (1888): „Some observations upon the sentence-length in English prose. “ *University of Nebraska Studies*, 1; 119-130.

SICHEL, H. S. (1974): „On a distribution representing sentence-length in prose.“ In: Journal of the Royal Statistical Society, A-137, S. 25-34.

STEINER, P**.** (1995): Effects of Polylexy on Compounding. In: Journal of Quantitative Linguistics, 2.2, S. 133-140.

STEINER, P. (2002): Polylexie und Kompositionsaktivität in Text und Lexik. In**:** KÖHLER, R. (Hrsg.): Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik (URL: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/01_koehler_hg.pdf), S. 209-251.

STORRER, A. (1992): Verbvalenz. Theoretische und methodische Grundlagen ihrer Beschreibung in Grammatikographie und Lexikographie. Tübingen: Niemeyer.

STRAUSS, U. et al.: Word Length and Word frequency. In GRZYBEK, P. (Hrsg.): Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Dordrecht: Springer, 277-294.

TESNIÈRE, L. (1980): Grundzüge der strukturalen Syntax. Hrsg. u. übers. v. Ulrich Engel. Stuttgart: Klett-Cotta. [Die Originalausgabe erschien 1959 unter dem Titel „Eléments de syntaxe structurale“.]

TEŠITELOVÁ, M. (1967): On the role of nouns in lexical statistics. In: Prague Studies in Mathematical Linguistics 2, S. 121-131.

TEUPENHAYN, R. & ALTMANN, G. (1984): Clause length and Menzerath’s law. In: BOY, J. & KÖHLER, R. (Hrsg.): Glottometrika 6, S. 127-138.

THOIRON, P. (1986): Indice de diversité et mesure de la richesse lexicale. In: MULLER, CH. (Hrsg.): Méthodes quantitatives et informatiques dans l'étude des textes. Colloque international de CNRS à l'Université de Nice, 5-8 Juin 1985. Paris u.a.: Champian u.a., S. 831-840.

TULDAVA, J. (1995): Methods in Quantitative Linguistics. Trier: WVT.

TULDAVA, J. (1998): Probleme und Methoden der quantitativ-systemischen Lexikologie. Trier: WVT.

WEBER, H. J. (1997): Dependenzgrammatik. Ein interaktives Arbeitsbuch. 2., überarbeite Aufl. Tübingen: Narr.

WHEELER, E.S. (2003): Multidimensional scaling to visualize text preparation. In: Glottometrics 6, S. 65-69.

WILLIAMS, C. B. (1940): „A note on the statistical analysis of sentence-length as a criterion of literary style.“ In: Biometrika 31, S. 356-361.

WIMMER, G. & ALTMANN, G. (2001): Models of Rank-Frequency Distributions in Language and music. In: ALTMANN, G. et al. (Hrsg.): Text as a linguistic Paradigm: Festschrift in honour of Ludek Hrebicek. Trier, S. 283-294.

YULE, G. U. (1939): „On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship.“ In: Biometrika 30; 363-390.

ZIPF, G.K. (1932). Selected studies of the principle of relative frequency in language. Cambridge, Mass.: Harvard Univ.Press.

ZIPF, G.K. (1935): The Psycho-Biology of Language. Cambridge (Mass.).

ZIPF, G. K. (1972), Human behavior and the principle of least effort. An introduction to human ecology. New York: Hafner reprint.

ZIPF, G. K. (1949), Human behavior and the principle of least effort. Cambridge u.a.: Addison-Wesley.

ZÖRNIG, P., KÖHLER, R. & BRINKMÖLLER, R**.** (1990). Differential equation models for the oscillation of the word length as a function of the frequency. In: HAMMERL, R. (Hrsg.): Glottometrika 12, S. 25-40.

ZÖRNIG, P. & BORODA, M. (1990): Zipf-Mandelbrot’s law in coherent text: Towards the problem of validity. In: HAMMERL, R. (Hrsg.): Glottometrika 12, S. 41-60.

ZÖRNIG, P. & BORODA, M. (1992): The Zipf-Mandelbrot Law and the Interdependencies between Frequency Structure and Frequency Distribution in Cohent Texts. In: Glottometrika 13, S. 205-218.

ZWIRNER, E. & ZWIRNER, K. (1935): Lauthäufigkeit und Zufallsgesetz. In: Forschungen und Fortschritte 11, Nr. 4., S. 43-45.

ZWIRNER, E. & ZWIRNER, K. (1938): Lauthäufigkeit und Sprachvergleichung. In: Monatsschrift für höhere Schulen 37, S. 246-253.

Nachschlagewerke

JUILLAND, A. et al. (1970): Frequency dictionary of French words. The Hague u.a.: Mouton.

Le Petit Larousse illustré 2007. Paris: Larousse.

Softwaren

Altmann-FITTER (1997): Lüdenscheid: RAM-Verlag.

NLREG (Nonlinear Regression Analysis). Copyright 1992-2005 Phillip H. Sherrod.