QUANTITATIVE LINGUISTICS

Volume 59

Juhan Tuldava

Probleme und Methoden der quantitativ-systemischen Lexikologie

Aus dem Russischen
von
Gabriel Altmann
Reinhard Köhler

Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Tuldava, Juhan:
Probleme und Methoden der
quantitativ-systemischen Lexikologie/
Juhan Tuldava. -
Trier : WVT Wissenschaftlicher Verlag Trier, 1998
(Quantitative linguistics ; Vol. 59)
Einheitssacht.: Problemy i metody kvantitativno-sistemnogo issledovanija leksiki <dt.>
ISBN 3-88476-314-8

Grafiken: Johannes Kiel, Petra Vock
Umschlag: Brigitta Disseldorf
(M. Nottar, Agentur für Werbung und Design, Konz)

ISBN 3-88476-314-8

WVT Wissenschaftlicher Verlag Trier
Bergstraße 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504
Internet: http://www.wvttrier.de
e-mail: wvt@wvttrier.de
Internet QL: http://www.ldv.uni-trier.de:8080/~iqla.ql.books.html

Inhalt

Vorwort

Die Untersuchung der Lexik mit quantitativen Mitteln gehört zu den ältesten Forschungstraditionen innerhalb der quantitativen Linguistik; sie nimmt bereits im Werk von George Kingsley Zipf, der allgemein als der Begründer dieser linguistischen Disziplin gilt, eine zentrale Rolle und breitesten Raum ein. Der Wortbestand einer Sprache bzw. eines Textes als Untersuchungsgegenstand ist weit vielfältiger und vielschichtiger, als es auf den ersten Blick den Anschein haben mag, umfaßt er doch prinzipiell alle Eigenschaften und Zusammenhänge, die Wörtern und anderen lexikalischen Einheiten zukommen bzw. zugeschrieben werden können: phonetisch/phonologische, morphologische, semantische, pragmatische, syntagmatisch-distributionelle u.a. – und damit sowohl Charakteristika der Verwendung, die nur im einzelnen, konkreten Kontext definiert sind, als auch Eigenschaften, die von der einzelnen Verwendung abstrahieren (wie z.B. die Polysemie). Gerade die Gewinnung solcher abstrakten Eigenschaften und die Arbeit mit ihnen machen – wie vielleicht kein anderer linguistischer Untersuchungsaspekt – die Konsequenzen des Verhältnisses zwischen beobachtbaren Instanzen sprachlicher Äußerungen einerseits und linguistischen Konstrukten andererseits deutlich, bei dem also *Rede* (*Text*) – als Reales (Realisiertes) – *Sprache* – als Potentiellem (bzw. als Konstrukt) – gegenübersteht.

Trotz der auf der Hand liegenden Analogie zwischen diesem Verhältnis und dem zwischen Stichprobe und Population in der Statistik gibt es grundlegende methodologische und epistemologische Vorbehalte, die eine einfache Anwendung inferenzstatistischer Verfahren zum Schließen vom Text(korpus) auf die Sprache "als Ganze" höchst problematisch machen. Um nur ein methodologisches Problem zu nennen: Es ist kein einziger Fall einer linguistischen Untersuchung bekannt geworden, bei dem die in anderen empirischen Wissenschaften vielfach automatisch als gegeben vorausgesetzten Bedingungen erfüllt gewesen wären. Dazu gehören vor allem

(a) Repräsentativität: Keine Stichprobe kann repräsentative Sprachdaten in dem Sinne liefern, daß in dem in der Statistik üblichen Sinne gültige Schlußfolgerungen auf das "Sprachganze" möglich wären. Durch Vergrößerung der Stichprobe, z.B. durch Erweiterung eines Textkorpus um weitere Texte, ver-

größert sich dagegen die Diversität der Daten im Hinblick auf viele Parameter (Thematik, Stilistik, Genre/Funktionalstil etc.) und damit die Inhomogenität der Daten (s. Punkt b);

(b) die Homogenität der Daten: Nur homogene Stichproben sind für viele der meistverwendeten statistischen Verfahren geeignet. Diese Bedingung ist für Sprachdaten nur selten erfüllt, z.B. im Fall von Briefen, die spontan, ohne Unterbrechung und ohne nachträgliche Überarbeitung geschrieben wurden und nicht zu lang sind, so daß über den gesamten Prozeß der Textgenerierung konstante Randbedingungen angenommen werden können. Solche einzelnen, kurzen Texte sind allerdings gerade wegen ihrer Kürze nur bedingt aussagefähig (vgl. Punkt a);

(c) die Normalverteiltheit der Zufallsvariablen und der Abweichungen: Die wichtigsten Testverfahren, auf der eine Schlußfolgerung von der Stichprobe auf die Grundgesamtheit ja beruht, setzen voraus, daß die beobachteten Abweichungen von den erwarteten Werten der Zufallsvariablen normalverteilt sind. Diese Voraussetzung ist in der Sprache jedoch gewöhnlich nicht erfüllt, so daß eigentlich für jeden einzelnen Fall gesonderte Tests abgeleitet werden müßten (eine mathematisch äußerst unbequeme und in der Praxis nicht durchführbare Forderung);

(d) die Homöoskedastizität: Auch diese Bedingung, die gleichbleibende Varianz über alle Werte der betrachteten Zufallsvariablen, wird von Sprachdaten nicht generell erfüllt und muß besonders sorgfältig überprüft werden, bevor übliche Verfahren der Statistik angewendet werden dürfen.

Ein zentrales und für die Sprache typisches Phänomen ist auch die extreme Schiefe der Häufigkeitsverteilung von Wörtern in Texten, die dazu führt, daß im Bereich der seltenen Wörter stets – wie groß die analysierte Textbasis auch sei – eine nicht vernachlässigbare Unterrepräsentation vorliegt. Berücksichtigt man nun, daß viele andere linguistische Größen wie Länge, Polysemie, Polytextie etc. direkt oder indirekt funktional mit der Frequenz verknüpft sind, wird klar, daß sich die entsprechenden Besonderheiten von Sprachdaten auf *jede* linguistische Untersuchung auswirken können. Dies gilt für Signifikanztests von Verteilungsanpassungen und Regressionen ebenso wie für Verfahren des Textvergleichs u.a. Für die korrekte Anwendung statistischer Verfahren auf Sprachdaten (bzw. die Entwicklung neuer Methoden für solche Daten) ist daher außer einer guten Kenntnis der mathematischen Statistik allgemein auch immer die genaue Prüfung der jeweiligen Randbedingungen im Einzelnen erforderlich.

Zu unserem heutigen Kenntnisstand in bezug auf diese Probleme, zur Klärung vieler theoretischer und praktischer Fragen und zur Entwicklung eines umfangreichen Instrumentariums zu Analyse, Charakterisierung, Vergleich und Klassifikation von Texten und Stilen, für phonologische und lexikologische Untersuchungen und zu weiteren Gebieten der quantitativen Linguistik hat zu einem erheblichen Teil der Autor des vorliegenden Bandes, Juhan Tuldava, beigetragen. Ihm verdanken wir darüber hinaus zahlreiche systematische Darstellungen, die auch als Lehrtexte für Seminare der quantitativen Linguistik geeignet sind. Die meisten seiner Artikel und Bücher sind in Estnisch und Russisch geschrieben. Dies und die Behinderung des wissenschaftlichen Austausches durch die politischen Grenzen in den letzten Jahrzehnten tragen die Schuld daran, daß seine wichtigen Arbeiten bei uns noch immer zu wenig bekannt geworden sind. Die Buchreihe *Quantitative Linguistics* ist unter anderem gerade zu dem Zweck gegründet worden, diesen Austausch durch die Organisation von Übersetzungen (hierzu trug die Stiftung Volkswagenwerk mit finanzieller Hilfe entscheidend bei) möglich zu machen und zu beleben. Auf diese Weise gelang es auch, Teile von J. Tuldavas Werk in den westlichen Ländern zugänglich zu machen (vgl. z.B. den 1995 in dieser Reihe erschienenen Band *Methods in Quantitative Linguistics*, der zehn wichtige Arbeiten dieses Autors in englischer Übersetzung versammelt). Die Wertschätzung, die J. Tuldava in allen Teilen der Welt entgegengebracht wird, zeigt sich auch in der überaus großen Beteiligung an der Festschrift zu seinem 75. Geburtstag, die in Form von einem und einem halben Jahresband der internationalen Zeitschrift *Journal of Quantitative Linguistics* (1997/98) erschien.

Mit der nun vorliegenden Monographie wird unserer Leserschaft ein weiterer Teil der Forschungsergebnisse J. Tuldavas vorgestellt: eine überarbeitete, ergänzte und autorisierte Übersetzung des in russischer Sprache verfaßten und 1987 in Tallinn erschienenen Buchs *Проблемы и методы квантитативно-системного исследования лексики*. Es handelt sich um eine gründliche Einführung in die linguistisch-theoretischen und methodologischen Grundlagen der quantitativen Lexikologie und zugleich um eine umfassende Darstellung des heutigen Forschungsstands, die auch eine große Anzahl von Arbeiten bespricht, die im Westen aufgrund des Sprachproblems trotz ihrer Bedeutung bisher leider weitgehend unbekannt sind.

Mit dem Titel des Bandes unterstreicht der Verfasser eine – vor allem in der formalen Linguistik nicht unbedingt verbreitete – Einsicht: Die Lexik einer Sprache ist nicht die ungeordnete Zusammenfassung des "unsystematischen Rests", der übrigbleibt, wenn man die regelhaften Bereiche in der Grammatik dieser Sprache beschrieben hat. Zum einen ist es notwendig, einen fortschrittlicheren

"System"-Begriff zugrunde zu legen, der die längst überholte Bindung an den Determinismus aufgibt (und der uns mit den quantitativen Bereichen der Mathematik und vor allem mit der modernen Systemtheorie zur Verfügung steht), zum anderen gilt es, die in der Lexik vorhandenen nichtdeterministischen Regularitäten und gesetzesartigen Zusammenhänge zu entdecken, mathematisch zu beschreiben und (Gesetzes-)Hypothesen aufzustellen und zu testen, die für diese Zusammenhänge und Verteilungen verantwortlich sind.

J. Tuldava führt in seinem Buch in diese moderne sprachwissenschaftliche Disziplin ein, illustriert seine Aussagen und Befunde mit Hilfe von Originaldaten aus verschiedenen Sprachen, gibt Beispiele für praktischen Anwendungen der Resultate und weist auf Querverbindungen zu anderen Forschungsrichtungen hin. Es ist als Lehr- und Lernbuch für Anfänger bestens geeignet und zugleich eine reiche Fundgrube für den erfahrenen Forscher. Ich wünsche der deutschen Fassung dieses exzellenten Werks weite Verbreitung und intensive Nutzung.

Reinhard Köhler

1. Theoretisch-methodologische Grundlagen der quantitativ-systemischen Untersuchung der Lexik

Unter den Bedingungen der wissenschaftlich-technischen Revolution und der Informationsexplosion, die für unsere Zeit so charakteristisch sind, bekommen die theoretischen und methodologischen Probleme der Wissenschaft eine vorrangige Bedeutung. Gegenwärtig erreicht die quantitative Linguistik einen Zustand, in dem es unumgänglich ist, die theoretischen Konsequenzen zusammenzufassen, die Begriffe und Methoden der Forschung zu präzisieren, aber auch einige neue Aspekte des Forschungsgegenstandes zu erhellen und Wege zu seiner Erklärung zu skizzieren. Insbesondere die Untersuchung der Lexik vom systemischen Blickwinkel aus und unter Anwendung quantitativer Methoden verlangt die Ausarbeitung der methodologischen Grundlagen eines komplexen quantitativ-systemischen Ansatzes. Dies wird uns ermöglichen, einige grundlegende Probleme der quantitativen Linguistik von einem einzigen Gesichtspunkt aus zu durchdenken.

1.1. Begründung der quantitativ-systemischen Untersuchung der Lexik

Das Objekt und die Ausgangsprinzipien der Untersuchung

Wenn man als Objekt der lexikologischen Untersuchung Struktur und Funktion der Lexik einer gegebenen Sprache (oder einer Gruppe von Sprachen) betrachtet, dann muß man angesichts der Komplexität, der Vielseitigkeit und Vielschichtigkeit des Forschungsobjektes den Untersuchungsbereich im Hinblick auf einen konkreten Ansatz abgrenzen und den Aspekt oder die Seite des Aspekts, die man erforschen will, präzisieren. Das bedeutet, daß das Erkennen des Objekts als einer ontologischen Entität einen vielseitigen Prozeß darstellt, in dem unterschiedliche gnoseologische Aspekte eine Rolle spielen. Die Struktur und die Funktion der Lexik als Untersuchungsobjekt äußern sich durch diverse Erscheinungsformen, die unterschiedliche Forschungsobjekte ergeben, welche wiederum mit spezifischen Methoden erforscht werden. Es gibt z.B. den historisch-vergleichenden, den strukturalistischen, den linguostatistischen Ansatz zur Untersuchung der Lexik. Das gegebene Objekt wird bereits am Anfang der Analyse in einen neuen Erklärungskontext gestellt, was dem Prinzip der Beschreibungspluralität der Objekte der reellen Welt

entspricht und letzten Endes das Verhältnis zwischen Vielfalt und Einheit der Welt widerspiegelt. Um ein Objekt ganzheitlich zu erfassen, muß man verschiedene Aspekte miteinander verbinden, eine Synthese durchführen, deren Voraussetzung eine tiefer gehende Einzelanalyse unterschiedlicher Aspekte des Objekts war.

Die Bestimmung des Forschungsobjekts hängt sowohl von der gestellten Aufgabe als auch vom Umfang und Tiefe des Wissens im gegebenen Bereich ab, über das die Wissenschaft bereits verfügt. Geht man von der allgemeinen Aufgabe aus, die quantitativen Aspekte der Lexik einer konkreten Sprache im Zusammenhang mit der Erhellung einiger wichtiger Gesetzmäßigkeiten der Struktur und Funktion der Lexik allgemein im Sprechprozeß zu beschreiben, so muß man den gegenwärtigen Entwicklungsstand der quantitativen Linguistik und die Möglichkeiten der Integration des vorhandenen Wissens in ein konsistentes System zugleich betrachten. Wir akzeptieren die Definition der quantitativen Linguistik von Piotrowski, Bektaev und Piotrovskaja (1977:8), die unter diesem Begriff die Erforschung und Erklärung linguistischer Erscheinungen mit Hilfe der quantitativen Mathematik (Wahrscheinlichkeitstheorie, mathematische Statistik, Informationstheorie u.a.) verstehen.[1] Die quantitative Linguistik kann man der "kombinatorischen" Linguistik gegenüberstellen, die sich auf den "nichtquantitativen" Teil der Mathematik stützt (Mengenlehre, mathematische Logik, Theorie der Algorithmen usw.). Die quantitative und die kombinatorische Linguistik stellen zwei Seiten des allgemeineren Oberbegriffs "mathematische Linguistik" dar. Eine besondere Bedeutung in der mathematischen Linguistik kommt der statistisch-kombinatorischen Modellierung zu (Andreev 1967), in der sich der mengentheoretische, der algorithmische und der statistische Ansatz zur Spracherforschung vereinigen.

Die praktischen und die theoretischen Untersuchungen im Bereich der mathematischen Linguistik haben im letzten Jahrzehnt einen derartigen Aufschwung erlebt, daß sich daraus ein fruchtbarer Boden für einen neuen integrativen Ansatz zur quantitativen Erforschung der Lexik ergab. Es haben sich zwei grundlegende Prinzipien einer derartigen Forschung herauskristallisiert: *Das Prinzip des systemischen und das des probabilistisch-statistischen Charakters* der Organisation der Lexik. Die Vereinigung dieser Prinzipien liefert uns die systemisch-probabilistische Herangehensweise, die in vielen Arbeiten der letzten Zeit mit Erfolg anwendet wurde (Alekseev 1977; Bektaev 1978; Neljubin 1983 u.a.). In dieser neueren Phase der Entwicklung der quantitativen Linguistik wurde klar, daß es möglich ist, Begriffe wie "probabilistisches System in der Linguistik" und "Methoden der Erforschung probabilistischer Gesetzmäßigkeiten und ihrer Funktion" noch exakter zu bestimmen und noch tiefer zu untersuchen. Es hat sich gezeigt, daß außer probabilistisch-statistischen und informationstheoretischen Methoden auch andere Methoden der

[1] "Linguostatistik" oder "Sprachstatistik", d.h. die Erforschung linguistischer Objekte mit Hilfe traditioneller statistischer Methoden, ist ein spezieller Fall der quantitativen Linguistik.

quantitativen Mathematik, wie z.B. Analysis und Funktionentheorie (Piotrowski u.a. 1977) eine große Rolle spielen können. In letzter Zeit verwendet man bei der Untersuchung der Systeme und Subsysteme in der Linguistik zunehmend auch die Theorie der unscharfen Mengen (Zadeh 1965, 1976; Piotrowski 1979, 1984). Ein charakteristischer Zug der neueren Forschung in der quantitativen Linguistik ist das Bemühen, die theoretischen Grundlagen der quantitativen Texttypologie im Rahmen einer allgemeinen Textlinguistik auszuarbeiten (z.B. Alekseev 1981).

Die Erweiterung des Arsenals der Forschungsmethoden der quantitativen Linguistik durch Übernahme neuer quantitativer Methoden aus der Mathematik als Erweiterung der traditionellen probabilistisch-statistischen Methoden macht es notwendig, die Bezeichnung für den allgemeinen Ansatz zur Erforschung quantitativer Eigenschaften linguistischer Objekte etwas zu präzisieren. Es scheint, daß die geeignetste verallgemeinernde Bezeichnung *quantitativ-systemische Analyse* wäre (Tuldava 1979), da sie den systemischen Charakter des mit Hilfe verschiedener Methoden der quantitativen Mathematik untersuchten Objekts unterstreicht. In Übereinstimmung mit unserer Konzeption bleibt auch die These über die Analyse der Lexik (und anderer linguistischer Objekte) als eines probabilistischen Systems, charakterisiert nicht nur durch die Zufälligkeit der Parameter (niedrigere Organisationsstufe), sondern auch durch eine gewisse Stabilität und Regularität der Masse zufälliger Ereignisse (höhere Organisationsstufe), in Kraft. In probabilistischen Systemen werden daher nach aktueller Auffassung deterministische Zusammenhänge nicht ausgeschlossen, aber sie werden auf eine höhere, allgemeinere Organisationsstufe übertragen (Sačkov 1971). Im philosophischen Sinne wird hier ein Zusammenhang zwischen Zufall und Notwendigkeit angenommen, der zusammen mit der allgemeinen Verbundenheit der Erscheinungen der objektiven Realität (als Grundlage des systemischen Ansatzes zur Untersuchung der Objekte und Erscheinungen der realen Welt) die philosophisch-methodologischen Postulate des in dieser Arbeit dargestellten quantitativ-systemischen Ansatzes zur Erforschung linguistischer Objekte darstellt.

Die *Forschungsobjekte* bei einem quantitativ-systemischen Ansatz sind (bei probabilistischer Interpretation) die quantitativen Eigenschaften und Gesetzmäßigkeiten von Struktur und Funktion der Lexik (oder anderer linguistischer Objekte), die man unter systemischer Perspektive und Betonung der probabilistischen Natur der Sprache untersucht. Die Lexik wird als ein probabilistisches System betrachtet, mit allen Eigenschaften, die solchen Systemen eigen sind, darunter Stabilität und Variabilität. Ähnlich wie andere Systeme enthält auch die Lexik Subsysteme und kann als ein vielschichtiges, multilaterales Gebilde untersucht werden; gleichzeitig ist sie selbst ein Subsystem des Gesamtsystems Sprache (Sprechtätigkeit) und stellt eine bestimmte Ebene in der Hierarchie sprachlicher Erscheinungen dar.

Der quantitativ-systemische Ansatz zur Erforschung der Lexik, der sich auf den Systemcharakter und den probabilistisch-quantitativen Aspekt der Sprache

stützt, stellt den zentralen Begriff und das zusammenfassende Prinzip der vorliegenden Untersuchung dar. Dieses Grundprinzip soll alle anderen Begriffe, Entscheidungen und Gesetze in einem bestimmten Ganzen vereinen und der folgenden Untersuchung empirischer Tatsachen eine gemeinsame Richtung verleihen. Bevor wir aber an die Untersuchung des empirischen Materials und der grundlegenden Gesetze der Lexik herangehen, müssen wir einige allgemeinere Begriffe und Kategorien, die das oben erwähnte grundlegende methodologische Prinzip beleuchten und präzisieren, detaillierter darstellen. Man sollte dabei nicht vergessen, daß das Objekt der Beschreibung ein bestimmter Aspekt des Sprechprozesses (s. Kap. 1.2) ist und daß man letzten Endes ein linguistisches Problem lösen möchte. Es ist bekannt, daß sämtliche Prinzipien und Kategorien nur dann der wissenschaftlichen Erforschung eines Objekts dienen, wenn sie vorher entsprechend überarbeitet und interpretiert wurden, im konkreten Fall im Lichte der Bedürfnisse der quantitativ-systemischen Analyse der Lexik. Betrachten wir im folgenden unter diesem Aspekt den Begriff der Systemizität der Lexik, ihre quantitativen und probabilistischen Aspekte und eine Reihe von grundlegenden Begriffen und Annahmen, die zum konzeptionellen Apparat des quantitativ-systemischen Ansatzes zur Untersuchung der Lexik gehören.

Der Systemcharakter der Sprache und der Lexik

Die Mehrheit der Linguisten unserer Zeit stimmt darin überein, daß Systemizität eine der wesentlichsten Eigenschaften der Sprache ist. "Der Sprache Systemizität abzusprechen wäre grundfalsch", sagt B.A. Serebrennikov (1973:295), "da die Systemizität der Sprache nicht nur auf ihrer Funktion als Kommunikationsmittel beruht; vielmehr ist sie notwendigerweise auch mit einigen rein physiologischen und psychologischen Eigenheiten des Menschen verbunden, und zu ihrer Begründung kann die Existenz der Systemizität der den Menschen umgebenden Welt dienen." Methodologische Probleme des systemischen Ansatzes wurden von Linguisten oft anhand von konkreten Beispielen der Entwicklung und der Funktion der Sprache diskutiert. Es wurde festgestellt, daß im Prinzip "darin Übereinstimmung herrscht, daß Sprache zu den systemischen Gebilden gehört. Allerdings werden die Begriffe 'System' und 'systemisch' in unterschiedlichen Werken unterschiedlich verstanden" (Solncev 1977:12).

Gleichzeitig weisen viele Autoren darauf hin, daß der Systemcharakter der *Lexik* (des Wortschatzes) bezüglich der allgemeinen Systemizität der Sprache ihre Eigenheiten hat. Dies hängt damit zusammen, daß die Lexik ein kompliziertes und widersprüchliches Objekt darstellt, das sich nur schwer in eine Systematisierung und eine Klassifikation fügen läßt. Die Systemizität der lexikalischen Ebene wird viel seltener untersucht als die der anderen Sprachebenen.

Die Schwierigkeiten der systemischen Analyse der Lexik erklären sich aus ihren "wesentlichen Charakteristika", wie die Unzählbarkeit ihrer Einheiten, die unbeschränkte Kombinierbarkeit der Wörter, die Komplexität und Heterogenität der Wortzusammenhänge in der Sprache und in der Rede, die außerlinguistische Determiniertheit der Wörter usw. Dies alles trägt zu der Schwierigkeit der Analyse der Lexik bei, die uns als "schlecht organisiertes, diffuses System" erscheint, die aber nichtsdestoweniger erfolgreich im Prozeß der menschlichen Kommunikation funktioniert. Es entsteht der Eindruck, daß die Lexik keine streng reguläre und starre Organisation braucht, um zu funktionieren, da gerade ihre Elastizität dem Sprachprozeß seine Wendigkeit und Manövrierfähigkeit verleiht.

Die Ansichten über die Systemizität der Lexik (und der Sprache im Ganzen) hängen letzten Endes davon ab, wie man den Begriff "System" definiert. Wenn einige Forscher von Erscheinungen der "Antisystemizität" in der Sprache (Budagov 1978; Filin 1979) oder von Einheit und Verwobenheit von systemischen und nichtsystemischen Prozessen sprechen, dann verstehen sie offensichtlich unter "systemisch" eine starre Regularität der Beziehungen zwischen Elementen, Stabilität und Widerpruchsfreiheit des Zusammenwirkens der Komponenten des Ganzen. Dies sind im Grunde die Merkmale von sogenannten dynamischen *deterministischen* Systemen, die im Unterschied zu den probabilistischen (statistischen) Systemen etwas andere Eigenschaften haben. In der vorliegenden Arbeit geht es um einen systemischen Ansatz, wobei wir der Auffassung sind, daß sowohl deterministische als auch stochastische Ansätze zugelassen sind.

Das Aufkommen des systemischen Ansatzes in unserer Zeit hängt vor allen Dingen mit der Notwendigkeit zusammen, große komplexe Systeme zu untersuchen, die in der Regel schwach strukturiert sind und teilweise nichtformalisierbare Elemente enthalten, wobei diese Systeme oft unter Bedingungen der Unbestimmtheit funktionieren. Die Sprechtätigkeit im Ganzen und ihre Subsysteme, darunter auch die Lexik, gehören zu solchen komplexen Gebilden, deren Erforschung einen systemischen Ansatz dringend benötigt.

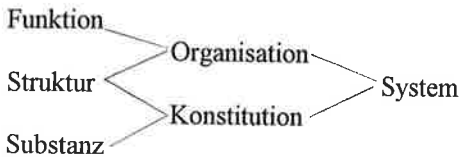
In der allgemeinsten Form kann man ein *System* als "ein ganzheitliches Objekt, das aus miteinander verknüpften Elementen besteht" bezeichnen (Solncev 1977: 14). Die Gesamtheit der Beziehungen zwischen den Elementen bezeichnet man als *Struktur* des Systems, wobei man bei der Definition des Systems den stabilen Charakter der Beziehungen zwischen den Elementen hervorhebt, was eine Voraussetzung für Stabilität und erfolgreiches Funktionieren des Systems darstellt. "Die Struktur" – sagt Ovčinnikov (1966:268) – "garantiert den Fortbestand des Systems, daher stellt sie das dar, was in der wandelbaren Existenz eines Gegenstands relativ konstant bleibt."

Die Struktur des Systems oder das Netz der stetigen Beziehungen zwischen den Elementen stellt also den wichtigsten Aspekt des Systems dar, und die Aufgabe des Forschers besteht darin, eine adäquate Art der Beschreibung zu finden. Es muß

bemerkt werden, daß - obwohl die Struktur unter bestimmten Bedingungen als Forschungsobjekt ohne Berücksichtigung der Systemsubstanz untersucht werden kann (dies ist das Objekt der Allgemeinen Systemtheorie) - die Struktur ohne die Systemsubstanz (Elemente) nicht existieren kann, d.h. Struktur und Substanz muß man (in empirischen Wissenschaften) in ihrem Zusammenhang untersuchen.

Die materielle Zusammensetzung (Substanz, Elemente) und die Beziehungen zwischen den Elementen (Struktur) bilden zusammen die "Konstitution" des Gesamtsystems. Der systemische Ansatz bedingt jedoch die gleichzeitige Untersuchung der Konstitution und der *Funktion* des systemischen Objekts. Unter der Funktion des Systems versteht man - im weitesten Sinne - die Interaktion des Systems mit seiner Umgebung, darunter auch die mit anderen Systemen und Subsystemen. Im engeren Sinne ist die Funktion des Systems das Zusammenwirken der Elemente im Rahmen der konstituierenden Struktur. Die Funktion stellt offensichtlich einen Bestandteil des Systems dar, der mit der Konstitution des Systems untrennbar verbunden ist. Erst ihre Synthese, die Einheit der drei Hauptbestandteile des Systems, nämlich die Menge der Elemente, die Struktur und die Funktion, machen ein System aus. Dabei kann man die Struktur und die Funktion gemeinsam als die *Organisation* des Systems bezeichnen. In diesem Sinne ist ein System die Menge seiner Elemente plus ihrer Organisation (Struktur und Funktion).

Zusammenfassend:



Der Begriff *Lexik* hat zwei Grundbedeutungen. Zum einen wird die Lexik als eine Menge von separaten Einheiten (Wortformen oder Lexemen) betrachtet. In dem Falle untersucht man die systemischen Zusammenhänge und Gesetzmäßigkeiten des Vokabulars und des Textes in der Richtung von den konstituierenden Teilen hin zum Ganzen (in der Allgemeine Systemtheorie bezeichnet man das als *Identifikationsproblem*). Zum anderen betrachtet man die Lexik als eine einheitliche Gesamtheit mit verschiedenen Parametern (Gesetze der Zusammenhänge der Komponenten usw.), die zum Objekt als Ganzem gehören. In diesem Falle geht die Richtung der Untersuchung vom Ganzen zu den Komponenten (in der Allgemeinen Systemtheorie als *Rekonstruktionsproblem* bezeichnet). Diese zwei Grundrichtungen (Aspekte) der systemischen Lexikologie kann man als analytisch bzw. synthetisch bezeichnen.

Der systemische Charakter der Lexik als einer einheitlichen Gesamtheit offenbart sich vor allen Dingen in der Aufteilung der Wörter in unterschiedliche lexikalische Gruppen - Subsysteme, in denen es bestimmte Beziehungen und Verbindungen zwischen den Elementen gibt. Man kann sagen, daß die Lexik vor allen Dingen eine *Gesamtheit verschiedener Subsysteme* ist. Die innere Struktur und die Bedingungen der Funktion dieser Subsysteme ist wegen der Vielzahl und der Diversität der Elemente und der Komplexität ihrer gegenseitigen Beziehungen noch nicht hinreichend bekannt. Es gibt keine speziellen Untersuchungen, in denen gezeigt worden wäre, wie diese Subsysteme untereinander systemisch verbunden sind. Üblicherweise begnügt man sich mit der Bemerkung, daß "unterschiedliche Glieder des Systems gegenseitig verbunden und bedingt, aber nicht im stabilen Gleichgewicht sind" (Serebrennikov 1972:52). Für das lexikalische System insgesamt (Subsystem im Gesamtsystem Sprache) ist ein relatives oder Fließgleichgewicht charakteristisch. Dabei kann man feststellen, daß es ein "System ist, das spontan über Jahrtausende gewachsen ist und sich verändert hat. Daher gibt es in jeder Sprache viel "Unlogisches", "Irrationales" und "Widersprüchliches" (Maslov 1975:33).

Es entsteht die Frage, wie ein Modell dieses Systems aussehen sollte, das in einem Komplex die gesamte Mannigfaltigkeit des Objektes adäquat erfassen soll und dabei gleichzeitig auch die Aspekte der Regularität und Irregularität (Stabilität und Instabilität usw.) der Sprache als Ganzem und ihrer Subsysteme, speziell der Lexik, einschließen soll.

Probabilistisches System

Wenn man unter *System* eine strenge Regularität der Beziehungen und Zusammenhänge unter den Elementen, eine starre funktionale Verbindung der Komponenten des Gesamtgebildes versteht, dann ist es offensichtlich, daß man es nur mit *einem* der vielen Ausprägungen von Systemen zu tun hat, nämlich mit dem sogenannten dynamischen (deterministischen) System. In der heutigen Systemforschung untersucht man neben dynamischen (deterministischen) auch *probabilistische* Systeme, in denen Ganzheitlichkeit und Stabilität des Systems mit hinreichend großer Autonomie der Teile verknüpft sind (Sačkov 1971; Kravec 1976). Bei solchen Systemen kann man von einem regulären Wechsel "zwischen streng deterministischer und probabilistisch-statistischer Regulation" sprechen (Blauberg 1977:10). In den Untersuchungen zur Theorie probabilistischer Systeme hat sich gezeigt, daß ihre Parameter zu unterschiedlichen Ebenen gehören, als ob sie in zwei Klassen aufgeteilt wären: Zufallsereignisse (auf niedrigerer Ebene) und Gesetzmäßigkeiten, Regularitäten in der Masse der Zufallsereignisse (auf höherer Ebene). Die Charakteristika einer höheren Ebene, die die ganzheitliche Struktur des Systems bestimmen, wirken sich dabei nicht direkt auf jedes konkrete Zufallsereignis aus, besser gesagt, sie

bestimmen diese Zufallsereignisse (d.h. Charakteristika der niedrigeren Ebene) nur allgemein, ganzheitlich. Bei einem derartigen Ansatz wird klar, daß "nicht-systemische" Erscheinungen sich in das Gewebe des Systems als eines ganzheitlichen Gebildes organisch einfügen, welches sowohl reguläre als auch irreguläre (zufällige, vorübergehende, variable usw.) Eigenschaften besitzt.

Manchmal spricht man von "unterschiedlichen Graden der Systemizität" sprachlicher Erscheinungen (Serebrennikov 1972:4). Hier betrachtet man offensichtlich verschiedene Grade der Regularität, Symmetrie usw. auf den verschiedenen Ebenen des Sprachsystems. In diesem Sinne kann man vom "Zentrum" und von der "Peripherie" des Systems sprechen, wobei man der letzteren alle Irregularitäten und nicht-systemischen Erscheinungen zuschreibt. In der Tat kann man in der Struktur und in der Funktion der Sprache sowohl reguläre, eindeutig determinierte, als auch völlig zufällige oder irreguläre Erscheinungen und Prozesse sehen. Sie bilden, sozusagen, Gegenpole des Systems, wobei die meisten realen Sprachen und Sprechprozesse eine Zwischenposition einnehmen. Dies kann man als die Konkretisierung des Grundprinzips der probabilististischen Systeme in dem Sinne betrachten, daß die Parameter der niedrigeren Ebene, die zufällige Ereignisse darstellen, in unterschiedlichem Maße von den Parametern der höheren Determinationsstufe reguliert werden. Man muß sich dessen bewußt sein, daß eine volle Determiniertheit eines komplexen Systems (oder der Teile eines solchen Systems) eher die Ausnahme als die Regel darstellen. Volle Determiniertheit ("dynamische Gesetzmäßigkeit") ist im Grunde eine statistische Gesetzmäßigkeit, die mit der Wahrscheinlichkeit 1 gilt. Dies gibt uns die Möglichkeit, den folgenden *probabilistischen Ansatz in dem Sinne zu verallgemeinern, daß er den deterministischen als Spezialfall enthält.*

Als wichtig ist festzuhalten, daß die probabilistische Behandlung realer Phänomene nicht nur auf der Vorstellung vom probabilistischen Charakter des Wissens, sondern vor allen Dingen auf der Vorstellung beruht, daß "selbst das Objekt der Erkenntnis in seiner Bewegung und Veränderung, in seinen gegenseitigen Beziehungen zu anderen Objekten, probabilistischen Gesetzen folgt" (Stoff 1972:131). Der probabilistische Ansatz zur Erforschung von Systemen ist folglich durch das Forschungsobjekt selbst bedingt, nämlich durch das komplexe und vielseitige Phänomen, in dem der notwendige Zusammenhang einzelner Komponenten, ihre kausale Bedingtheit, offensichtlich nicht in Form eines reinen Determiminismus dargestellt werden kann, sondern eher als ein Zusammenhang von Zufälligkeiten und den hinter ihnen verborgenen Notwendigkeiten. Diese Charakterisierung kann man in vollem Maße auf unser Forschungsobjekt, die Lexik, übertragen.

Der probabilistische Ansatz zur Erforschung von Systemen hat eine philosophische Begründung. Sie beruht auf der Konzeption der Verbindung zwischen den Kategorien des Zufalls und der Notwendigkeit (Monod 1970). Aus dieser philosophischen Grundlage erwächst die sich immer mehr verbreitende "probabilistische Denkweise", die die Zufälligkeit als eine Erscheinungsform der Notwendigkeit in die Struktur theoretischer Systeme einbezieht. Dies erlaubt uns, zur Erforschung sehr komplexer Objekte vorzudringen, für die die gegenseitige Überlagerung strikter Kausalität mit Wahrscheinlichkeit in ihrer Struktur charakteristisch ist.

Es ist zu bemerken, daß der Begriff der *Wahrscheinlichkeit* an sich kein rein mathematischer Begriff ist. Die Wahrscheinlichkeit kann man nicht nur durch statistische Charakteristika der Objekte (z.B. relative Häufigkeit) sondern durch sehr unterschiedliche Eigenschaften der Objekte (Nutzen, Wert usw.) ausdrücken. Verallgemeinernd kann man sagen, daß der Begriff der Wahrscheinlichkeit "beliebige Prozesse oder Situationen, in denen alternative Ausgänge objektiv möglich sind" charakterisiert (Suslov 1978). Beim quantitativen Ansatz ist man vor allem an dem quantitativen Ausdruck der Wahrscheinlichkeit interessiert, und dies führt zu der logisch-mathematischen Auffassung der Wahrscheinlichkeit; laut derer ist sie "eine objektive Charakteristik des Grades der Möglichkeit des Eintretens eines Ereignisses unter den gegebenen Bedingungen, die sich unzählige Male wiederholen können" (Filosofskaja ėnciklopedia, vol. 1: 244). Dies bedeutet, daß der Begriff der Wahrscheinlichkeit auf Massenerscheinungen, die sehr oft vorkommen, angewandt wird. Die *massenhafte* Anwendung linguistischer Einheiten ist aufgrund der Funktion der Sprache als Kommunikationsmittel gegeben. Bei dem probabilistischen Ansatz wird der Faktor der Massenhaftigkeit üblicherweise mit dem Begriff der *Zufälligkeit* assoziiert. Diesen Begriff benutzt man zur Bestimmung der Spezifizität der Massenerscheinung und charakterisiert dadurch bestimmte objektive Zustände. Zufälligkeit entsteht "infolge des Zusammenwirkens von Tatsachen oder Überschneidung von Notwendigkeiten" (Ruzavin 1978:227). Es ist bekannt, daß Wörter im Text nicht nur rein zufällig vorkommen, sondern auch die Folge einer gezielten Auswahl sind. Bei der Masse linguistischer Ereignisse jedoch (z.B. bei der Generierung eines Textes) unterliegt die gezielte Auswahl dem Einfluß einer derartig großen Anzahl sehr unterschiedlicher, darunter auch außerlinguistischer Faktoren, daß man praktisch von zufälligem Auftreten linguistischer Einheiten im Sprechfluß ausgehen darf. Allerdings stellt der Fluß individueller "zufälliger" Ereignisse nur die äußere Auswirkung der inneren, d.h. "notwendigen" Tendenzen dar. Dies spiegelt das Wesen der probabilistischen Systeme wider, in denen Zufälligkeit und Notwendigkeit eng verbunden sind.

Das Moment der *Notwendigkeit* erscheint im probabilistischen System auf zwei Arten: Erstens, als relative Stabilität der Häufigkeiten unterschiedlicher Elemente oder Gruppen von Elementen (dies ist eben die innere Eigenschaft der Wahrscheinlichkeit, die in der Realität als "die Tendenz, sich unter gegebenen Bedingungen um einen konstanten Wert zu gruppieren" zutage tritt, vgl. Kolmogorov 1956:274)[2]. Zweitens, in Form einer festen Verteilung der Elemente, die die Exi-

[2] Zu der sog. von Mises-Fundierung der Wahrscheinlichkeitstheorie s. Alimov (1980).

stenz einer inneren Ordnung im System ausdrückt. Die Verteilung als ein verallgemeinerter, integraler Begriff ist die wichtigste Charakteristik eines probabilistischen Systems. Unten werden wir zeigen, daß stabile Verteilungen zum Objekt unserer Untersuchung, dem quantitativen Aspekt des lexikalischen Systems gehören.

Die Darstellung des Untersuchungsobjekts in Form eines probabilistischen Systems verlangt eine detailliertere Besprechung der Bedingungen und Möglichkeiten des quantitativen Ansatzes für die untersuchten Erscheinungen, im konkreten Fall für die lexikalischen Phänomene und die Klärung der fundamentalen Frage des Zusammenhangs zwischen qualitativen und quantitativen Aspekten der Sprache.

Der quantitative Aspekt

In unserer Zeit besteht keine Notwendigkeit mehr, die Anwendung quantitativer Kriterien in linguistischen, darunter lexikologischen, Untersuchungen zu verteidigen. Quantitative (besonders statistische) Methoden sind schon seit langem in der Linguistik etabliert. Um quantitative Methoden bei der Untersuchung von Forschungsobjekten anwenden zu können, reicht es, "daß die Eigenschaften dieser Objekte Wiederholbarkeit, Periodizität und in einem bestimmten Ausmaß invariante Beziehungen sowie eine gesetzmäßige Verteilung ihrer Parameter usw. besitzen" (Sadovskij 1974:42-43). Insbesondere die *Wiederholbarkeit* (Rekurrenz, Periodizität) sprachlicher, darunter lexikalischer, Einheiten, ihre Reproduktion in verschiedenen Texten, ist die wichtigste Bedingung für die Quantifizierung des sprachlichen Materials und für die Anwendung verschiedener mathematischer Methoden bei seiner Analyse.

Die Nützlichkeit und Wichtigkeit des quantitativen Ansatzes bei der Erforschung linguistischer Objekte wurde bereits von vielen prominenten Linguisten der Vergangenheit und der Gegenwart betont. "Man muß vermehrt quantitatives, mathematisches Denken in der Sprachwissenschaft einsetzen", sagte Baudouin de Courtenay (1963:17). Bei der Erörterung der Verwendungshäufigkeit unterschiedlicher Worttypen in unterschiedlichen schriftlichen und gesprochenen Stilen bemerkte V.V. Vinogradov (1938:176-177), daß "exakte Ermittlungen in diesem Bereich helfen würden, strukturell-grammatische und teilweise auch semantische Unterschiede zwischen den Stilen festzustellen". Jarceva (1970) schreibt: "Häufigkeit gehört zu der funktionalen Seite des Sprachsystems [...] Die Erfassung der Häufigkeit einer sprachlichen Erscheinung ist eine nützliche Methode der Analyse". Gleichzeitig aber weisen einige Forscher auf die bekannte Einschränkung mathematischer Methoden in der Linguistik hin, wobei sie ihre Wichtigkeit nicht bestreiten. "Mathematische Methoden der Analyse sprachlicher Erscheinungen", schreibt Filin (1979:27), "sind für die Entwicklung unserer Wissenschaft vielversprechend, haben aber ihre Grenzen." Filins Ansicht nach "hat die Sprache nicht

nur eine quantitative Seite, die man berechnen kann. Sie unterscheidet sich von Computer- und algorithmischen Sprachen [...] wesentlich dadurch, daß ihre Elemente (Wörter, Sätze, grammatische Formen u.a.) mehrdeutig sind, die Fähigkeit haben, neue übertragene Bedeutungen und Bedeutungsnuancen zu bilden, ganz zu schweigen von der unendlichen Vielfalt ihrer Anwendungen [...]. Die assoziative Verwobenheit der Sprachelemente ist dermaßen kompliziert und grenzenlos (wie unsere Erkenntnis), daß sie sich auch den ausgeklügelsten Methoden verschließt".

Diese Aussage kann man in dem Sinne als korrekt akzeptieren, daß mathematische Methoden - im gegebenen Fall quantitative Methoden - in der Tat nicht imstande sind, beliebige Probleme der Analyse sprachlicher Erscheinungen zu lösen. Der quantitative Ansatz ist nur imstande, einen bestimmten Aspekt der Sprache und der Rede zu erfassen. Dies ist aber ein fundamentaler Aspekt der Sprache, der eine Reihe von wichtigen Seiten der Sprechtätigkeit, die man nicht rein qualitativ aufdecken kann, widerspiegelt. Bei der quantitativen Analyse muß man manchmal die linguistischen Tatsachen vereinfachen (z.B. wenn man bei statistischen Zählungen die Polyvalenz und die vielfältigen Nuancen der Bedeutung außer acht läßt). Bei dieser Analyse kann man aber im Prinzip einen viel differenzierteren Zugang zu der Polysemie, Polyvalenz und zu anderen Eigenschaften der Wörter bekommen, als es bei einer rein qualitativen Analyse möglich ist. Oft zeigt sich aber, daß "die assoziative, unendliche Verwobenheit der Sprachelemente dermaßen kompliziert ist", daß sie weder mit quantitativen noch mit qualitativen Analysen völlig erfaßbar ist. Außerdem muß man darauf hinweisen, daß eine rein qualitative Analyse komplexer Erscheinungen oft im Bereich subjektiver, willkürlicher Interpretationen verharrt.

Wie bekannt ist, sind *Qualität* und *Quantität* komplementäre Begriffskategorien und durch sie "wird ein Gegenstand vollständig, restlos erfaßt" (Šeptulin 1980: 36). Folglich gilt, daß eine allseitige Analyse eines komplexen Objekts notwendigerweise sowohl qualitative als auch quantitative Momente einschließt Bei der quantitativen Analyse linguistischer Objekte muß man inhaltlichen Fragen sowohl bei der Quantifizierung des Materials (besonders bei der Bestimmung der Maßeinheiten) als auch bei der Interpretation der Resultate besondere Aufmerksamkeit widmen. Auf diese Weise setzt jegliche quantitative Charakterisierung sprachlicher Erscheinungen die qualitative Charakterisierung voraus. Gleichzeitig muß man betonen, daß umgekehrt die qualitative Bestimmtheit eines linguistischen Objekts wesentlich von der Anzahl der Elemente, aus denen es besteht, von der Anwendungshäufigkeit oder von der Assoziation (Korrelation) der Elemente abhängt. Man kann eine enge Verknüpfung von qualitativen und quantitativen Charakteristika der Sprache feststellen; ihre gleichzeitige Betrachtung eröffnet breite heuristische Möglichkeiten der Erforschung sprachlicher Prozesse und Erscheinungen. Eine derartige Analyse erlaubt beispielsweise, die Qualitätsprognosen auf Grundlage der Quantität und umgekehrt. Es ergibt sich die Möglichkeit der Erforschung gesetzmäßiger Kor-

relationen quantitativer Charakteristika mit qualitativen Faktoren (z.B. des Zusammenhangs zwischen Häufigkeit und morphologischer Struktur der Wörter). In vielen Fällen können quantitative Charakteristika als Signale dienen, die die Aufmerksamkeit des Forschers auf einige - der einfachen Beobachtung verborgene - Eigentümlichkeiten und Gesetzmäßigkeiten von Individual- und Funktionalstilen auf sich ziehen.

Man kann daraus folgern, daß die quantitative Erforschung sprachlicher Erscheinungen, besonders in Verbindung mit dem systemischen Ansatz, nicht nur eine Ergänzung der qualitativen Analyse ist, sondern etwas Umfangreicheres, da man auf diesem Wege zu einer tieferen Erkenntnis des linguistischen Objekts und seiner qualitativen Gestaltung kommen kann.

1.2. Linguistische Grundlagen der Untersuchung

In diesem Abschnitt werden einige theoretisch-linguistische Aspekte der quantitativ-systemischen Analyse betrachtet und konkretisiert, die mit der Abgrenzung von Sprache und Rede und mit der Gliederung des Systems der Sprechtätigkeit verbunden sind.

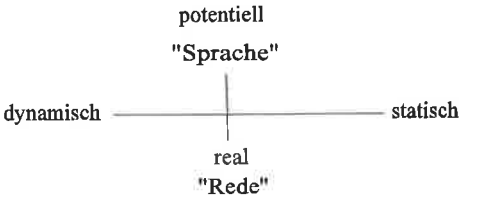
Sprache und Rede

Wir gehen von der allgemeinen Annahme aus, daß Sprache im Gegensatz zur Rede "unterschiedliche Interpretationen eines und desselben materiellen Objekts sind" (Leont'ev 1974:44). Dieses Objekt kann man als "allgemeines Sprachsystem" oder "Sprechtätigkeit" bezeichnen, wobei man im letzteren Fall den aktiven, tätigkeitsbezogenen Charakter der Sprachverwendung als das Hauptmittel für Erkenntnis und Kommunikation der menschlichen Gesellschaft hervorhebt. Sprechtätigkeit im weiten Sinne, die man zwecks terminologischer Abgrenzung als "das System der gesprochenen Kommunikation" bezeichnen kann, umfaßt außer dem Prozeß der Sprachproduktion auch den Prozeß der Rezeption (Empfang der Nachricht), die wir an dieser Stelle nicht untersuchen werden.

Die Möglichkeit und Notwendigkeit der Abgrenzung der beiden Seiten der Sprechtätigkeit ("Sprache" und "Rede") beruht auf der offensichtlichen Tatsache, daß man in dieser Tätigkeit zwei miteinander verbundene, jedoch evident abtrennbare Komponenten sehen kann: das Mittel (Instrument) und seine Verwendung. Allerdings wird die Präzisierung des Inhalts und der Bedeutung dieser Komponenten in der linguistischen Theorie und Praxis unterschiedlich gehandhabt. Üblicherweise definiert man paarweise assoziierte Merkmale, wie z.B. sozial vs. individuell, ideell vs. materiell, allgemein vs. spezifisch, abstrakt vs. konkret, usw.,

die einzeln oder in Kombination miteinander der Definition und Interpretation der Begriffe Sprache und Rede dienen. Es ist charakteristisch, daß die erwähnten Merkmalspaare sich eng berühren und einander oft ergänzen. In einigen Fällen können sie sich überkreuzen. Auswahl und Kombination dieser Merkmale sowie die Abgrenzung der Sprach- und Redesphären werden in Abhängigkeit vom theoretisch-methodologischen Ansatz und von der Besonderheit der konkreten wissenschaftlichen Untersuchung durchgeführt.

Unter Berücksichtigung der Spezifizität der quantitativ-systemischen Untersuchung der Sprache allgemein und der Lexik speziell kann man den Komplex "Sprache - Rede" in Form sich kreuzender Hauptachsen darstellen: Die Achse mit der Opposition von potentiell - real und die Achse mit der Opposition von dynamisch - statisch:



Die Opposition von *potentiell* und *real* betrachtet man in diesem Fall als eine genuine Sprache-Rede-Beziehung, d.h. "Sprache" betrachtet man als ein System potentieller Möglichkeiten, und "Rede" als die Aktualisierung dieser Möglichkeiten. In der Beziehung potentiell vs. real ist offensichtlich ein hierarchisches Moment in Form einer Folge oder Ordnung enthalten, und daher scheint es gerechtfertigt, diese Beziehung als eine Ordnungsrelation zu betrachten. Wir vereinbaren, die Potenzialität als eine höhere "Sprachebene" und die Realisierung als eine niedrigere "Sprechebene" zu betrachten. Es muß bemerkt werden, daß eine derartige Abgrenzung im logischen Sinne der ontologischen Beziehung zwischen Priorität und Sekundarität nicht existiert. Man vergleiche dazu allgemein bekannte Aussagen darüber, daß "historisch die Rede vor der Sprache existiert" (de Saussure 1977) und daß "die Sprache sich eben in der Rede, die durch Sätze realisiert wird, formiert und gestaltet". (Benveniste 1974:140). Rede als die konkrete Realisierung der Sprache ist das einzige direkt beobachtbare Objekt der Linguistik. Die Schlußfolgerungen über die Ebene der Sprache werden aufgrund von Verallgemeinerungen der Resultate der Redeanalyse gemacht. In der linguistischen Praxis ist jedoch auch der "synthetische Ansatz" möglich, nämlich dann, wenn man die Erschließung des Sprachsystems über die Rede und die Untersuchung der Rede über das Sprachsystem durchführt (Vannikov 1979:17-18).

In der quantitativen Linguistik und speziell bei der quantitativ-systemischen

Untersuchung der Lexik hat die Opposition von Potenz und Realisierung in vielen Hinsichten einen direkten praktischen Sinn.

Mit der Potentialität und der Realisierung (Möglichkeit und Realität) ist vor allen Dingen die Idee des "vollständigen Systems" verknüpft, d.h. einer vollständigen Gruppe von Ereignissen, die unter den gegebenen Bedingungen vorkommen können, im Unterschied zu einer begrenzten Stichprobe realisierter Ereignisse. Eine quantitativ ausgedrückte Beziehung zwischen Potenz und Realisierung kann eine heuristische Bedeutung haben und z.B. als ein sinnvolles typologisches Kriterium verwendet werden. Einige Forscher verbinden die Potentialität und Realisierung im obigen Sinne mit der Beziehung zwischen einer statistischen Grundgesamtheit und der Stichprobe aus dieser Gesamtheit, wobei sie die Grundgesamtheit der Sprache und die Stichprobe der Rede zuschreiben. Bei solch einer Interpretation kann man Sprache und Rede mit den Kategorien des Allgemeinen und des Speziellen assoziieren. Die Opposition von Potenz und Realisierung benutzt man auch in informationstheoretischen Untersuchungen der Sprache, speziell bei der Ermittlung der Redundanz des Systems.

Weiter kann man die Realisierung als Aktualisierung oder Wahl einer Variante aus allen in der gegebenen Situation möglichen betrachten. Beispielsweise die Wahl eines geeigneten Wortes aus einem gegebenen semantischen Feld oder die Aktualisierung einer der virtuellen Bedeutungen des Wortes im Text.

Schließlich sind mit den Begriffen Potenz und Realisierung als Charakteristika von Sprache und Rede die Begriffe Wahrscheinlichkeit und Häufigkeit verbunden. "Sprache ist ein probabilistisches, Rede ein Frequenzphänomen" (Golovin 1968: 39). Vor diesem Hintergrund unterscheidet man beispielsweise Sprach- und Redestile: Im ersten Fall berücksichtigt man Stilwahrscheinlichkeiten (als Sprachgesetz) und im zweiten die Vorkommenshäufigkeiten der Einheiten in einzelnen Texten der gegebenen Sprache. Wenn wir Wahrscheinlichkeit als Potentialität auffassen, dann sondern wir in ihr das "Notwendige" aus, nämlich die Tendenz der Häufigkeiten (Häufigkeiten als "Zufälliges"), sich um einen stabilen, systemisch notwendigen Wert zu gruppieren. Daraus folgt, daß man das "Potentielle, Sprachliche" und das "Realisierte, zur Rede Gehörende" unter bestimmten Bedingungen den Kategorien der Notwendigkeit und Zufälligkeit zuordnen kann.

Auf diese Weise sieht man, daß die Opposition von Potentialität und Realisierung als Korrelate der Sprache (Instrument) bzw. Rede (Anwendung) gleichzeitig mit philosophischen Kategorien wie *Möglichkeit* vs. *Tatsache, allgemein* vs. *spezifisch, Notwendigkeit* vs. *Zufälligkeit* in Wechselbeziehung gebracht werden kann. Alle diese Kategorien, die sich teilweise überschneiden, charakterisieren Sprache und Rede objektiv und erlauben es, die Gegenüberstellung und Wechselbeziehung der beiden grundlegenden Aspekte sprachlicher Tätigkeit tiefer zu durchdenken.

Was die de Saussursche Opposition vom Sozialen (Gemeinschaftlichen) als

Merkmal der Sprache und des Individuellen als Merkmal der Rede betrifft, so ist hier auch ein anderer Ansatz möglich, bei dem man diese Merkmale sowohl der Sprache als auch der Rede zuordnet. Sprache als Potenz existiert in "allgemeiner, gemeinschaftlicher" Form, als Invariante für alle Sprecher der gegebenen Sprache. Vom Gesichtpunkt der quantitativen Linguistik aus gesehen wird Sprache in diesem Fall durch Durchschnittsmaße charakterisiert. Gleichzeitig hat jedes Individuum seine eigene individuelle Sprache, ein "inneres System" im Gegensatz zu dem "äußeren System oder der Sprache der anderen" (Coseriu 1963:300); in der quantitativen Linguistik registriert man entsprechend die variablen Eigentümlichkeiten der individuellen Sprache, des Idiolekts. Auf diese Weise kann man Sprache aus der Menge individueller Akte (Corpora) oder aufgrund individueller Kreativität erforschen.

In unserem Modell gibt es neben der Potenz-Realisierungsachse noch eine andere Achse, die sich mit der ersten kreuzt und die Ebenen der Sprache und der Rede (Potenz und Realisierung) in Subebenen oder Sphären der *Dynamik* und *Statik* aufteilt. Die Charakteristika der Dynamik und der Statik werden oft bei der Unterscheidung der Rede als Prozeß und der Rede als Resultat dieses Prozesses verwendet. Die Ebene der Sprache betrachtet man üblicherweise rein statisch, als "Inventar sprachlicher Mittel und die Menge der Regeln". Nur in der Psycholinguistik betrachtet man speziell auch den erzeugenden Mechanismus, der zur Sprachebene gehört ("Kompetenz" in Chomskys Auffassung, "Sprachfähigkeit" in Leont'evs Auffassung). Den Unterschied zwischen den Sphären Dynamik und Statik sieht man auch darin, daß Dynamik (Mechanismus und Prozeß) mit der Tätigkeit des Gehirns zusammenhängt, während Statik (Sprache als "Objekt") sich außerhalb des Menschen befindet.

Ausgehend von unserer Vorstellung der Überkreuzung dieser beiden Achsen und dadurch auch ihrer Vereinigung in der Sprechtätigkeit scheint es zweckmäßig, den dynamischen (erzeugenden, prozessualen) und den statischen (resultativen, inventarisierenden) Aspekt sowohl auf der Ebene der Sprache als auch auf der der Rede zu unterscheiden. In dem System der Sprechtätigkeit kann man folglich vier Grundaspekte (Subsysteme) unterscheiden: auf der sprachlichen Ebene die "sprachliche Kompetenz" und das "Sprachschema", auf der Ebene der Rede den "Sprechprozeß" (Akt) und das "Redeprodukt" (Text).

Mit anderen Worten, die vier genannten Aspekte oder Subsysteme sind das Resultat der Überlagerung der assoziierten Aspekte Potenz - Realisierung und Dynamik - Statik. Durch ihre Kreuzung entsteht eine Struktur, die einer 2x2-Kontingenztafel der Assoziation alternativer Merkmale entspricht. Die Beziehungen zwischen den einzelnen Aspekten zeigt Tabelle 1.

Alle genannten Aspekte der Sprechtätigkeit sind eng miteinander verbunden, aber für spezifische Forschungsziele kann man sie getrennt untersuchen. Aufgrund der Unterscheidung von Dynamik und Statik kann man in der quantitativen Linguis-

tik zwei grundlegende Forschungsbereiche isolieren: Arbeiten, die mit der Erforschung der Redegenerierungsprozesse verbunden sind, wenn man z.B. den Text als einen stochastischen Prozeß betrachtet, einerseits und die bisher zahlenmäßig überwiegenden Arbeiten über die Charakteristika des Sprachmaterials (von Texten) in der Statik andererseits. Dementsprechend steht die quantitative Linguistik beim Übergang von der Redeebene zu der theoretischen Sprach-Ebene vor zwei Typen

Tabelle 1
Aspekte der Sprechtätigkeit

Merkmale	Dynamik	Statik
Potentialität ("Sprache")	Sprachkompetenz	Sprachschema
Realisierung ("Rede")	Sprechprozeß	Sprechprodukt

von Aufgaben: Einerseits die Beschreibung der Redeerzeugung auf der Grundlage der Sprachkompetenz und andererseits die Abstraktion des "Sprachschemas" aus dem Textmaterial, d.h. der statisch-sprachlichen Gesetzmäßigkeiten der Texterzeugung. Diese beiden Aufgaben kann und muß man zusammenführen, um eine vollständige Vorstellung und Klärung der quantitativen Seiten der Rede-Sprachtätigkeit zu erhalten.

Unten werden wir die genannten Aspekte (Subsysteme) der Sprechtätigkeit einzeln besprechen. Wir beginnen mit dem Sprachschema, das der "traditionellen Auffassung des Sprachsystems" in der Statik am nächsten kommt (beim systemischen Ansatz muß man auch alle anderen Teile des allgemeinen Obersystems als Systeme betrachten, wenn man sie separat untersucht).

Das Sprachschema

Unter *Sprachschema* verstehen wir das System der Sprachelemente und der Relationen zwischen ihnen. Die Relationen können sowohl zur Paradigmatik als auch zur Syntagmatik gehören (Bogdanov 1973). Sie charakterisieren die Grammatik im weitesten Sinne (als die Menge der Operationsregeln der Elemente). Das Sprachschema (S) kann man als eine Struktur

(1) $S = \langle M; R \rangle$

veranschaulichen, wobei M die Menge der Elemente (Inventar) und R die Menge

der systemischen Relationen (Grammatik) darstellt. Das Sprachschema stellt an sich ein taxonomisches Phänomen dar, das geordnete Objekte enthält, z.B. das Lexikon und andere Inventare lexikalischer Einheiten (lexikalisch-semantischer Gruppen usw.) zusammen mit ihren formalen, semantischen, Valenz-bezogenen u.a. Beziehungen und entsprechenden quantitativen Charakteristika.

Das Sprachschema kann man als das statistische System der gegebenen Sprache als Ganzer betrachten, d.h. als die Grundgesamtheit linguistischer Elemente und Beziehungen zwischen ihnen. In der Realität existiert jedoch jede natürliche Sprache in vielen Variationen, in Form von gesonderten Subsystemen, die zu verschiedenen Subsprachen oder Funktionalstilen gehören. Eine Subsprache definiert man als "die Menge der Sprachelemente und ihrer Relationen in Texten mit gleichartiger Thematik" (Andreev 1967:23), d.h. der Begriff der Subsprache ist mit einer bestimmten Sphäre der Wirklichkeit (Subsprache der Publizistik, der Wissenschaft und Technik, Geschäftsbriefe usw.) assoziiert. Funktionalstile definiert man als "Sprachvarianten, die durch Unterschiede in Kommunikationsbereichen und Grundfunktionen der Sprache (Darstellung, Ausdruck, Appell) bedingt sind" (Vinogradov 1967:5-6). Aufgrund dieser Definitionen muß man feststellen, daß die Begriffe der Subsprache und des Funktionalstils sich nicht decken. In der Definition des Funktionalstils wird der stilistisch-funktionale Aspekt der Sprache (Darstellung, Ausdruck und Appell) betont, während die Subsprache nur vom Gesichtspunkt der Darstellung aus betrachtet wird. Eine direkte Korrelation zwischen Subsprache und Funktionalstil kann nur dann entstehen, wenn eine Subsprache ausschließlich durch einen für sie eigenen Stil charakterisiert wird. Die Frage nach der Beziehung zwischen Subsprache und Funktionalstil ist kompliziert und kann nur mit Hilfe von konkreten Untersuchungen an umfangreichem Material beantwortet werden. Auf jeden Fall können sowohl thematische Subsprachen als auch Funktionalstile für Forschungszwecke als mit eigenen spezifischen (darunter quantitativen) Besonderheiten ausgestattete sprachliche Subsysteme dargestellt werden. Man kann sich sogar vorstellen, daß man die verallgemeinerte Bezeichnung "Subsprache" unterschiedlichen sprachlichen Subsystemen gibt, die aufgrund des allgemeinen Kriteriums "Darstellungsbereich" abgegrenzt wurden. Vom Gesichtspunkt der quantitativ-systemischen Untersuchung der Lexik aus ist es wichtig festzuhalten, daß Subsprachen sich voneinander vor allem durch die "probabilistischen Spektren ihrer Lexik" unterscheiden (Andreev 1967:23).

Folglich erscheint das Sprachschema als ein komplexes, vielschichtiges Phänomen, das in Abhängigkeit von der Darstellungssphäre in bestimmtem Maße seine Struktur ändert, d.h. in verschiedenen Varianten auftritt. Gleichzeitig zeichnet sich das Sprachschema durch seine Beständigkeit im Rahmen einer Sprache aus. In ihm zeichnet sich beispielsweise die allgemeine, allen Subsprachen gemeinsame Lexik ab, und es weist stabile Zusammenhänge und Gesetzmäßigkeiten auf der gemeinsprachlichen Ebene auf. Den Sprachgebrauch, der sich in gemeinschaftlicher Praxis

herauskristallisiert hat, allgemein akzeptiert wird, sich regulär in der bestimmten Kommunikationssphäre wiederholt und sich als stabiler "Kern" des Subsystems im Sprachschema erkennen läßt, bezeichnet man als die *Norm* der gegebenen Sprache oder Subsprache (bzw. des Funktionalstils). In der quantitativen Linguistik kann man die Norm der Sprache als die wahrscheinlichste Zusammensetzung und die wahrscheinlichsten Zusammenhänge zwischen den Elementen sowie dasjenige Konfidenzintervall, das bei der Realisierung eines entsprechende Textkorpus am wahrscheinlichsten ist, bezeichnen.[3] Die Norm als statisches Phänomen auf der Ebene des Sprachschemas spielt die Rolle des Reglers in der dynamischen Komponente der Sprachkompetenz (s. unten).

Sprachkompetenz

Die Sprachkompetenz ist die Folge der Abbildung des Sprachschemas im menschlichen Bewußtsein, d.h. sie ist die Gesamtheit der Elemente und der systemischen Beziehungen zwischen ihnen, die mit den Elementen und Beziehungen zwischen ihnen im Sprachschema übereinstimmen (auf der Abbildungsebene) *plus* eine besondere dynamische Komponente, die für die Realisierung der Sprache nötig ist. Das Modell der Sprachkompetenz (SK) kann man durch die folgende Struktur darstellen:

(2) $SK = \langle M'; R'; G \rangle$,

wobei M' und R' die Abbildung der Gesamtheit der Elemente bzw. der Beziehungen des Sprachschemas im menschlichen Bewußtsein sind und G die dynamische Komponente (der "Generator") ist. Die Komponente G kann man sich als einen Erzeugungsmechanismus im weitesten Sinne vorstellen, als einen Komplex von nicht nur sprachlichem (einschließlich der Kenntnis der "Norm"), sondern auch vom sozial-pragmatischem Wissen. Dieser Komplex beruht teilweise auf der sozial-pragmatischen Erfahrung der Menschen und teilweise auf phylogenetischen Ursprüngen der besonderen Konstruktion des menschlichen Gehirns. Hier berühren wir den Grenzbereich zwischen Linguistik, Psycholinguistik und Psychophysiologie. Es zeigt sich, daß die Linguistik ohne Rückgriff auf benachbarte Wissenschaften nicht imstande ist, die wahre Natur und die inneren Gesetzmäßigkeiten der Sprache im Ganzen zu erfassen. Ebenso können die quantitativen Gesetzmäßigkeiten der Rede (die universellen Besonderheiten der statistischen Struktur des Textes,

[3] Parallel zur Norm definiert man manchmal den Usus, womit man die unbewußte und nicht kodifizierte Norm bezeichnet. Hier werden sie nicht unterschieden. Unter dem Terminus "Norm" werden wir beide Begriffe zusammenfassen.

die Existenz von stabilen Verteilungen, die regulären/gesetzmäßigen Korrelationen mit der Umgebung usw.) nur unter Zuhilfenahme extralinguistischer Vorstellungen und Kriterien erfaßt werden.

Der Sprechprozeß

Der Sprechprozeß oder Sprechakt ist die Realisierung der Sprachpotentialität, d.h. die eigentliche Funktion des Mechanismus der Redeerzeugung. Die Beziehung des Sprechprozesses zur Sprachkompetenz kann man auch als die Beziehung des geregelten Subsystems zum regelnden betrachten. Der Sprechprozeß liefert unmittelbar eine lineare Sequenz von Sprecheinheiten[4], aber der Prozeß der Äußerungserzeugung selbst stellt einen komplizierten, vielschichtigen Sprech-Denk-Akt dar. Er besteht aus den Phasen der Transformation der Intention über innere Sprache in das Schema der Sprechäußerung und ihre Umformung in den phonetischen, lexikalisch-semantischen und logisch-grammatischen Kode der Sprache (vgl. Leont'ev 1969; Luria 1979).

Das neue Moment, das in das System der Sprechtätigkeit von *außen* hineingetragen wird - das Objekt, über das sich die Tätigkeit vollzieht - ist die Intention, hervorgerufen durch ein bestimmtes Bedürfnis und modifiziert durch das Motiv und das Ziel.[5] Bezeichnen wir diese modifizierte Absicht als *Thema*. Die zweite Komponente außersystemischen (extralinguistischen) Ursprungs ist die *Situation*, d.h. die Bedingungen, unter denen die sprachliche Mitteilung geschieht. Die Situation schließt die Sphäre der Mitteilung sowie begleitende zufällige Faktoren (z.B. Störung), Kontext im engen Sinne und Rückkopplung ein. Das Thema und die Situation kann man als entscheidende Komponenten des Sprechprozesses in dem Maße betrachten, wie sie in widergespiegelter Form im menschlichen Bewußtsein (oder Unterbewußtsein) als Elemente des Sprech-Denk-Prozesses erscheinen. Hier zeigt sich besonders klar die enge Verbindung des Sprechprozesses mit dem Denken und die Wechselwirkung mit der Umgebung, d.h. mit der umgebenden Welt.

[4] In Abhängigkeit von den jeweiligen Forschungaufgaben in der quantitativen Linguistik kann man als Sprecheinheiten die Buchstaben, Wörter, Phrasen usw. betrachten. Auf der lexikalischen Ebene ist die Grundeinheit das Wort, das der kleinsten bedeutungstragenden Einheit in der Zusammensetzung der Äußerung entspricht.

[5] Das Ziel als Vorwegnahme des Resultats der Tätigkeit wird aufgrund des Motivs formuliert (einschließlich Identifikation), hinter dem ein Bedürfnis steht (innerer oder äußerer Stimulus zur Tätigkeit). Man kann allgemeine und momentane Stimuli unterscheiden, z.B. kann dem momentanen Ziel eine Mitteilung (Äußerung einer Idee) entsprechen, während dem allgemeinen Ziel die gesamte aus den einzelnen Äußerungen bestehende Mitteilung (Äußerung eines ganzen Komplexes von Ideen) entspricht.

Den Sprechprozeß (SP) kann man allgemein in Form der Struktur

(3) $SP = \langle T; S; G' \rangle$

darstellen, wobei T das Thema (Absicht, Motiv, Ziel), S die Situation (enschließlich der Sphäre der Mitteilung, Kontext, Rückkopplung und begleitende Faktoren), G' den sprachlichen Mechanismus in Tätigkeit darstellen.

Der gesamte Prozeß der Sprecherzeugung (Erzeugung der Äußerungen) stellt eine Folge von unterschiedlichen Sprechakten oder Tätigkeiten dar. Sprecherzeugung ist im Grunde ein zyklischer Sprech-Denk-Prozeß, der in groben Zügen an das funktionale Verhaltenssystem von Anochin (1962) erinnert. Den Stellenwert dieses Prozesses im Gesamtsystem der Sprechtätigkeit und seine Verbundenheit mit anderen Teilen des Gesamtsystems kann man mit dem Schema in Abb. 1 illustrieren. Betrachtet man den einzelnen Akt der Äußerungserzeugung als Sprech-Denk-Tätigkeit, so wird klar, daß dieser Akt (der eine SP-Struktur im Mikroformat hat) aus einer Folge von Entscheidungen besteht, wobei in einzelnen Stadien eine

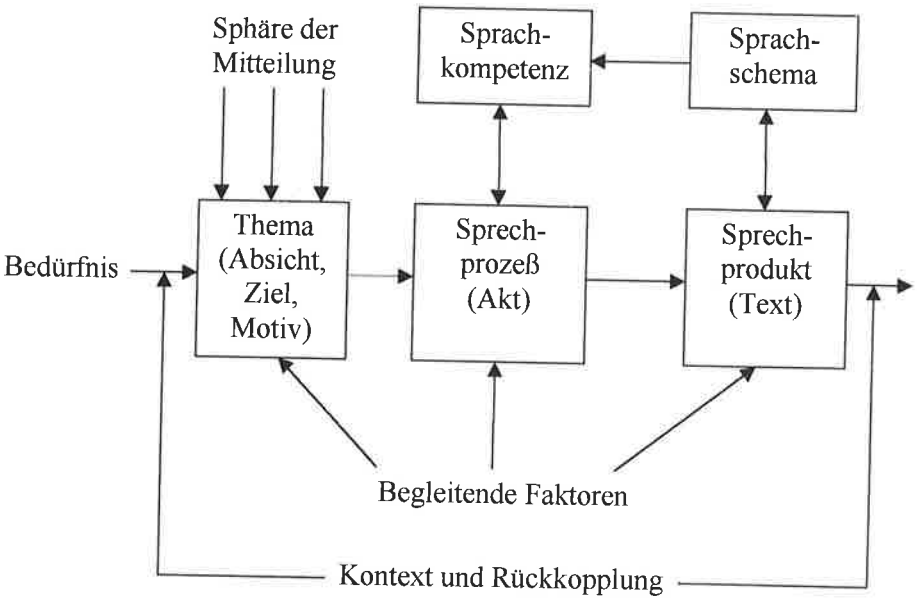


Abbildung 1

Wahl zwischen verschiedenen Varianten getroffen werden muß. Die Wahl der Variante (z.B. des Wortes in der Schlußphase der Sprech-Denktätigkeit) kann in einigen Fällen völlig von der gegebenen Situation und den Möglichkeiten (Zuständen) der Sprachkompetenz determiniert sein, in anderen Fällen agiert der Sprecher unter den Bedingungen "unvollständiger Determiniertheit", wobei nur der Bereich (das "Feld"), in dem man die entsprechende Variante suchen muß, bekannt ist (beispielsweise bei der Wahl des Wortes aus einer Menge von Synonymen). In solchen Fällen stützt sich der Sprecher auf die probabilistische Erwartung (Frumkina 1969; vgl. auch den Begriff der probabilistischen Schätzung bei Nalimov 1979), und in den meisten Fällen wählt er eine für die gegebene Situation typische Variante, wobei Schwankungen in bestimmten Grenzen möglich sind. Die Wahrscheinlichkeit der Wahl einer bestimmten Variante (z.B. eines Wortes) wird in diesem System durch einen Komplex aus den erwähnten Faktoren bedingt (Thema, Situation, Zustand der Sprachkompetenz). Folglich wird die gesamte Sprecherzeugung als Resultat einer *komplexen Wechselwirkung von determinierten und probabilistischen (zufälligen) Faktoren* verwirklicht. Insgesamt kann man den gesamten Prozeß als ein probabilistisches System betrachten, das sich durch eine bestimmte Stabilität und Regularität in der Masse der Zufallsereignisse auszeichnet.

Bei der Lösung konkreter Aufgaben bei der mathematischen Modellierung des Prozesses der Redeerzeugung benutzt man üblicherweise komplizierte Modelle des Markov-Typs (z.B. Lounsbury 1965). In das probabilistische Schema der Textproduktion (und Texterkennung) kann man noch ergänzende situative Momente einfügen (Piotrowski 1975:56; Hoffmann, Piotrowski 1979:40-41). In einigen Modellen betrachtet man den Erzeugungsprozeß als das Resultat der Überlagerung von Zufalls- und deterministischen Prozessen (Gačečiladze, Cilosani 1971), oder als Einbettung von Zufallsbeziehungen in die deterministische Information, wobei wir vermerken, daß das Zusammenwirken zufälliger und determinierter Zusammenhänge dem System erlaubt, sich zu entfalten und zu optimieren (Zubov 1980).

Das erhöhte Interesse an mathematischen, insbesondere probabilistisch-statistischen Modellen der Sprechproduktion (und -Erkennung) ist heutzutage bedingt durch die praktischen Bedürfnisse in Gebieten wie der automatischen Textverarbeitung und muß im Zusammenhang mit der Lösung einiger Probleme der linguistischen Absicherung von KI-Systemen gesehen werden. Die Aufmerksamkeit der Forscher konzentriert sich gleichzeitig auf die Suche nach der theoretischen Klärung quantitativer Eigenschaften der Sprechproduktion. In diesem Zusammenhang wurde die Meinung geäußert, daß eine adäquate mathematische Modellierung des Sprechprozesses sich auf die Theorie der sog. optimalen Prozesse stützen kann, in Anbetracht dessen, daß (nach Bernstein) "das Wesen der Sprechtätigkeit als Tätigkeit überhaupt auf Optimierung unter gegebenen Randbedingungen beruht" (Leont'ev 1974:80).

Produkt der Sprechtätigkeit

Das Produkt des Redegenerierungsprozesses ist das Sprechprodukt oder der *Text*, der in der allgemeinsten Form als ein auf bestimmte Art fixiertes Segment des Sprechkontinuums verstanden wird. Auf der Realisierungsebene steht der Text in direkter Wechselbeziehung zu der Sprachebene als die Abbildung des Sprachschemas in Form eines besonderen Systems: <M"; R">, wo *M"* und *R"* Teilmengen der Elemente bzw. Beziehungen zwischen ihnen sind (M" ⊂ M, R" ⊂ R). In seiner Beziehung zum Sprechprozeß wird der Text so charakterisiert, daß von ihm die prozedurale Form der Sprech-Denk-Tätigkeit abstrahiert wird, auch wenn er die Beziehung zu dieser Tätigkeit aufrechterhält. Da der Text das einzige direkt beobachtbare Objekt der linguistischen Analyse ist, kann er bei entsprechendem Ansatz auch als Modell für die Erforschung der dynamischen Aspekte der Rede (d.h. des Redeprozesses) dienen. Den Textbegriff kann man in dem Sinne erweitern, daß man in seine Beschreibung die von ihm widergespiegelten Komponenten des Sprechprozesses, nämlich Thema und Situation, einschließt. Man kann sogar noch weiter gehen und bei der Analyse des Textes versuchen, die Widerspiegelung der Sprachkompetenz, die sich in Form von bestimmten stilistischen und grammatischen Merkmalen äußert, zu beleuchten. Folglich kann man das gesamte System des Textes in Form des folgenden allgemeinen Modells darstellen:

(4) $Text = \langle M''; R''; T'; S'; G'' \rangle$

wo T' und S' Thema und Situation und G'' die Sprachkompetenz darstellen.

Da wir die Redeerzeugung (Texterzeugung) als einen komplexen stochastischen Prozeß (der sich aus der Wechselwirkung von zufälligen und deterministischen Faktoren zusammensetzt) verstehen, kann man schließen, daß auch das Ergebnis eines derartigen Prozesses probabilistisch charakterisiert werden kann. Den Text und das dazugehörige Vokabular kann man im allgemeinen als probabilistische Systeme betrachten. In der Empirie äußert sich dies auf der einen Seite durch die Existenz von stabilen Verteilungen, durch stetige intra- und intersystemische Korrelationen, durch die Bildung eines "Kerns" in geschlossenen Gruppen usw. und, auf der anderen Seite, durch periphere Erscheinungen (infolge von unscharfen Grenzen) und verschiedenartige zufällige Fluktuationen.

Aufgrund der Etablierung der sich überschneidenden Achsen potentiell-real und dynamisch-statisch kann man das gesamte System der Sprechtätigkeit in getrennte Teile aufteilen, die man im Funktionsplan des Systems als grundlegende Subsysteme betrachtet. Gleichzeitig etabliert sich die innere Struktur, die Wechselbeziehung zwischen den einzelnen Subsystemen. Das ganze System der Sprechtätigkeit kann man anschaulich in Form eines "Aggregatmodells" darstellen, das aus vier Teilen besteht:

(1) $Sprachschema = \langle M; R \rangle$
(2) $Sprachkompetenz = \langle M'; R'; G \rangle$
(3) $Sprechtätigkeit = \langle T; S; G' \rangle$
(4) $Text = \langle M''; R''; T'; S'; G'' \rangle$.

Hier sieht man folgende strukturell-funktionalen Systemkomponenten, die in verschiedenen Kombinationen in primärer oder sekundärer Form in der Zusammensetzung der Subsysteme auftauchen: *M* - Menge der Sprachelemente, *R* - Menge der Beziehungen zwischen den Elementen, *G* - erzeugender Mechanismus, *T* - Thema, *S* - Situation.

Zum Abschluß der Analyse des Systems der Sprechtätigkeit kehren wir nochmals zu dem ursprünglichen Modell des Sprachschemas zurück. Im entsprechenden Abschnitt haben wir erwähnt, daß das Sprachschema in der Realität in verschiedenen Varianten erscheint, die man als Subsprachen (oder Funktionalstile) bezeichnen kann. Nachdem wir die Verlaufsbedingungen des Sprechprozesses und die Rolle der Komponenten *T* (Thema) und *S* (Situation) analysiert haben, wurde klar, daß diese Komponenten auch mit dem Sprachschema verbunden sind und aufgrund der komplexen Wechselwirkung zwischen Rede- und Sprachebene bei der Aufteilung des Sprachschemas in thematische und situative Varianten mitwirken. Das zugrunde liegende Modell des Sprachschemas (s. Formel 1) kann man also präzisieren, indem man die Faktoren *T"* und *S"* hinzufügt:

(1a) $Sprachschema = \langle M; R; T''; S'' \rangle$,

wo *T"* und *S"* als Repräsentationen von typischen Themen und Mitteilungssphären bei der Bildung der Varianten des Sprachschemas (Subsprachen oder Funktionalstile) betrachtet werden. Auf dieser Grundlage kann man beispielsweise im Sprachschema stilistische Schichten des Lexikons unterscheiden.

Das Modell der Sprechtätigkeit stellt zweifellos Tätigkeiten aus einer bestimmten Sicht vereinfacht dar. Eine Weiterentwicklung des Modells ist mit der detaillierten Ausarbeitung bestimmter Seiten (in Abhängigkeit von den Forschungszielen) verbunden, sowie mit der Vertiefung der Analyse der inter- und intrasystemischen Beziehungen.

Das Modellieren des allgemeinen Sprachsystems und die Erforschung der Eigenschaften und Wechselbeziehungen der einzelnen Teile und Komponenten des Systems ermöglichen uns, den Gegenstand der quantitativen Linguistik deutlich abzugrenzen und ihre Aufgaben zu formulieren. Gleichzeitig steht aber das gesamte System als ein ganzheitliches Gebilde vor dem Forscher. Eine konkrete Untersuchung führt man aufgrund dieser Aufteilung durch. Man kann beispielsweise die quantitativ-systemische Untersuchung der Lexik auf dem Gebiet der Dynamik und Statik oder im Bereich ihrer Wechselwirkungen durchführen. Den Text kann man

als Folge von grammatisch und semantisch verknüpften lexikalischen Einheiten (auf der <*M''*; *R''*>-Ebene) oder als zusammenhängendes Ganzes mit Thema und stilistischen oder pragmatischen Besonderheiten (wenn man die Komponenten *T'*, *S'*, *G''* ergänzt) untersuchen. Die Feststellung der quantitativen Eigenschaften der Texte und ihrer Vokabulare, die Etablierung der systemischen Beziehungen zwischen Wörtern in Paradigmatik und Syntagmatik, die Erforschung der Umstände der Texterzeugung usw. müßten zu einer darauffolgenden Verallgemeinerung, Ordnung und zum Verständnis des empirischen Materials auf einer höheren, theoretischen Ebene führen. Letzten Endes muß man den synthetischen, integralen Ansatz zur Erforschung und die Aufdeckung der quantitativen Eigenschaften des Sprechtätigkeitssystems im Ganzen und in kontinuierlichem Zusammenwirken von quantitativer Analyse mit qualitativer Interpretation verwirklichen.

1.3. Die Untersuchungsmethode

In diesem Abschnitt werden Fragen erörtert, die die Untersuchungsmethode betreffen: Quantifizierung des Materials, Einheiten und Ebenen der Analyse, Methoden der Beschreibung des lexikalischen Materials in Form von lexikalischen Gruppen und die Modellierung mit Hilfe von Verteilungen; besondere Aufmerksamkeit wird dabei der Interpretation linguistischer Verteilungen gewidmet.

Status der Methodik

Unter Untersuchungsmethodik verstehen wir eine geordnete Gesamtheit oder ein *System* unterschiedlicher Methoden (konkreter Lösungswege einer Aufgabe) und auch deren Anwendung. Vom Standpunkt der wissenschaftlichen Methodologie aus befinden sich konkrete Untersuchungsmethoden auf der Ebene der "Methodik und Technik der Forschung" (Sadovskij 1979). Diese Ebene steht tiefer in der Hierarchie, die auch höhere Ebenen der Methodologie einschließt, nämlich die konkret-wissenschaftliche, die allgemein-wissenschaftliche und die philosophische. Die Zugehörigkeit von Methodik und Technik zur "niedrigeren" Ebene bedeutet jedoch nicht, daß hier die Beziehung zu den höheren Ebenen, in diesem Falle zu den allgemeinen theoretisch-methodologischen Prinzipien der quantitativ-systemischen Analyse (konkret-wissenschaftliche und allgemein wissenschaftliche Ebene) oder zu der Ebene der philosophischen Methodologie unterbrochen würde. Die angewandten Methoden müssen in ihrer Gesamtheit ein System von wissenschaftlich ausgearbeiteten Regeln und Forschungsverfahren darstellen, die als ein organischer Teil in ein allgemeineres System der wissenschaftlichen Methodologie eingeht.

Für die quantitativ-systemische Analyse (sowie jede systemische Analyse) ist

nicht ein spezifischer Apparat konkreter Methoden charakteristisch, sondern ein geordneter, logisch begründeter Ansatz zur Verwendung vorhandener Methoden, die bereits im Rahmen anderer Disziplinen ausgearbeitet wurden (Mathematik, Linguistik u.a.). Wahl und Kombination dieser Methoden stellen eines der wichtigsten Momente sowohl in der Phase der Beobachtung und des Experiments als auch in der Phase der Analyse und der theoretischen Verallgemeinerung der Forschungsresultate dar. Besonders zu unterstreichen ist die Ausarbeitung der Methodik "im interdisziplinären Bereich", darunter in der quantitativen Linguistik, wo die Wahl der Methoden einen wichtigen Schritt auf dem Weg zur Entdeckung neuer Phänomene und Abhängigkeiten in der Linguistik bedeuten kann.

Im Rahmen der Darstellung und Beschreibung der hier verwendeten Methoden der quantitativ-systemischen Analyse der Lexik ist es vor allen Dingen notwendig, die allgemeinen Grundlagen, die mit der Quantifizierung des linguistischen Materials als der empirischen Basis der Untersuchung zusammenhängen, zu erläutern. Bei der Erläuterung der Probleme der Untersuchungsmethodik werden wir nicht die einzelnen Methoden, die in entsprechenden Lehrbüchern und Einzeluntersuchungen (vgl. z.B. Golovin 1971; Bektaev, Piotrovskij 1973-1975; Piotrovskij, Bektaev, Piotrovskaja 1977; Piotrowski, Bektaev, Piotrovskaja 1985; Altmann 1980a; Dugast 1980) hinreichend ausführlich beschrieben wurden, betrachten, sondern wir konzentrieren uns vor allem auf die logischen Grundlagen und die allgemeinen Prinzipien der Anwendung quantitativer Methoden in der Linguistik.

Quantifizierung, Quantierung, Messung

Die Voraussetzung für die Anwendung quantitativer Methoden bei der Erforschung sprachlicher Erscheinungen ist die *Quantifizierung* des untersuchten Materials. Quantifizierung – im weitesten Sinne des Wortes[6] – bedeutet eine quantitative Darstellung qualitativer Eigenschaften, d.h. eine Prozedur, bei der man Erscheinungen, die zunächst als qualitativ aufgefaßt wurden (z.B. linguistische Objekte), quantitative Werte zuschreibt, woraufhin man sie als quantitative Objekte untersuchen kann. In der Regel schließt die Quantifizierung eine vorangehende "Quantierung" des Objekts (die Umformung des Objekts in eine für die Messung geeignete Form, Bestimmung der Meßeinheiten auf verschiedenen Ebenen) und die anschließende Messung ein. Daher kann man im Rahmen der Methodik der quantitativ-linguistischen Untersuchung die Quantifikation im Ganzen als eine abstrakte Transformation (Überführung der Qualität in Quantität) betrachten, die durch Quantierung

[6] Im engeren Sinne (in der Logik) versteht man unter Quantifizierung die exakte Darstellung und Bestimmung des Umfangs des Subjekts und des Prädikats des Schlusses sowie die Anwendung von Operatoren, genannt Quantoren, in logischen Ausdrücken (Kondakov 1971:211).

und Messung erfolgt.

Wie wir schon bemerkten, stellt die *Quantierung* die erste Phase der Quantifizierung dar, die dann notwendig ist, wenn die Zähleinheiten nicht direkt beobachtbar sind oder wenn man sie gemäß der Untersuchung sbedingungen modifizieren muß. So können beispielsweise stetige Variablen durch Quantierung in diskrete Form gebracht werden, d.h. durch Zerlegung ihres Mutungsbereiches auf disjunkte Intervalle. Es ist nur natürlich, daß die Quantierung sprozeduren linguistischer Objekte inhaltlich begründet und eindeutig interpretierbar sein müssen.

Messung kann man als eine Prozedur bezeichnen, die aufgrund vorbestimmter Regeln den betreffenden Objekten numerische Werte zuschreibt. Es läßt sich zeigen, daß die logischen Grundlagen der Messung die Kategorien *Eigenschaft* und *Relation* bilden. Bei der Messung zeigt sich die Beziehung des Objekts (X) aufgrund einer Eigenschaft (P) zu einem Wert (Y), der quantitativ ausgedrückt wird. Diese abstrakte Struktur kann man schematisch wie folgt darstellen:

(1) $X \overset{P}{\rightarrow} Y$

wobei die Komponente "$\rightarrow$ " die "Abbildungsrelation" darstellt[7], bei der sich das Objekt (X) mit dem Wert (Y) aufgrund einer quantitativen Eigenschaft (P) assoziiert, d.h., mit einer Eigenschaft, die eine quantitative Beurteilung zuläßt. Quantitative Eigenschaften sind beispielsweise Umfang, Anzahl, Häufigkeit, Länge usw. Als Beispiel, das das Schema (1) illustriert, kann man die folgende Aussage heranziehen: Ein Text (X) ist mit dem quantitativ ausgedrückten Wert (Y) in dem Sinne verknüpft, daß der Textumfang (P) diesem Wert gleicht. In einer geläufigeren Form kann man diese Aussage auch folgendermaßen ausdrücken:

Der Umfang (P) der Textes (X) ist Y.

Andere Beipiele sind:

Die Häufigkeit (P) des Wortes (X) ist gleich Y;
die Länge (P) des Wortes (X) ist Y.[8]

[7] Vgl. Rakitovs (1977:240) Aussage zum Wesen der Messung: "Messungen sind im Grunde bestimmte Funktionen, die isomorphe oder homomorphe *Abbildungen* [Hervorhebung J.T.] von Elementen, Situationen, Prozessen oder Relationen eines Systems [...] in die Elemente eines anderen - numerischen - Systems realisieren."

[8] Es ist bemerkenswert, daß Aussagen des Typs "der Apfel ist rot" , d.h. "Apfel" (X) ist verknüpft mit "rot" in dem Sinne, daß die "Farbe" (P) des Apfels dem "rot" (Y) gleicht, von der Form her damit identisch sind (vgl. Kohonen 1980:15).

Diese Beispiele zeigen, daß bei der quantitativen Darstellung eines Objekts aufgrund einer Eigenschaft die numerischen Werte nicht dem Objekt selbst zugeschrieben werden, sondern der Eigenschaft, dem *Merkmal* des Objekts. Den Begriff des Merkmals (Parameters, Charakteristik) kann man als die "Materialisierung" des logischen Begriffs der Eigenschaft im Bereich der Beobachtung und des Experiments betrachten. Bei der praktischen Arbeit ist es zweckmäßig, den "Bezeichner des Merkmals" (P) und den "Wert des Merkmals" (Y) als wichtigste Komponenten des betrachteten logischen Schemas (1) zu unterscheiden. Das Meßresultat, d.h. den Wert des Merkmals (Y) kann man als Funktion

(2) $y = f[P(X)]$

darstellen. Der Begriff "Wert des Merkmals" (*Y*) besteht in der Regel aus zwei Elementen: aus einer Quantität (Zahl) und der Bezeichnung der Meßeinheit, die dem Merkmal zugeordnet wird, zum Beispiel: "die Länge des Wortes beträgt *6 Buchstaben*".[9] Man kann auch darauf hinweisen, daß neben dem rein quantitativen (numerischen) Wert auch ein gemischter oder "intermediärer" Merkmalswert existiert (Rubaškin 1976): der qualitativ geschätzte Wert (z.B. sehr wenig, wenig, durchschnittlich, viel, sehr viel) und der binär geschätzte (es gibt - es gibt nicht; viel - wenig u.a.). Diese Schätzungen kann man unter bestimmten Umständen (z.B. in unscharfen Mengen) oder in bestimmten Phasen der Analyse verwenden. Wenn es nötig ist, kann man sie durch numerische Werte ersetzen, wenn beispielsweise die qualitative Schätzung durch Grade, die eine geordnete Skala bilden, ausgedrückt wird, oder wenn man den binären Merkmalen die Werte 0 und 1 zuschreibt (wobei sie eine dichotomische Skala bilden).

Weiter kann man darauf hinweisen, daß es außer den *primären* (direkten) Messungen auch *abgeleitete* (indirekte) Messungen gibt. Messungen sind primär, wenn sie nicht auf vorangehenden Messungen basieren; sonst bezeichnet man sie als abgeleitet (Suppes, Zinnes 1968:25). Als klassiches Beispiel der abgeleiteten Messung gilt die Messung der Geschwindigkeit der Bewegung, die sich aus der Beziehung der Länge der Strecke zu der verstrichenen Zeit berechnet. In der quantitativen Linguistik sind abgeleitete Messungen z.B. der "Diversitätsindex" (das Verhältnis zwischen Vokabularumfang und Textumfang); die "Korrelationsfunktion" als das Verhältnis zwischen bedingter Wahrscheinlichkeit und nicht-bedingter Wahrscheinlichkeit (Andreev 1967:22) und viele andere Funktionen oder Häufigkeitsverhältnisse (z.B. die stilometrischen Koeffizienten von Golovin 1971:140-154).

Analog zu der Unterscheidung in primäre und sekundäre Messungen kann man auch quantitative Eigenschaften in primäre oder *einfache* und sekundäre oder

[9] Für eine neue Auffassung des Problems der Maßeinheiten s. Köhler (1995).

komplexe aufteilen. Letztere zeichnen sich durch ihre komplexe Struktur aus. Wie die Erfahrung zeigt, kommen in der praktischen Arbeit komplexe Eigenschaften (Merkmale) sehr oft vor. Von linguistischen Objekten kann man sagen, daß man sie durch eine Vielzahl von verschiedenen (primären und sekundären) quantitativen Eigenschaften charakterisieren kann, wobei die Aufgabe der systemisch-quantitativen Untersuchung der Lexik darin besteht, die "systemischen Eigenschaften" zu finden und zu erfassen, d.h. solche Eigenschaften, die als Grundlage für die Entdeckung quantitativ-systemischer Gesetzmäßigkeiten in der Lexik dienen.

Unter dem Gesichtspunkt der Theorie und der Praxis der Messung ist es wichtig zu wissen, daß man Messungen auf verschiedenen *Skalen* (Nominal-, Ordinal-, Intervall- und Verhältnissskala) durchführen kann. Jede Skala hat ein entsprechendes Zahlensystem und zugelassene Operationen (Stevens 1960).

Einheiten und Ebenen der Analyse

Bei der quantitativen Untersuchung eines Systems müssen die Einheiten der Analyse, die als sich wiederholende Komponenten (Elemente) des gegebenen Systems auf der gegebenen Untersuchungsebene zählbar sind, unbedingt definiert werden.[10] Die Bedingung der Wiederholbarkeit der Einheiten hängt natürlich mit der Forderung nach Invarianz dieser Einheiten zusammen. Bei der Erforschung linguistischer Objekte bedeutet dies, daß Spracheinheiten bei ihren verschiedenen Vorkommen als gleich identifizierbar sein müssen, damit man von ihrer Wiederholbarkeit sprechen kann.

Wenn man, wie üblich, das *Wort* als die lexikalische Grundeinheit des lexikalischen Systems betrachtet, dann muß man diesen Begriff beim quantitativen Ansatz als Untersuchungseinheit konkretisieren. Man muß zwei Aspekte der Existenz und der Funktion des Wortes in der Sprechtätigkeit unterscheiden: Die Einheit des *Vokabulars* und die Einheit des *Textes*. Damit unterstreicht man den Umstand, daß die Erforschung der Lexik nicht nur die Erforschung des Wortschatzes als strukturierter Gesamtheit lexikalischer Einheiten umfaßt, sondern auch als eines funktionierenden Kommunikationssystems. Es bleibt, die Begriffe *Vokabulars* und *Text* im Hinblick auf ihre quantitativen Eigenschaften zu präzisieren. Im lexikalischen Bereich kann man diese Begriffe folgendermaßen bestimmen.

---

[10] Die Begriffe "Einheit" und "Element" sind miteinander verwandt. Sie unterscheiden sich darin, daß die "Einheit" etwas Invariantes impliziert, das für die Bestimmung der quantitativen, zu messenden Eigenschaften notwendig ist, während der Begriff "Element" (in der systemischen Terminologie) eine Wechselbeziehung zu anderen Elementen impliziert , die das Gesamtsystem ausmachen. Einheiten und Elementen ist gemeinsam die Nichtzerlegbarkeit (Elementarität) auf der gegeben Zerlegungsebene des Systems (vgl. auch Köhler 1995, Altmann 1996).

Unter *Vokabular* versteht man die Gesamtheit der verschiedenen Wörter, die man üblicherweise in Form einer Liste darstellt. Die Vokabulareinheiten ("types") können in den folgenden zwei Arten vorliegen. Im ersten Fall kann man eine Liste der *Wortformen* aufstellen, d.h. Wörter in der Form, wie sie in realen Texten vorkommen. Im zweiten Fall bringen wir die unterschiedlichen Formen auf einen Nenner, üblicherweise unter die sogenante Grundform (bei Substantiven der Nominativ, bei Verben der Infinitiv). Diese Einheiten nennen wir *Lexeme.*

Unter Text verstehen wir im allgemeinen eine lineare Folge bestimmter Spracheinheiten: Wörter, Morpheme usw. Im lexikalischen Bereich stellt man den Text als eine Gesamtheit von Einheiten dar, die man in der Alltagssprache als Wörter und in der quantitativen Linguistik als *Wortverwendungen* ("tokens") bezeichnet. Formal werden sie durch zwei Zwischenräume (oder andere Trenner wie Interpunktionszeichen) im Text begrenzt.

Die gerade erwähnten Termini - Wortform, Lexem, Wortverwendung - werden bei der quantitativen Analyse der Lexik streng auseinandergehalten. In den Fällen, wo die Unterscheidung dieser Termini unwesentlich ist, benutzt man die allgemeine Bezeichnung *Wort.*

Die oben angegebene Definition der Vokabular- und Texteinheiten bezieht sich auf den Fall, in dem das Wort nur aufgrund seiner äußeren Form identifiziert wird oder als Einheit von Form und lexikalischer Bedeutung. Hält man aber an der Invarianz der Einheiten fest (aufgrund derer man sie identifizieren kann und folglich von ihrer Wiederholbarkeit sprechen kann), dann kann man sich auch den anderen Fall vorstellen, in dem die unterschiedlichen Bedeutungen des Wortes das Unterscheidungskriterium bilden; als Zähleinheit betrachtet man dann die gegebene lexikalisch-semantische Variante. Auf diese Weise kann man das Wort in Abhängigkeit vom jeweiligen Untersuchungsziel entweder vom formalen oder vom semantischen Gesichtspunkt aus betrachten.

In einigen Fällen kann man als Lexikoneinheiten bestimmte Wortklassen betrachten, z.B. "einsilbige Wörter, zweisilbige Wörter,...", "Substantive, Verben,...", die man aufgrund phonetischer, morphologischer u.a. Charakteristika identifiziert und zählt. Mit anderen Worten, das Kriterium der Wortidentifikation ist seine Zugehörigkeit zu einer *Klasse von Wörtern*, definiert durch eine (qualitative oder quantitative) Eigenschaft. Solche Eigenschaften sind beispielsweise Wortbildungsmodell, Zugehörigkeit zu einer Wortart, zu einer lexikalisch-semantischen Gruppe usw. Die Eigenschaften, aufgrund derer man das Wort als Analyseeineheit identifizieren kann, sind auf bestimmte Weise mit den unterschiedlichen Sprachbereichen verbunden. Daraus kann man folgern, daß der Bereich der Analyse (Beschreibung) der Lexikonerscheinungen, d.h. der gesamte lexikalische Bereich, in *Teilbereiche* aufgeteilt werden kann, beispielsweise in die phonetisch-lexikalischen und die grammatisch-lexikalischen Teilbereiche sowie den eigentlichen lexikalischen oder "logo-lexikalischen" Bereich. In allen diesen Fällen kann man Ausdrucks- und

Inhaltsseite unterscheiden, auch wenn dieser Unterschied am prägnantesten im eigentlichen lexikalischen Bereich hervortritt (man unterscheidet den formal-lexikalischen und den semantisch-lexikalischen Aspekt). Sogar die stilistische Erforschung der Lexik kann eine Aufgabe der quantitativ-systemischen Analyse sein.

Je nach Ziel der Analyse erscheint das Wort als Einheit mit jedesmal anderen Eigenschaften. Man kann feststellen, daß die Untersuchung der Lexik unter verschiedenen Aspekten (auf unterschiedlichen Unterebenen) eigentlich die Untersuchung der *Subsysteme* der Lexik bedeutet, wobei die Einheit (Element) dieser Subsysteme immer das Wort in allen seinen Erscheinungsformen ist.

Gleichzeitig kann man das Wort als ein "System von Formen und Bedeutungen" betrachten (Vinogradov 1947:15), d.h. als ein Systemobjekt, das aus Elementen besteht und eine bestimmte Struktur hat. In einem solchen Fall betrachtet man das Wort im Hinblick auf seine innere Struktur, wobei die Zähleinheit nicht das Wort selbst, sondern seine Komponenten sind: Phoneme, Morpheme, Silben usw. Die quantitative Untersuchung der inneren Struktur des Wortes wird üblicherweise zur Phono- und Morphostatistik gezählt. Eine derartige Analyse hängt aber eng mit dem lexikalischen Bereich der Sprache in dem Sinne zusammen, daß sie die unvermeidliche Vorphase auf dem Weg zur Untersuchung der Lexik auf phono- und morphologischen Unterebenen darstellt, wobei das Wort als Klassenvertreter die Zähleinheit ist. Eine der Aufgaben der quantitativ-systemischen Analyse der Lexik ist außerdem die Untersuchung der Wechselwirkung zwischen verschiedenen Bereichen der linguistischen Analyse, beispielsweise die Erfassung von phonologischen und lexikalischen Korrelationen oder die Gegenüberstellung der wortbildenden und semantischen Strukturen des Wortes in ihren quantitativen Beziehungen. Die Erforschung von Zusammenhängen zwischen verschiedenen Bereichen ist eines der wichtigsten methodologischen Prinzipien der systemischen Analyse ("die Verbindung des Systems zur Umwelt ") und hat auch Bedeutung für die Lösung einiger allgemeiner theoretisch-linguistischer Aufgaben, beispielsweise in der Sprachtypologie.

Lexikalische Gruppen

Der systemische Charakter der Lexik äußert sich besonders deutlich bei der Aufteilung der Wörter in verschiedene Gruppen (Klassen, Reihen, Felder usw.) aufgrund von Übereinstimmungen und Differenzen zwischen ihnen. Die Bildung von lexikalischen Gruppierungen beruht auf Klassifikationsprozeduren und hängt recht stark von dem Ziel der Untersuchung ab. In diesem Sinne kann man aufgrund des gegebenen Materials verschiedene Klassifikationsaufgaben lösen. Die vom Forscher erstellten Gruppierungen sollte man als Konstrukte betrachten, die es erlauben, die Struktur des Vokabulars von einem gegebenen Standpunkt aus zu modellieren. Dies bedeutet aber nicht, daß unterschiedliche lexikalische Gruppierungen keinen ontologischen Hintergrund hätten, wenn die realen Zusammenhänge im lexikalischen System betrachtet werden.

Die Gruppierung von Wörtern und die Erforschung der Beziehungen innerhalb und zwischen den Gruppen sind besonders nützliche Verfahren der lexikalischen Analyse. In vielen solchen Fällen kann man eine noch nicht beschriebene Seite des lexikalischen Systems der Sprache entdecken und ihre weitere Analyse in die Wege leiten. Bekannt ist die Rolle der lexikalischen Gruppierungen (Klassifikationen) in der Lexikographie, in der Stilistik und in der Methodik des Sprachunterrichts sowie bei der Informationssuche in der automatischen Übersetzung.

Zuerst müssen wir einige Begriffe und Termini präzisieren. In Abhängigkeit vom Charakter ihrer Wechselbeziehungen fassen wir Wörter zu lexikalischen Subsystemen zusammen, die auf unterschiedlichen Beschreibungsebenen unterschiedliche Varianten aufweisen und auch "lexikalische Gruppen" genannt werden. Aufgrund ihrer phonologischen oder morphologischen Ähnlichkeit werden "lexikalisch-formale Gruppen" gebildet. Die Gemeinsamkeit der grammatischen Bedeutung ist die Grundlage für die Bildung von "lexikalisch-grammatischen Gruppen". Aufgrund der semantischen Ähnlichkeit werden "lexikalisch-semantische Gruppen" gebildet. Neben diesen Grundtypen kann man Gruppen aufgrund von anderen (etymologischen, stilistischen u.a.) Kriterien bilden. Man kann die Kriterien auch mischen.

Lexikalische Gruppen, die man in einem beliebigen Bereich mit Hilfe einer Klassifikationsprozedur aufstellt, haben substantielle *quantitative Eigenschaften*. So kann man z.B. im Rahmen eines endlichen Vokabulars den Umfang, d.h. die Anzahl der Elemente der lexikalischen Gruppen feststellen, wobei die numerischen Werte der Umfänge unterschiedlicher lexikalischer Gruppen in verschiedenen Bereichen typische Verteilungen bilden (s. unten). Bei ungleicher Größe der Umfänge der lexikalischen Gruppen kann man eine derartige Verteilung als Rangverteilung darstellen, die die einzelnen lexikalischen Gruppen nach "Gewicht" ordnet und dadurch eine linguistische Gesetzmäßigkeit signalisiert. Weiter kommen bei der Verwendung in der Rede die lexikalischen Gruppen unterschiedlich häufig vor, und auch innerhalb einer lexikalischen Gruppe selbst kommen die Wörter unterschiedlich häufig vor, was es uns ermöglicht, das Zentrum, den Kern, und die Peripherie der gegebenen Gesamtheit zu bestimmen. Bedenkt man die Vagheit der Grenzen von lexikalischen Gruppen (aufgrund der möglichen Überschneidung) und den probabilistischen Charakter der Verteilung der Verwendungshäufigkeiten der Wörter, dann scheint es gerechtfertigt, neben der qualitativen Analyse auch die quantitativ-systemische Analyse durchzuführen, wobei die Verteilung von quantitativen Charakteristika der lexikalischen Gruppen sowohl im Vokabular (Inventar) als auch im Text (Rede) als probabilistische Systeme betrachtet werden, die Stabilitäts- und Variabilitätseigenschaften besitzen. Das grundlegende Instrument der quantitativ-systemischen Analyse der Lexik, darunter auch die lexikalischer Grup-

pen, ist die Modellierung mit Hilfe von Verteilungen. Diesem Instrument widmen wir den folgenden Abschnitt.

Modellieren mit Hilfe von Verteilungen

Das Modellieren, d.h. die Aufstellung und Analyse von Modellen (Analogien des realen Objekts, des Originals) ist eine Methode der systemischen Erforschung der inneren Struktur des Objekts oder seines Verhaltens. Die Aufstellung des Modells ist im Grunde der Versuch, in die "Architektonik" der Zusammenhänge des untersuchten Systems einzudringen. Es ist bekannt, daß es Modelle verschiedener Art gibt (vgl. z.B. Stoff 1972). Für uns ist es wichtig, die Prinzipien der Aufstellung und Analyse solcher Modelle zu erarbeiten, die mit der quantitativen Untersuchung der Lexik als eines probabilistischen Systems verbunden sind. Dem entspricht in erster Linie das Modellieren mit Hilfe von *Verteilungen*, da eine Verteilung als der verallgemeinernde, integrale Begriff die wichtigste strukturelle Charakteristik probabilistischer Systeme ist. "Mit Hilfe des Begriffs der Verteilung - sagt Sačkov (1971:112) - charakterisiert man die Elemente von probabilistischen Systemen und ihre Wechselbeziehungen und begründet ihre Zugehörigkeit zum System selbst und zu anderen Systemen im Ganzen." Wichtig ist nicht nur, daß die Verteilung die Existenz einer inneren Ordnung im System ausdrückt, sondern auch daß sie die Wechselwirkung zwischen den Elementen und die Gemeinsamkeit ihres Verhaltens, d.h. die Ganzheitlichkeit des Systems und auch die Stabilität und Regularität in der Masse der Zufallsereignisse erfaßt. In Anbetracht dessen, daß eine Verteilung in beträchtlichem Maße von den inneren Eigenschaften der Systemelemente bestimmt wird, kann man davon ausgehend auch die Eigenschaften einzelner Elemente (z.B. Wörter oder Wortklassen) untersuchen.

Den Begriff der Verteilung kann man im weiten Sinne als eine geordnete Gesamtheit quantitativ ausgedrückter Werte, d.h. Meßergebnisse, gewöhnlich mit Angabe der Ausprägung (Häufigkeit, Wahrscheinlichkeit, Rang) dieser Werte in der gegebenen Gesamtheit, auffassen. Im engeren Sinne wird eine (Wahrscheinlichkeits-)Verteilung als die "Aufzählung der Werte der Zufallsvariablen und ihrer Wahrscheinlichkeiten" bestimmt (Veneckij, Kil'dišev 1975:110). Aber auch in diesem Fall gibt es noch keine Eindeutigkeit, z.B. fällt unter die Bestimmung der Wahrscheinlichkeitsverteilung auch die einzelne Wahrscheinlichkeit (die Wahrscheinlichkeit von *A* kann man nur bei gleichzeitiger Bestimmung der Wahrscheinlichkeit von non-*A* in der gegebenen Gesamtheit festgelegt werden). Folglich erfaßt die Verteilung sowohl im weiten als auch im engeren Sinne einen weiten Kreis von Erscheinungen bei der quantitativen Untersuchung von Objekten; sie kann einfachere Formen der Messung (ein Element in der gegebenen Gesamtheit) als auch komplexere Formen der Beziehungen zwischen Messungen einschließen.

Das wichtigste hier ist, daß man die Verteilung selbst systemisch interpretiert, in unserem Fall unter dem Gesichtspunkt der strukturellen Charakteristika probabilistischer Systeme (Wechselbeziehungen der Elemente, Stabilität und Variabilität).

Um die grundlegenden Möglichkeiten der Darstellung (Modellierung) der Daten einer quantitativen Untersuchung der Lexik in Form von Verteilungen zu erläutern, ist es zweckmäßig, von dem methodologischen Prinzip der Opposition von *Sprach-* (Potenz-) und *Rede-* (Realisierungs-) Bereich und der Sphären der *Statik* und *Dynamik* im Rahmen des allgemeinen Systems der Sprechtätigkeit (s. Abschnitt 1.2) auszugehen.

Aufgrund der Unterscheidung in Sprach- und Redebereich kann man von *theoretischen (sprachlichen)* und *empirischen (Rede-)* Verteilungen sprechen. Wie bereits erwähnt, haben wir neben der Abgrenzung von Sprach- und Redebereich im System der Sprechtätigkeit gleichzeitig die Sphären der Statik und der Dynamik unterschieden. Entsprechend kann man die Verteilungen in *statische* und *dynamische* aufteilen. Diese Unterscheidung hat vor allen Dingen einen inhaltlichen Sinn, da sich die Verteilungen ihrer äußeren Form nach nicht unterscheiden.

Statische Verteilungen, die man auch als "synchron" bezeichnet, drücken vor allem die synchronen und die paradigmatischen Aspekte der Analyse linguistischer Erscheinungen aus. Hierher gehören beispielsweise Rangverteilungen von Wörtern oder Häufigkeitsverteilungen von Wörtern im Text ("Häufigkeitsspektrum" der Lexik). Zu den statischen Verteilungen gehören aber auch Zerlegungen einiger Gesamtheiten (z.B. lexikalisch-semantischer Felder) in verschiedene Gruppen (Klassen, Cluster), die nach ihren probabilistischen Eigenschaften geordnet sind.

Dynamische Verteilungen in der Linguistik unterscheiden sich von den statischen dadurch, daß sie entweder eine Prozessualität, die mit den Erscheinungen der Redeerzeugung zusammenhängt (in der Synchronie), oder Veränderung, Entwicklung der Sprache (in der Diachronie) ausdrücken. In allen Fällen tritt Zeit in der Verteilung explizit oder implizit in Erscheinung, daher bezeichnet man dynamische Verteilungen oft als "diachronische" im weitesten Sinne des Wortes.

Neben der Klassifikation der Verteilungen in theoretische (sprachliche) und empirische (redebezogene) Verteilungen sind für die Linguistik daher auch statische (synchrone) und dynamische (diachrone) Verteilungen charakteristisch. Zusätzlich zu diesen grundlegenden linguistischen Verteilungen kann man noch eine Reihe von Untertypen oder Abarten unterscheiden, die wir nachfolgend besprechen werden.

Rein technisch gesehen kann man jede Verteilung in tabellarischer Form (als "Verteilungsreihe"), in graphischer Form oder in Form einer Formel (als funktionale Abhängigkeit) angeben. Dabei kann man die Tabelle, den Graphen oder die Funktion in *Differential-* oder *Integralform* (kumulativ) darstellen (man unterscheidet die Differentialfunktion oder "Dichtefunktion" und die Integralfunktion; vgl. Mitropol'skij 1971:209)

Aufgrund einiger theoretischer und praktischer Überlegungen drückt man manchmal eine *diskrete* Verteilung linguistischer Daten in Form einer *stetigen* Verteilung (Bektaev, Piotrovskij 1973:131) aus.

Ohne auf die Details der Anwendung konkreter technischer Methoden der vorbereitenden Datenaufbereitung (Erfassung, Aufteilung in Intervalle, Berechnung der Dispersionscharakteristika usw.) einzugehen, bleiben wir hier bei den Varianten der Datendarstellung in Verteilungsform, die den o.a. Regeln der Quantifizierung des linguistischen Materials entsprechen. Wir gehen von den Messungsformeln (1) und (2) aus, wo die drei Grundkomponenten: Objekt (X), Merkmal (P) und Wert des Merkmals (Y) aufeinander bezogen werden. Die Meßresultate kann man dann in Form von Tabellen darstellen, die als Grundlage für die Ermittlung der entsprechenden Verteilungen dienen. Man unterscheidet drei Grundschemata bei der Aufstellung von Tabellen (Verteilungsreihen).

Schema 1. *Einzelobjektverteilung*[11]
(ein Objekt, mehrere Merkmale)

	P_1	P_2	...	P_n
X	y_1	y_2	...	y_n

Nach diesem Schema werden dem Objekt Resultate von Messungen unter unterschiedlichen Bedingungen zugeschrieben. Beispielsweise, wenn X eine konkrete linguistische Einheit (oder eine Klasse von Einheiten), P_i die Häufigkeit im Text i und y_i die entsprechende Häufigkeit ist, dann kann man das "Verhalten" einer bestimmten linguistischen Einheit in einer Reihe von Experimenten untersuchen. Wenn man die quantitativen Werte y_i (und entsprechend P_i) nach wachsendem oder fallendem Wert ordnet, dann kann man:

[11] Die Einzelobjektverteilung entspricht der "horizontalen" Verteilung von Alekseev und der "Einobjektverteilung" von Martynenko (1982), auch wenn es prinzipiell verschiedene Lösungen für die Klassifikation linguistischer Verteilungen gibt.

(a) eine Variationsreihe mit der Angabe der Häufigkeiten oder Wahrscheinlichkeiten des Vorkommens von y_i aufstellen; in dem Falle konstruiert man das sogenannte Spektrum oder die spektrale Verteilung.

(b) den Werten y_i Ränge zuschreiben, wodurch man eine Rang-Häufigkeitsverteilung erhält.

Nach der spektralen Einzelobjekt-Verteilung spezifiziert man den Typ der Verteilung (bei linguistischen Objekten sind es meistens Verteilungen der Gauß-Familie, z.B. Binomialverteilung, Normalverteilung, Poissonverteilung usw.). Die Rangverteilungen bei diesem Schema der Einzelobjekt-Verteilungen sind in der Linguistik nicht besonders interessant. Ein spezieller Fall entsteht bei der Rangierung von y_i nach qualitativen Kriterien, beispielsweise bei der Untersuchung dynamischer (diachroner) Prozesse. In solchen Fällen kann man die Verteilung als "Trend" betrachten, wobei die Werte y_i von $P(t_i)$ funktional abhängen, wo t_i die Zeit bedeutet.

Schema 2: *Mehrobjektverteilung*[12]
(ein Merkmal – mehrere Objekte)

	P
X_1	y_1
X_2	y_2
.	.
.	.
.	.
X_m	y_m

Nach diesem Schema mißt man ein gemeinsames Merkmal an verschiedenen Objekten. Beispielsweise sind X_i unterschiedliche linguistische Einheiten (oder Klassen von Einheiten), P ist die Häufigkeit in einem Text, y_i sind die entsprechenden Häufigkeiten. Nach diesem Schema wird beispielsweise das übliche Häufigkeitswörterbuch von Wörtern oder Wortklassen (Wortarten, phonetischer, morphologischer, semantischer u.a. Worttypen) aufgestellt sowie jede andere Häufigkeitsliste von linguistischen Einheiten. Dabei erhält man die Verteilung von Häufigkeiten der Einheiten in Bezug zueinander in der gegebenen Gesamtheit (z.B. in dem gegebenen Text).

So wie bei der Einzelobjektverteilung gibt es auch hier zwei Unterarten (Darstellungsformen):

[12] Vgl. die "vertikale Verteilung" von Alekseev (1978, 1985) und die "Mehrobjektverteilung" von Martynenko (1982).

(a) die Spektralverteilung, wenn man gleiche Meßresultate in Gruppen mit Angabe der Anzahl der Objekte mit dem jeweiligen Meßresultat zusammenfaßt; beispielsweise wenn es einen Zusammenhang zwischen der Häufigkeit eines Wortes im Text und der Anzahl der Wörter mit der gegebenen Häufigkeit gibt (d.h. Häufigkeits- oder lexikalisches Spektrum);

(b) die Rangverteilung, wobei man den geordneten Häufigkeiten y_i ihre Ränge (i) zuschreibt und den Zusammenhang zwischen y_i und i untersucht (beispielsweise die Rangverteilung von Worthäufigkeiten).

Sowohl die Einzelobjekt-Spektralverteilung als auch die Einzelobjekt-Rangverteilung gehören üblicherweise in der Linguistik zu der "nicht-Gaußschen Familie". In ihnen äußert sich eine der charakteristischen Eigenschaften kommunikativer Systeme: Die asymmetrische Verteilung der Elemente nach deren "Bedeutsamkeit", wobei die grundlegende funktionale Belastung auf einigen wenigen dominierenden Elementen (dem "Kern") ruht.

Schema 3. *Komplexe (mehrdimensionale) Verteilung*
(mehrere Objekte, mehrere Merkmale)

	P_1	P_2	...	P_n
X_1	y_{11}	y_{12}	...	y_{1n}
X_2	y_{21}	y_{22}	...	y_{2n}
...				
X_m	y_{m1}	y_{m2}	...	y_{mn}

Dieses Schema ist eine Kombination der beiden obigen Schemata. In einer einfacheren Form ist die Zahl der Objekte (X) und die der Merkmale (P) gleich zwei, wenn man beispielsweise die Häufigkeiten unterschiedlicher Wörter in zwei Texten untersucht, oder die Häufigkeiten verschiedener Wortklassen von zwei Gesichtspunkten aus betrachtet: im Wörterbuch und im Text. Aufgrund des Schemas der komplexen Verteilung kann man "multidimensionale" Aufgaben in Angriff nehmen: Die Wechselbeziehung und gleichzeitig die Variation einer Reihe von Objekten oder Merkmalen. Die Beziehung zwischen den Verteilungen quantitativer Werte (horizontal oder vertikal) kann in Form einer funktionalen Abhängigkeit (Regressionsgleichung) ausgedrückt werden, und die Stärke der Beziehung kann mit dem (linearen oder nicht-linearen) Korrelationskoeffizienten gemessen werden. Die internen Beziehungen in der ganzen Gesamtheit kann man mit Hilfe von Faktorenanalyse, Clusteranalyse usw. untersuchen.

Ein spezifisches Merkmal dieses Ansatzes zur Klassifikation (Typologie) linguistischer Verteilungen ist die Tatsache, daß man hier von qualitativen Vorstellungen über die Verteilung als einem Meßresultat ausgeht, wobei drei Komponenten aufeinander bezogen werden: Objekt (X), Merkmal (P) und der Wert des Merkmals (y). Auch wenn man im strengen Sinne nur die Reihe y als Verteilung betrachten kann, ob nun in spektraler oder in Rangform, darf man bei der konkreten Analyse die "Provenienz" dieser Reihe nicht vergessen, d.h. die Beziehungen der Komponente y zu den Komponenten X und P. Dies ist nicht nur für die Erarbeitung von vorläufigen Hypothesen (in Form einer Verteilung u.ä.) notwendig, sondern auch für die inhaltliche Interpretation der Resultate der quantitativen Analyse unter Berücksichtigung der Spezifik und der Aufgabe der *linguistischen* Untersuchung. Außerdem, wie oben gezeigt wurde, ermöglicht es gerade die Berücksichtigung aller Komponenten der Meßprozedur (X, P und y), die grundlegenden linguistischen Einzelobjekt-, Mehrobjekt und die komplexen linguistischen Verteilungen deutlich und natürlich zu unterscheiden.

In der Praxis wird eine "elastische" Handhabung der Begriffe des Objekts, des Merkmals und des Merkmalswertes zugelassen. Das Objekt kann eine individuelle Einheit (z.B. ein konkretes Wort) oder eine Klasse von Einheiten (Wortarten usw.) sein. Das Merkmal (oder exakter der Name des Merkmals) kann einfach oder zusammengesetzt sein, z.B. "Häufigkeit im Text T_i". Der Wert des Merkmals kann auf der quantitativ-proportionalen Intervall- oder Ordinalskala angegeben werden. Infolge dieser Flexibilität bei der Bestimmung der Meßkomponenten gibt es verschiedene Varianten der Klassifikation linguistischer Verteilungen in Abhängigkeit von den konkreten Bedingungen und Aufgaben der Untersuchung.

Alle diese Unterarten von Verteilungen können der quantitativen Untersuchung und Modellierung linguistischer Objekte dienen. Dabei sind unterschiedliche Darstellungs- und Beschreibungsarten der Verteilungen möglich, darunter in Form von funktionalen Abhängigkeiten. Einige Typen von Verteilungen kann man auch in Begriffen der Theorie der unscharfen (vagen) Mengen beschreiben (Zadeh 1976, 1980; Lesochin 1982; über die Möglichkeit des probabilistischen Zugangs zur Theorie unscharfer Mengen s. Nalimov 1979a). Bei der Untersuchung einiger dynamischer Erscheinungen (z.B. des Prozesses der Redeerzeugung) kann man auch spezifische Methoden der Untersuchung von Zufallsprozessen unter Zuhilfenahme der Theorie der "optimalen Prozesse", der Theorie der "dynamischen Programmierung" (Ventcel' 1976) usw. verwenden.

Die Orientierung an der Modellierung mit Hilfe von Wahrscheinlichkeitsverteilungen beim quantitativ-systemischen Ansatz erklärt sich durch die Spezifik der systemischen Untersuchung. Sie besteht darin, daß diese Analysemethode gerade den Aspekt untersucht, der ein Objekt – gemäß den Prämissen dieses Ansatzes – zum System macht. In diesem Sinne ist eine linguistische Verteilung eine *Modellbeschreibung* der sprachlichen Objekte, die man sich als probabilistische Systeme vorstellen kann.

Die Erforschung linguistischer Verteilungen fängt in der Regel mit der Ermittlung der empirischen Verteilung (oder Häufigkeitsverteilung) an, die man als eine erste Approximation des probabilistischen Modells betrachten kann. Eine Häufigkeitsverteilung kann eine hinreichend gute Vorstellung von den stabilen Eigenschaften in der Struktur und der Funktion des untersuchten Systems geben. Die Modellierung von Daten mit Hilfe einer theoretischen Verteilung hebt die Untersuchung auf eine höhere, allgemeinere Ebene. Die wichtige Rolle theoretischer Verteilungen bei der Wiedergabe der Gesetzmäßigkeiten der materiellen Welt ist hinreichend bekannt, und man kann voraussetzen, daß in bezug auf einige Klassen linguistischer Objekte eine derartige Modellierung zahlreiche Möglichkeiten nicht nur für die Lösung vieler praktischer Aufgaben, sondern auch für die Entdeckung tieferer quantitativer Strukturen und Funktionen der Sprache eröffnet.

Der Modellierungsprozeß verläuft vom Objekt zum Modell (Modellaufstellung), und dann vom Modell zum Objekt (Interpretation des Modells, Gewinnung neuer Erkenntnisse über das Objekt).

Interpretation linguistischer Verteilungen

Linguistische Verteilungen sind an sich nicht rein formale Modelle, man sollte sie als interpretierte Zeichensysteme betrachten. Die Erörterung der Eigenschaften und Beziehungen sowie die der inneren Gesetzmäßigkeiten, die den Charakter linguistischer Verteilungen insgesamt bestimmen, muß durch eine qualitative (inhaltliche) Analyse der Untersuchungsresultate begleitet werden.

Die Interpretation linguistischer Verteilungen als Modelle der untersuchten probabilistischen Systeme kann man sich als einen vielschichtigen (multistadialen) Prozeß der qualitativ-quantitativen Analyse und Synthese vorstellen. Die Aufgabe der Interpretation besteht darin, aufzudecken, was hinter dem Phänomen steht, jedoch gibt es hier mehrere Wege, dieses Ziel zu erreichen. Man unterscheidet den induktiven und den deduktiven Ansatz, wobei die wissenschaftliche Erklärung eine höhere Form der Interpretation bildet, die zeigen muß, daß "die gegebene wissenschaftliche Tatsache die Konsequenz eines Gesetzes ist oder daß das zu erklärende Gesetz aus einem noch allgemeineren Gesetz (oder einer Theorie) folgt" (Drujanov 1980:53). In Zusammenhang mit den Bedingungen und Aufgaben der Untersuchung und aufgrund des Charakters des Materials kann man linguistische Verteilungen auf drei verschiedene Weisen interpretieren: strukturell-funktional, pragmatisch (stilistisch) und genetisch (kausal). Da diese drei Interpretationen in Wechselbeziehung zueinander stehen und sich überschneiden, kann man sie in einer konkreten Untersuchung gemeinsam finden.

1. Strukturell-funktionale Interpretation

Eine wichtige Bedingung des Modellierens ist die Analyse der physischen Gestaltung des untersuchten Objekts in der Vormodellierungsphase, um einige apriori-Informationen für die Ableitung der Form des gesuchten Modells zu erhalten. Die einfachste und effektivste Methode zur Darstellung der Information über die Verteilung ist die *graphisch-geometrische Methode.* Graphiken erlauben es, die Daten in anschaulicher Form mit minimalem Aufwand darzustellen. Aufgrund einer graphischen Darstellung kann man Schlüsse auf die allgemeine Form der Verteilung und auf den Charakter der Wechselbeziehung der Elemente des untersuchten Systems (Symmetrie - Asymmetrie, Linearität - Nichtlinearität, Eingipfligkeit - Mehrgipfligkeit usw.) ziehen. Eine S-förmige Kurve sagt uns, daß die gegebene Verteilung möglicherweise einen dynamischen Prozeß widerspiegelt, der logistisch verläuft, und eine Hyperbel kann auf "Konzentration und Streuung" der Systemelemente hindeuten. Graphiken kann man in dem üblichen kartesischen Koordinatensystem darstellen oder in modifizierter Form, z.B. in logarithmische Transformation, wobei man sich vergewissern kann, ob die Daten einem bestimmten Verteilungsgesetz folgen. Sehr wertvoll sind Graphiken beim Versuch, neue Information über Abweichungen von der allgemeinen Tendenz in bestimmten Intervallen zu gewinnen, über Wendepunkte usw. So lieferte beispielsweise die Analyse der Verteilungsform der Worthäufigkeiten in doppeltlogarithmischen Koordinaten den Impuls für die Entdeckung eines besonderen "nichtlinearen" Rangverteilungstyps der Lexik in großen Texten und zur Formulierung der These über "die vierte Approximation des Zipfschen Gesetzes" (Alekseev 1978).

Im Unterschied zu Graphiken haben *Formeln* den großen Vorteil, daß sie verschiedene Operationen zulassen, die schon an sich Ausgangspunkt für die Entdeckung neuer, unerwarteter Seiten des untersuchten Phänomens sein können. Eine Formel kann man als eine symbolische (analytische) Niederschrift der Struktur der gegebenen Erscheinung betrachten, aber die Formel kann auch eine Vorstellung von der Funktion des Systems, von dem dynamischen Prozeß, von seinem Wachstum usw. geben.

In der quantitativen Linguistik benutzt man üblicherweise verschiedene Funktionen oder Differentialgleichungen als analytische Ausdrücke von Verteilungen. Stellt man die Verteilung in Form einer Funktion dar (als funktionale Abhängigkeit)[13], dann beginnt man oft mit der Bestimmung der Funktionsform aufgrund der empirischen Daten. Ein derartiger *induktiv-empirischer* Zugang kann manchmal ganz akzeptable Resultate liefern. Wenn beispielsweise festgestellt wurde, daß die

[13] In probabilistischen Systemen behandelt man eine Funktion als Wahrscheinlichkeitsfunktion. Formal gibt es keinen Unterschied zwischen ihnen, solange wir den mathematischen Symbolen keine bestimmte Bedeutung zuschreiben.

gegebene empirische Verteilung mit Hilfe einer Potenzfunktion des Typs $y = ax^b$ (mit a und b als Parameter) darstellbar ist, dann kann man Schlüsse auf das "allometrische" Gesetz der Veränderung von y in Abhängigkeit von x ziehen, wobei der Parameter b ("der Koeffizient der relativen Elastizität") erlaubt, die durchschnittliche Prozentzahl der Veränderung von y bezüglich der Veränderung von x um 1% zu bestimmen, und der Parameter a die Anfangsgröße von y bei $x = 1$ angibt. Die qualitative Interpretation der Formel und ihrer Parameter erfordert natürlich die Analyse der Erscheinung oder des Prozesses aufgrund ihrer inneren Logik oder aufgrund ihrer physischen Gestaltung, um zu einer adäquaten Vorstellung von der untersuchten Erscheinung zu gelangen. Weiter, wenn wir diese Formel in der Differentialform schreiben, z.B. als Gleichung $(dy/y)/(dx/x) = b$, dann sehen wir, daß das Verhältnis der relativen Zuwächse von x und y konstant ist ("stabil bleibt") und sich dadurch die Erscheinung dem Gesetz des "stabilen relativen Zuwachses" unterordnet (Land 1977: 388), eines der wichtigsten Gesetze, die für einige Typen von selbstorganisierenden Systemen charakteristisch sind.

Analog kann man die Exponentialfunktion als "lawinenartiges Wachstum" interpretieren, die logarithmische Funktion als das "Gesetz der adaptiven Inhibition" (Nalimov, Mul'čenko 1969:41) oder das "Gesetz des proportional abnehmenden relativen Wachstums" (Land 1977:388), die logistische Funktion als "Wachstum mit Initialbeschleunigung und anschließender Verlangsamung (mit Sättigung)", die Weibullfunktion als das "verallgemeinerte Modell des progressiven Wachstums" (vgl. Dobrov 1969:158) usw.

Auf einer anderen Ebene der wissenschaftlichen Analyse verwendet man *hypothetisch-theoretische* Modelle von probabilistischen Systemen, wenn man nämlich Funktionen/Modelle aufgrund theoretischer Postulate mit größerer oder geringerer Allgemeinheit hypothetisch ableitet. Die Interpretation solcher Modelle enthält sowohl die Ausgangpostulate als auch die möglichen Konsequenzen, die man aufgrund des konkreten empirischen Materials zieht. In der quantitativen Linguistik sind die Versuche bekannt, das Zipfsche Gesetz aufgrund eines stochastischen Prozesses (vgl. Simon 1955) oder aufgrund einer Analogie zur Thermodynamik (vgl. Mandelbrot 1954) abzuleiten. Einigen Forschern ist es gelungen, Formeln der quantitativen Struktur des Textes aufgrund von kombinatorisch-variationalen Prinzipien abzuleiten (vgl. Arapov, Šrejder 1978; Krylov 1987). Das Modell des "verallgemeinerten Zipf-Mandelbrotschen Gesetzes" (Orlov 1976) wurde theoretisch abgeleitet, aber die These von der besonderen Rolle des "Zipfschen Umfangs" wurde aufgrund empirischer Befunde begründet und interpretiert, wobei man den Schluß gezogen hat, daß die Befolgung des Zipfschen Gesetzes (entsprechend dem "Zipfschen Umfang") von einem "hohen Maß an Organisiertheit des vollständigen Textes" zeugt.

In der quantitativen Linguistik praktiziert man auch den intermediären *hypothetisch-empirischen* Ansatz, wobei man das Ausgangsmodell aufgrund von relativ

selbständigen theoretischen Schemata (Hypothesen) erhält und die konkrete Form der Verteilung auf dem empirischen iterativen Wege bestimmt, indem man sich an die Forderungen des theoretischen Modells hält (vgl. Abschnitt 2.3).

Es gibt weitere interessante Versuche, theoretische Modelle zu konstruieren und zu interpretieren, indem man von der Analyse *gegenseitiger Abhängigkeiten* (der *"Synergetik"*) zwischen den linguistischen Objekten ausgeht (Altmann 1980; Köhler 1986). Das höchste Ziel dieser Untersuchungen ist die Aufstellung einer adäquaten linguistischen Theorie im Rahmen der allgemeinen Theorie selbstorganisierender Systeme.

2. Pragmatische (stilistische) Interpretation

Unter diesem Begriff versteht man die Erklärung, die auf dem Zusammenhang zwischen formalen (quantitativen) und einigen pragmatisch-stilistischen Kenngrößen beruht, darunter etwa Schätzungscharakteristika der untersuchten Erscheinungen. In der quantitativen Linguistik, insbesondere in einer ihrer Spezialdisziplinen, der quantitativen Linguostilistik oder Stilometrie, operiert man mit inhaltlichen (stilistischen) Begriffen wie "Vokabularreichtum" oder "Differenziertheit" des Vokabulars, "Vollständigkeit" und "künstlerische Abgeschlossenheit" des Textes, "Stereotypie" und "Ökonomie" der Mitteilung usw. Alle Begriffe dieser Art haben das Ziel, einige Eigenschaften realer Objekte zu charakterisieren, auch wenn man diese Eigenschaften nicht direkt beobachten kann und dadurch auch nicht direkt messen kann. In solchen Fällen betrachtet man äußerlich beobachtbare Aspekte der untersuchten Objekte (Wörterbücher, Texte) und versucht, indirekt über diese zum Kern des Problems zu gelangen.

Theoretisch ist diese Situation mit der Lage in der Meßtheorie zu vergleichen, wo man zwei Typen von Variablen hat: (1) *latente Variablen*, d.h. das, was der Forscher für seine Analyse auswählt und fixiert; (2) *Indikatoren,* d.h. das, was man direkt messen kann (Haitun 1983:16). Die notwendige Bedingung für die Analyse ist dabei die Existenz einer bestimmten Verbindung zwischen den latenten Variablen und den Indikatoren.

Zu den formalen Indikatoren gehören auch die linguistischen Verteilungen: Es ist z.B. bekannt, daß "Vokabularreichtum" aufgrund des Verlaufs der Rangverteilung der Wörter bestimmt wird (konkret aufgrund des Winkels der Kurve oder aufgrund des Zipfschen Parameters γ), oder aufgrund der komplexen Verteilung, die den Zusammenhang zwischen dem Vokabular- und dem Textumfang ausdrückt (davon ausgehend kann man die Wachstumstendenz, d.h. den "potentiellen Reichtum" des Vokabulars prognostizieren). Die asymmetrische Verteilung von Wörtern erklärt man als die Manifestation des Prinzips der "Präferenz" oder "Bedeutsamkeit" des gegebenen Vokabularteils unter den gegebenen Bedingungen, was zur

"Konzentration und Dispersion" der Einheiten führt. "Aktivität" und "Deskriptivität" des Stils beurteilt man aufgrund der Verteilung der Wortarten im Text usw.

Bei dieser Analyse darf man nicht vergessen, daß der Zusammenhang zwischen quantitativen Indikatoren und latenten Variablen (Eigenschaften, Charakteristika) einen probabilistischen Charakter hat. Das bedeutet erstens, daß die Interpretation der Daten nur aufgrund von repräsentativem Material gültig ist, d.h. bei hinreichend großen Stichproben und Reproduzierbarkeit der Resultate der Experimente. Es stellt sich auch die Frage, in welchem Maße die ausgewählten Indikatoren imstande sind, die fundamentalen Aspekte der Erscheinung zu erfassen und ob es bei falscher Wahl von Indikatoren oder bei unzureichender Repräsentativität der Indikatoren zu einer Verzerrung des Bildes kommen kann. Ob nun die Verbindung eines Indikators mit der gegebenen latenten Variablen gerechtfertigt ist, wird in jedem konkreten Fall in der Praxis getestet.

3. Genetische (kausale) Interpretation

Unter genetischer Interpretation versteht man den Versuch, die untersuchte Erscheinung von seiner Genese her zu erklären und dadurch direkt oder indirekt die Ursache seiner Existenz zu erfassen. Hier, genauso wie bei den vorherigen Interpretationstypen, muß man berücksichtigen, daß wir es hier mit einem probabilistischen Ansatz zur Untersuchung sprachlicher Erscheinungen zu tun haben. Dabei muß man von der probabilistischen Konzeption der Kausalität ausgehen, laut derer "Kausalität etwas ist, das in größerem oder geringerem Maße vorhanden sein kann, und nicht nur ausschließlich sein oder nicht sein kann" (Wiener 1964:309). Man muß auch das Prinzip der Mannigfaltigkeit der Ursachen zulassen, das auf der Mannigfaltigkeit der Zusammenhänge der Erscheinung mit anderen Erscheinungen beruht.

Bei der Untersuchung linguistischer Verteilungen kann man einiges mit "internen" *linguistischen* Ursachen erklären, beispielsweise mit Besonderheiten der morphologischen Struktur der gegebenen Sprache. Solche internen Ursachen sind jedoch immer mit äußeren (außerlinguistischen) Ursachen verwoben (Sprachkontakt, Gemeinschaftsbedürfnisse usw.). In letzter Zeit werden zur Erklärung linguistischer Verteilungen als Modelle bestimmter Seiten der Sprechtätigkeit besonders psychologische (psycholinguistische), psychophysiologische und phylogenetische Überlegungen herangezogen.

So versucht man beispielsweise die "hyperbolische" Verteilung der Worthäufigkeiten (in Form einer Potenzfunktion), bekannt als Zipfsches Gesetz, einerseits mit den Besonderheiten der *Psyche des Menschen*, seinem Kommunikationsbedürfnis und andererseits mit seiner Tendenz nach Minimierung seiner geistigen und physischen Anstrengung zu verbinden. Dieses allgemein bekannte Prinzip der

"geringsten Anstrengung" (Zipf 1949) begründet sich daher auf der Wechselwirkung zweier gegenseitiger Tendenzen in der Psyche (im Unterbewußtsein) des Menschen. Mit der Wechselwirkung konkurrierender Tendenzen bei der Redeerzeugung (Variabilität – Einschränkung der Variabilität) und mit assoziativen Eigenschaften des menschlichen Gedächtnisses (bei beschränktem Umfang des Kurzzeitgedächtnisses) erklärt man auch einige andere bekannte linguistische Verteilungen, beispielsweise die komplexe Verteilung, die das im Vergleich mit dem Anwachsen des Textumfangs langsamere Anwachsen des Vokabularumfangs ausdrückt (vgl. Abschnitt 2.3).

Die Vertreter des Psychophysiologie verbinden einige Typen linguistischer Verteilungen als Modelle der Sprechtätigkeit mit strukturellen Besonderheiten des Gehirns, besonders mit der raum-zeitlichen Organisation periodischer (zyklischer) *Prozesse im Gehirn*. Aufgrund der Vorstellung von der Kodierung von Wortbildern durch "Wellenpakete neuronaler Tätigkeit" leitete A.N. Lebedev eine Formel ab, die mit der Zipfschen Formel zur Beschreibung der Worthäufigkeiten in der Rede übereinstimmt, und eine andere, die den Zusammenhang zwischen dem Wachstum des Vokabulars und dem Wachstum des Textumfangs beschreibt (Lebedev 1983, 1986). Man kann feststellen, daß die Hypothesen über die Verbindung zwischen den Besonderheiten der quantitativen Struktur des Textes und einigen Gesetzmäßigkeiten der Gehirntätigkeit offensichtlich ihre Berechtigung haben und daß Untersuchungen in diesem Bereich zunehmend aktuell werden.

Im phylogenetischen Bereich erklärt man linguistische Verteilungen mit der *evolutionären Entwicklung* des menschlichen Gehirns, das sich im Laufe von Jahrtausenden an die äußere Umwelt angepaßt hat. Dieser evolutionäre Prozeß der Sprachentwicklung "ähnelt entfernt der organischen Evolution aufgrund natürlicher Selektion" (Panov 1980:147). Die Hierarchisierung als Systemeigenschaft grundlegender linguistischer Verteilungen kann man mit Adaptation an die Vergrößerung der Elementenanzahl (bei der Redeerzeugung) erklären, weil eine hierarchische Struktur die Zahl der Verbindungen grundsätzlich minimiert (vgl. Kozačkov 1978:15). Die Stabilität einiger linguistischer Verteilungen (die Aufrechterhaltung ihrer allgemeinen Form) weist auf eine Tendenz zum Gleichgewicht, zur Optimalität und Zweckmäßigkeit des komplexen selbstorganisierenden Systems der Sprache hin.

Viele der erwähnten Eigenschaften haben sicherlich einen allgemeinen Charakter, man findet sie in der lebenden und der leblosen Natur wieder. Dies zeugt von der "Einheit der Welt", die darauf beruht, daß "allgemeinere Gesetze der 'niedrigeren' Stufen des Seins auch in allen 'höheren' Stufen aufrechterhalten werden" (Brušlinskij 1979:47), wobei diese Universalität "die Existenz spezifischer Gesetzmäßigkeiten nicht nur ausschließt, sondern im Gegenteil voraussetzt" (ibidem).

Folglich kann man festhalten, daß die genetische Erklärung, die die Suche nach Ursachen der Erscheinungen an die erste Stelle setzt, sowie die pragmatische Me-

thode, die auf der Analyse latenter Variablen beruht, eine feste Stelle unter den Interpretationen linguistischer Verteilungen einnehmen.

2. Die statistische Organisation von Vokabular und Text

In diesem Kapitel werden Fragen der Kompilation und der Analyse von Häufigkeitswörterbüchern betrachtet und grundlegende Gesetzmäßigkeiten der statistischen Organisation von Vokabular und Text vom Gesichtspunkt des quantitativ-systemischen Ansatzes in der Lexikologie aus erörtert: das Verhältnis zwischen Wortform und Lexem im Text, die Schichtung der Wörter in bezug auf ihre Häufigkeit, die Frequenzstruktur des Textes, der Zusammenhang zwischen Vokabular und Text etc.

2.1. Häufigkeitswörterbücher

Kompilation von Häufigkeitswörterbüchern

Eine der wichtigsten Aufgaben der quantitativen Linguistik besteht im Erstellen von Häufigkeitswörterbüchern, deren effektive Verwendung bei der Lösung einer Vielzahl von Anwendungs- und Forschungsfragen ständig zunimmt. Ein Häufigkeitswörterbuch vermittelt einen Eindruck von der statistischen Struktur des Vokabulars und auch des Textmaterials, das bei der Erstellung des Häufigkeitswörterbuchs als Grundlage diente. Ein Häufigkeitswörterbuch kann als eine Art Modell für einen auf spezielle Weise gebildeten Text betrachtet werden, als Modell der Verteilung der Anwendungshäufigkeiten der Texteinheiten. Ein Häufigkeitswörterbuch stellt sich als geordnete Liste von Wörtern dar, die mit Daten über ihre Anwendung im Text (in der Rede) versehen sind. Im Hinblick auf die Anordnung der Einheiten werden zwei grundlegende Typen von Häufigkeitswörterbüchern unterschieden: (1) das alphabetische und (2) das rangierte Häufigkeitswörterbuch.

Im ersten Fall befinden sich die Wörter mit der Angabe der jeweiligen Häufigkeit in alphabetischer Anordnung, im zweiten Fall in der Reihenfolge der absteigenden Häufigkeit, wobei auch der Rang angegeben sein kann.

Eine spezielle Unterart ist das rückläufige alphabetische Häufigkeitswörterbuch, in dem die Wörter alphabetisch, aber in Sortierung vom Wortende her, angeordnet sind. Die Einträge eines Häufigkeitswörterbuchs können aus Wortformen oder Lexemen bestehen. Für besondere Zwecke werden auch Häufigkeitswörterbücher erstellt, deren Einträge Wortstämme, lexikalisch-semantische Lesarten,

Wortfügungen u.a. sind. Als Ausgangsmaterial für die Erstellung eines Häufigkeitswörterbuchs kann sowohl ein einzelner Text dienen als auch ein Textkorpus, wobei in beiden Fällen als Grundlage sowohl ganze (abgeschlossene) Texte als auch Textfragmente (Stichproben) verwendet werden. (siehe Alekseev, 1980, 1984 für Einzelheiten über die Typologie und die Anwendung von Häufigkeitswörterbüchern).

Bei der Erstellung eines Häufigkeitswörterbuchs und bei der anschließenden Analyse der Daten müssen einige allgemeine Anforderungen der quantitativen Linguistik beachtet werden. Große Bedeutung besitzt die Frage nach der Repräsentativität des Materials. Diese Frage wird in der einschlägigen Literatur näher beleuchtet; hier wollen wir nur anmerken, daß in der Praxis bei der Erstellung von Häufigkeitswörterbüchern die Aufmerksamkeit vor allem auf die Auswahl und die Bemessung des linguistischen Materials zu richten ist. Als Ausgangsbasis bei der Erstellung einer Textstichprobe nimmt man gewöhnlich eine "minimale" oder "Standard"-Stichprobe - ein Textfragment im Umfang von 1000 laufenden Wörtern (Andreev, 1967; Alekseev, 1968; Jakubajtis, 1981). Es bestätigt sich, daß eine solche minimale Stichprobe alle Züge besitzt, die einem zusammenhängenden Text zukommen (Jakubajtis, Skljarevič, 1978: 62), und daher für die Lösung der verschiedenartigsten Aufgaben der quantitativen Linguistik geeignet ist. Es ist wichtig, das Prinzip der Homogenität des Materials zu beachten; deshalb wird eine Untersuchung gewöhnlich auf eine einzige Subsprache oder sogar auf einzelne Texte und ihren Vergleich beschränkt. 'Häufigkeitswörterbuch der Gesamtsprache' ist ein relativer Begriff, da die quantitativen Eigenschaften eines solchen Wörterbuchs stark von der Zusammensetzung und der Proportion von Texten aus Subsprachen abhängt, die als Grundlage zu ihrer Erstellung herangezogen wurden. In der Praxis zeigen sich in Fragen der Umfangsfestlegung für große Häufigkeitswörterbücher beträchtliche Divergenzen. So wurde das Häufigkeitswörterbuch des Russischen (Zasorina, 1977) auf der Grundlage von Texten im Gesamtumfang von einer Million laufender Wörter erstellt, die zu etwa gleichen Anteilen aus vier verschiedenen Textsorten stammten: künstlerische Prosa (25,4%), dramaturgische Texte (27,2%), wissenschaftlich-publizistische Texte (23,6%) und journalistische Texte (23,8%). Das Häufigkeitswörterbuch der slowakischen Sprache (Mistrík, 1969) entstand auf der Grundlage von Texten im Umfang von einer Million laufender Wörter, die sich wie folgt verteilten: künstlerische Prosa (30,2%), wissenschaftlich-technische Texte (31,5%), journalistische Texte (14,6%), Poesie (13,2%) und Dialoge (10,5%). Das Häufigkeitswörterbuch der französischen Sprache (Juilland et al., 1970) hat als Grundlage Texte aus fünf Textsorten zu 100.000 laufenden Wörtern: künstlerische Prosa, Geschäftstexte, Publizistik, Drama und Essay. Die Grundlage des integrierten Häufigkeitswörterbuchs des Finnischen besteht aus Texten im Gesamtumfang von 400.000 laufenden Wörtern, die sich auf folgende Weise verteilen: künstlerische Prosa (11,5%), Rundfunksendungen (19,2%), Publizistik (26,0%), Verschiedenes (43,3%). Ein Gesamtumfang von 400.000 laufenden Wörtern liegt auch der Erstellung von vielen weiteren Wörterbüchern zugrunde, zum Beispiel des Häufigkeitswörterbuchs des Spanischen, das auf der Basis von Texten vier verschiedener Subsprachen zu unterschiedlichen Anteilen kompiliert worden ist (García Hoz, 1953).

Zur Frage des Textumfangs einer Subsprache extistieren verschiedene Ansichten. Gewöhnlich gilt, daß bei der Erstellung eines Häufigkeitswörterbuchs einer Subsprache ein Textkorpus im Gesamtumfang von nicht weniger als 200.000 laufenden Wörtern erforderlich ist (bei Einzelstichprobengrößen zu 1000 bis 10.000 laufenden Wörtern). Unsere Erfahrung zeigt, daß als ausreichende Untergrenze für eine einzelne Subsprache ein Umfang von 100.000 laufenden Wörtern gelten kann (bei einer Einzelstichprobengröße von 1000 bis 5000 laufenden Wörtern) unter der Bedingung, daß das Korpus dieses Umfangs nicht weniger als 2000 Wörter (Lexeme) mit einer Häufigkeit von $F \geq 5$ und einer Textabdeckung von ungefähr 80% enthält (Tuldava, 1977c:149). Der Häufigkeit $F = 5$ entspricht ein theoretisches Konfidenzintervall von 1...9, also mit einer Untergrenze > 0, wenn man voraussetzt, daß seltene Wörter ungefähr nach der Poisson-Verteilung verteilt sind. (Der Fehler kann mit Hilfe der Näherungsformel $2\sqrt{F}$ auf dem Signifikanzniveau 95% berechnet werden.) Diese statistischen Maße sind ausreichend für die Beschreibung des lexikalischen Kerns einer gegebenen Subsprache; dabei muß das Häufigkeitswörterbuch nach dem Prinzip des Verteilungswörterbuchs aufgebaut sein, das eine Stabilitätsbeurteilung der Worthäufigkeiten im Hinblick auf die Einzelstichproben gestattet. Bei Bedarf ist es zum Beispiel möglich, solche Wörter aus der Betrachtung auszuschließen, die in lediglich ein bis zwei der Einzelstichproben erscheinen. Anhand der Daten eines Verteilungswörterbuchs ist es möglich, Stabilitäts- und Gebrauchskoeffizienten für die Wörter zu berechnen (Juilland et al., 1970; Andrjuščenko, 1978) oder andere Schätzmethoden für die Verläßlichkeit der Häufigkeiten unter Berücksichtigung der tatsächlichen Verteilung der Wörter über die Einzelstichproben anzuwenden (z.B. Perebejnos, 1984). Die endgültige Antwort auf die Frage nach der für die Erstellung eines Häufigkeitswörterbuchs ausreichenden Korpusgröße hängt von den Zielen und Aufgabenstellungen der Untersuchung ab.

So wird, wenn die qualitative Zuverlässigkeit von Häufigkeitswörterbüchern von der Auswahl der Texte abhängt, die quantitative, das heißt statistische Zuverlässigkeit von Häufigkeitswörterbüchern ausgehend von spezifischen Anforderungen geschätzt, die in der modernen quantitativen Linguistik erarbeitet worden sind. Dabei muß angemerkt werden, daß die Vorstellung von einem Häufigkeitswörterbuch als einer Liste von Wörtern, die in strenger Häufigkeitsrangfolge angeordnet sind, sowohl vom praktischen als auch vom theoretischen Gesichtspunkt aus gesehen kaum gerechtfertigt ist. Für die Lösung der Mehrzahl der aktuellen Probleme in der quantitativen Linguistik, insbesondere bei der Untersuchung von Gesetzmäßigkeiten der statistischen Organisation des Vokabulars, ist die Verfügbar-

keit eines "absoluten Häufigkeitswörterbuchs" nicht unbedingt erforderlich. Es ist zu betonen, daß der Sinn der Erstellung von Häufigkeitswörterbüchern vor allem in der "Stratifikation von in ihrem statistischen Gewicht unterschiedlichen Schichten der Lexik" (Zasorina, 1966:70) besteht, das heißt in der Sichtbarmachung der fundamentalen lexikalischen Frequenzzonen. In Anbetracht der Besonderheit der quantitativen Untersuchung linguistischer Gegenstände macht es, wie es einige Forscher einschätzen, keinen Sinn, von exakten Wahrscheinlichkeiten von Wortvorkommen im Text zu sprechen - vielmehr stellen sie in der Eigenschaft der Invarianz von Vorkommen einen unscharfen Frequenzbereich für jedes einzelne Wort fest, das heißt ein Spektrum von zulässigen Häufigkeiten oder Rängen von Wörtern in der beschriebenen Textsorte (Arapov et al., 1978). Vom Standpunkt des quantitativ-systemischen Ansatzes aus wird die Rangierung von Wörtern nach Vorkommenshäufigkeit unter dem Aspekt der probabilistischen Systeme betrachtet, das heißt unter dem Aspekt der Wechselwirkung zwischen Stabilität und Variabilität, wobei ein Schwerpunkt auf der Untersuchung der systemischen Eigenschaften des Vokabulars liegt, insbesondere auf der Analyse der Interdependenzen und Gruppierungen von Wörtern und auf der Modellierung mit Hilfe von verschiedenen statistischen Verteilungen.

Fundamentale Charakteristika von Häufigkeitswörterbüchern

Die allgemeinsten statistischen Eigenschaften von Häufigkeitswörterbüchern sind: *N* - der Textumfang, auf dessen Grundlage das Wörterbuch erstellt worden ist (d.h. die Anzahl der Wortvorkommen im Text); *V* - die Anzahl der Wortformen; *L* - die Anzahl der Lexeme; dazu kommen relative Indizes: *V/N* (oder *L/N*) - das Verhältnis des Inventars zum Textumfang (es repräsentiert den relativen "Reichtum" oder die "Diversität" des Wörterbuchs); das inverse Verhältnis *N/V* (oder *N/L*) drückt die mittlere Worthäufigkeit (Wortwiederholung) im gegebenen Text aus.

Bei der Verwendung der erwähnten relativen statistischen Indizes ist zu beachten, daß sie vom Textumfang, aber auch vom Texttyp abhängen.

Als Beispiel führen wir die Daten des Häufigkeitswörterbuchs der Autorensprache zeitgenössischer estnischer künstlerischer Prosa an:

	N	*V*	*L*	*V/N*	*N/V*	*L/N*	*N/L*
Teilstichprobe	5000	2690	1953	0,54	1,9	0,39	2,6
Textkorpus	99898	30733	14654	0,31	3,3	0,15	6,8

Man kann feststellen, daß die relative Diversität der Lexik (das Verhältnis *V/N* oder

L/N) bei Vergrößerung des Textumfangs sinkt, während die mittlere Wortfrequenz (*N/V* oder *N/L*) wächst. Vergleicht man das Textkorpus von Autorensprache wissenschaftlicher Prosa (das aus 20 Teilstichproben zu je 5000 laufenden Wörtern aus Texten verschiedener Autoren besteht) mit einem Einzeltext in Autorensprache (hier ein Roman von A.H. Tammsaare, vgl. Villup, 1978), dann zeigt sich, daß sich die Werte aller relativen Indizes im Einzeltext signifikant von den entsprechenden Indizes für ein Textkorpus etwa gleichen Umfangs unterscheiden:

	N	*V*	*L*	*V/N*	*N/V*	*L/N*	*N/L*
Roman v. A.H. Tammsaare	114124	16750	7348	0,15	6,8	0,64	15,5

Besonders fällt der Unterschied in den Werten von *N/V* (und *N/L*) ins Auge: im Textkorpus (das aus Texten verschiedener Autoren besteht) ist die mittlere Wortfrequenz geringer und dementsprechend die Diversität größer als in dem Einzeltext.

Im Vergleich von Daten aus verschiedenen Sprachen (siehe Tabelle 2.1) wird ein typologischer Unterschied zwischen den Sprachen deutlich. Während zum Beispiel im synthetischen estnischen Text (des Umfangs von ungefähr 100000 laufenden Wörtern) die mittlere Wortformenfrequenz 3,3 beträgt, schwankt im analytischen englischen Text (des gleichen Umfangs) die mittlere Häufigkeit von Wortformen zwischen 7,4 (in einem allgemeinen literarischen Text) bis 12,7 (in einem wissenschaftlich-technischen Text). Bei funktioneller Homogenität der Texte und typologischer Nähe der Sprachen erweist sich die mittlere Frequenz der Wortformen als ähnlich, zum Beispiel in Texten über Elektronik in den analytischen Sprachen Englisch und Französisch (*N/V* = 12,7 bzw. 12,3). Der Einfluß des Funktionalstils zeigt sich beim Vergleich von Daten synthetischer Sprachen: Estnisch (künstlerische Texte - *N/V* = 3,3) und Kasachisch (Kinderliteratur - *N/V* = 4,2).

Das Verhältnis Wortform - Lexem

Ein wichtiger quantitativ-typologischer Index für eine Sprache ist das numerische Verhältnis *L/V* (oder *V/L*), das heißt die Proportion aus der Zahl verschiedener Lexeme zur Zahl der verschiedenen Wortformen (oder umgekehrt) im Text. Dieses Verhältnis kennzeichnet bekanntlich die morphologische Struktur einer Sprache vom quantitativen Standpunkt aus und erlaubt die Beurteilung des Grades von Analytismus/Synthetismus einer Sprache. Je größer das numerische Verhältnis *L/V*, desto analytischer ist die Sprache eines gegebenen Textes, da in diesem Fall die Anzahl der verschiedenen Lexeme sich der Zahl der verschiedenen Wortformen annähert und folglich im Text durchschnittlich weniger Flexionsformen je Lexem erscheinen. Umgekehrt, ein größeres numerisches Verhältnis *V/L* verrät, daß im Text auf jedes Lexem im Mittel mehrere Flexionsformen entfallen, und die Sprache

Tabelle 2.1
Fundamentale statistische Eigenschaften von Häufigkeitswörterbüchern verschiedener Sprachen (Alekseev, 1968; Bektaev, 1978; Grigori'eva, 1981; Häufigkeitswörterbuch der zeitgenössischen ukrainischen künstlerischen Prosa 1969; Kučera, Francis, 1967; Latviešu val. biež. vārdn., 1972).

Sprache	*N*	*V*	*L*	*V/N*	*N/V*	*L/N*	*N/L*
Estnisch (künstl. Prosa)	99898	30733	14654	0,31	3,3	0,15	6,8
Ukrainisch (künstl. Prosa)	100000	27570	13954	0,28	3,6	0,14	7,2
Lettisch (künstl. Prosa)	100000	-	11439	-	-	0,11	8.7
Kasachisch (Kinderlit.)	98040	23350	10076	0,24	4,2	0,10	9,7
Russisch (Briefe)	96800	15842	8064	0,16	6,1	0,08	12,0
Deutsch (wiss.-techn.)	100000	14434	-	0,14	6,9	-	-
Englisch (allgem. literarische Texte)	101566	13706	-	0,13	7,4	-	-
Englisch (Elektronik)	100000	7853	5197	0,079	12,7	0,052	19,2
Französisch (Elektronik)	100000	8108	4527	0,081	12,3	0,046	21,9

eines solchen Textes muß man als synthetischer kennzeichnen. Allerdings ist dabei zu beachten, daß die quantitativen Maße für Analytismus/Synthetismus sensitiv auf Veränderungen des Textumfanges N reagieren. Die Erfahrung zeigt, daß in dem Maße, wie N erhöht wird, das Verhältnis L/V (in bestimmten Grenzen) immer weiter sinkt, während das Verhältnis V/L entsprechend steigt. Diese Tendenz kann man am besten anhand eines Einzeltextes illustrieren (siehe Tabelle 2.2). Allerdings wird sich bei der Vergrößerung des Textumfangs bis hin zu einem großen Umfang mehr und mehr der Einfluß bestimmter Gesetzmäßigkeiten des Auftretens seltener Wörter zeigen: Beim Übergang auf Stichproben größeren Umfangs sind praktisch alle neu hinzukommenden Wörter seltene (siehe Frumkina, 1964). Gleichzeitig verringert sich die Geschwindigkeit, mit der sich das Verhältnis L/V verkleinert (beziehungsweise das Verhältnis V/L sich vergrößert). Wenn die Bedingungen der Datenerhebung gleichartig sind, erlaubt der Vergleich von Texten verschiedener Sprachen auf Grund des Verhältnisses L/V (oder V/L), analytischere von weniger analytischen Sprachen zu unterscheiden (Tabelle 2.3). Nun zeigt sich dabei, daß die Maße für Analytismus/Synthetismus nicht nur vom Textumfang und von der Sprache abhängen, sondern in bestimmtem Grade auch vom Funktionalstil. So ist zum Beispiel bei einem Textumfang von $N = 200000$ in der englischen Sprache der Analytismus-Koeffizient (das Verhältnis L/V) für Zeitungstext = 0,53, für wissenschaftlich-technischen Text dagegen 0,67. Man kann den Schluß ziehen, daß das numerische Maß für den Analytismus hauptsächlich vom Vokabularumfang abhängt: Im Zeitungstext ist die Größe des Vokabulars $L \approx 12000$, während in dem wissenschaftlich-technischen Text gleicher Länge die Anzahl verschiedener Lexeme nur etwa 7000 beträgt. Es ist klar, daß das Verhältnis der Zahl der Lexeme zur Zahl der

Wortformen bei einem solchen Unterschied im Vokabularumfang anders ausfällt und die Abhängigkeit vom Textumfang lediglich mittelbar ist, das heißt in dem Maße, in dem die Vokabulargröße vom Umfang einer bestimmten Textsorte abhängt.

Da das numerische Maß für Analytismus/Synthetismus in erster Linie durch den Umfang des Vokabulars eines Textes bestimmt wird, ist es zweckmäßig, den analytischen Zusammenhang zwischen der Anzahl von Wortformen V und der Anzahl von Lexemen L im Zusammenhang mit der Veränderung einer dieser Größen herauszufinden. Diese Frage hat nicht nur theoretische, sondern auch praktische Bedeutung. Insbesondere bei der Betrachtung von Wörterbüchern, die auf der Grundlage von Texten synthetischer Sprachen erstellt wurden, kann die Notwendigkeit entstehen, den Umfang des Lexeminventars aus der Kenntnis des Wortformeninventars vorherzusagen oder umgekehrt.

Tabelle 2.2
Die Veränderungsdynamik der Maße für Analytismus/Synthetismus in Daten der Autorensprache im ersten Band des estnischen Romans "Tôde ja ôigus" von A. H. Tammsaare

N	*V*	*L*	*L/V*	*V/L*
10000	3636	2114	0,58	1,72
20000	5944	3124	0,53	1,90
30000	7503	3781	0,50	1,98
114124	16750	7348	0,44	2,28

Wir gehen von der Voraussetzung aus, daß in gewissen Grenzen (unter Ausschluß eines sehr großen Textumfangs) im Prozeß der Rede- oder Texterzeugung ein gleichbleibendes Verhältnis zwischen der Geschwindigkeit des relativen Inventarwachstums der Lexeme und der Geschwindigkeit des relativen Zuwachses an Wortformen existiert. Dies ist eine völlig plausible Annahme, die dem "Allometriegesetz" oder dem des "konstanten relativen Wachstums" entspricht. Mathematisch läßt sich dieses Gesetz in Form einer Differentialgleichung ausdrücken:

$$\frac{dy/y}{dx/x} = b. \tag{2.1}$$

Diese Formel läßt sich auch folgendermaßen schreiben:

$$\frac{dy}{y} = b\,\frac{dx}{x}. \tag{2.2}$$

Tabelle 2.3
Maße für den Analytismus/Synthetismus verschiedener Sprachen
(die Daten wurden entnommen aus: Alekseev, 1975; Bektaev, 1978; Jablonskaja, 1976; Engwall 1974).

Sprache	*N*	*V*	*L*	*L/V*	*V/L*
Russisch (Elektronik)	200000	21648	6816	0,32	3,18
Rumänisch (Elektronik)	200000	14292	5708	0,40	2,50
Deutsch (Medizin)	200000	41041	20367	0,50	2,02
Französisch (künstl. Literatur)	200000	20531	10868	0,53	1,89
Englisch (Zeitungen)	200000	23595	12588	0,53	1,87
Englisch (Elektronik)	200000	10582	7160	0,67	1,48

Durch Integrieren erhalten wir

(2.3) $$\ln y = A + b \ln x,$$

woraus sich ein linearer Zusammenhang zwischen *ln y* und *ln x* ergibt. Setzt man $A = \ln a$, ergibt sich ein Potenzgesetz, die allometrische Wachstumsfunktion

(2.4) $$y = ax^b,$$

mit *a* und *b* als Parameter. Setzen wir für *y V* ein und für *x L*, können wir eine Formel aufstellen, die die Abhängigkeit zwischen der Zahl der verschiedenen Wortformen und der Zahl der verschiedenen Lexeme in einem Text ausdrückt und folgende Form hat:

(2.5) $$L = aV^b,$$

wo *a* und *b* wieder Parameter sind. Tests ergeben eine gute Übereinstimmung zwischen empirischen (beobachteten) und theoretischen (erwarteten) Größen auf Grund der Formel 2.5 für Daten des Estnischen, Russischen und einiger anderer Sprachen (siehe auch Tuldava, 1995, Kap. 9; zu anderen Ansätzen zur Analyse der Entsprechung zwischen Wortformen und Lexemen siehe Orlov, 1978; Nešitoj, 1975). Unter der Bedingung, daß das Datenmaterial homogen ist, kann man die Anzahl von Lexemen *L* bei vorgegebener Zahl von Wortformen *V* durch Anwenden der Formel 2.5 erfolgreich vorhersagen. Zum Beispiel ergibt sich auf der Grundlage der ersten beiden Stichproben von A. H. Tammsaare (Tabelle 2.2):

$N_1 = 10000$ $V_1 = 3636$ $L_1 = 2114$
$N_2 = 20000$ $V_2 = 5944$ $L_2 = 3124$

Die Berechnung der Parameter ergibt: $a = 3{,}1$; $b = 0{,}8$[1]. Die Vorhersage für den ganzen Roman (mit $N = 114124$ und $V = 16750$) ist nach Formel (2.5): $L = 3{,}1 \times 16750^{0{,}8} = 7423$, also sehr nah an der beobachteten Lexemzahl $L = 7348$.

Der Parameter *b* in (2.5) hat eine inhaltliche Bedeutung und repräsentiert eine reale Interdependenz zwischen *V* und *L*: wenn zum Beispiel $b = 0{,}8$, so bedeutet das, daß bei Erhöhung von *V* um 1% *L* durchschnittlich um 0,8% wächst. In diesem Sinne kann der Parameter *b* als Index für den Analytismus/Synthetismus einer Sprache interpretiert werden: offensichtlich ist bei einem großen Wert von *b* die Sprache eines gegebenen Textes analytischer (bei Erhöhung der Anzahl verschiedener Wortformen wächst die Zahl der Lexeme schneller). Man darf nicht aus dem Auge verlieren, daß beim Vergleich von Sprachen oder Texten die Beobachtungsbedingungen identisch gehalten werden müssen; insbesondere ist es wichtig, auf welche Weise die Zählung von Lexemen operationalisiert wird (zählt die Gesamtheit aller lexikalisch-semantischen Varianten als ein Lexem oder werden sie einzeln gezählt?). Man muß auch mit einem Einfluß des Funktionalstils rechnen. Nach vorläufiger Erfahrung bleiben die Parameter *a* und *b* in einem homogenen Textkorpus in den Grenzen $500 < N < 200.000$ hinreichend stabil.

In der Praxis kann die Notwendigkeit entstehen, die umgekehrte Relation festzustellen, das heißt die Abhängigkeit von *V* von *L*. In diesem Fall muß die Formel (2.5) transformiert werden, so daß sich

(2.6) $$V = \alpha L^{\beta}$$

ergibt, wo $\alpha = e^{-(\ln a/b)}$ und $\beta = 1/b$.

Schließlich kann man die Frage nach dem Zusammenhang zwischen dem Analytismus/Synthetismus-Index, das heißt dem Verhältnis *L/V* (oder *V/L*), und dem Textumfang *N* stellen. Zwischen ihnen besteht eine Korrelation; wie schon gesagt wurde, spielt der Vokabularumfang eines Textes eine bestimmende Rolle beim Aufstellen eines Maßes für den Analytismus/Synthetismus (wobei bei gleichartigen Textgrößen die Größen ihrer Vokabulare stark in Abhängigkeit von den Eigenarten des Funktionalstils oder des Individualstils variieren können). Das Verhältnis *L/V* sinkt im Mittel im Zusammenhang mit der Vergrößerung von *N* (siehe Tabelle 2.2). Es zeigt sich, daß der Zusammenhang zwischen diesen Größen nicht linear ist, sondern ebenfalls den Charakter einer Potenzfunktion besitzt, die mit der Formel

(2.7) $$L/V = cN^{-d}$$

[1] Die Parameter *a* und *b* kann man nach Linearisierung mit Hilfe der Methode der kleinsten Fehlerquadrate bestimmen (einen linearen Zusammenhang erhält man durch Logarithmieren der Größen *L* und *V*).

ausgedrückt werden kann (c und d sind Parameter).

Auf diese Weise erlaubt die Kenntnis der gesetzmäßigen Dynamik der Interdependenz zwischen Wortformen und Lexemen, die Größen V und L auseinander abzuleiten und innerhalb bestimmter Grenzen das Maß für Analytismus/Synthetismus für verschiedene Werte von N numerisch zu bestimmen. All dies kann bei der typologischen Untersuchung von Sprachen und bei der Lösung einiger Probleme der automatischen Textverarbeitung von Bedeutung sein.

Verteilung und Frequenzzonen des Vokabulars

Große Bedeutung für die quantitativ-linguistische Untersuchung der Lexik hat die Bestimmung des Verteilungstyps (Verteilungsgesetzes) für die Worthäufigkeiten in der "Horizontalen", also über die Teilstichproben aus den Daten der Gesamtheit (vgl. Kapitel 1.3).

Die Frage nach dem Verteilungstyp ist in der Literatur häufig erörtert worden (siehe z.B. Herdan, 1964; Bektaev, Luk'janenkov, 1971; Piotrovskij, Turygina, 1971; Kaširina, 1974; Jakubajtis, 1981). Vor allem hat sich herausgestellt, daß man auf der Grundlage des erkannten Verteilungstyps semantisch dominierende Wörter (Schlüsselwörter, Termini) von neutralen Texteinheiten unterscheiden kann ("Bektaev-Effekt"). Aber man muß dabei beachten, daß die Bestimmung des Verteilungstyps sehr von den Bedingungen der Untersuchung abhängt (z.B. von der Größe der Teilstichproben); die Veränderung von Untersuchungsbedingungen kann bewirken, daß an die Stelle des einen Verteilungstyps ein anderer tritt (Bektaev, Luk'janenkov, 1971:104). Nach den Daten einer Reihe von Untersuchungen folgt die Mehrzahl der gebräuchlichen hochfrequenten und mittelfrequenten Wörter der Normal- oder der Poissonverteilung bei Einzelstichproben eines Umfangs von nicht weniger als 2000 laufenden Wörtern. Die seltenen lexikalischen Einheiten folgen unabhängig von Stückelung und Normierung der Stichprobe der Poissonverteilung (Piotrowski, 1984). Die Tatsache, daß die Verteilung der Worthäufigkeiten der Normal- oder Poissonverteilung folgt, besagt, daß der betreffende Text in bezug auf das Verhältnis zu diesen Wörtern statistisch homogen ist. Das schließt jedoch nicht die Möglichkeit des Einflusses individueller Unterschiede in der Menge der Häufigkeitsdaten aus. Insbesondere ist es möglich, beim Vergleich von Individualstilen auf Grund von Abweichungen von der mittleren Häufigkeit sogenannte positive oder negative Wörter auszumachen (siehe z.B. Muller, 1968:87; McKinnon, 1980).

Wie schon oben erwähnt, ist eine der wichtigsten Aufgaben der Erstellung von Häufigkeitswörterbüchern die Auffindung von Häufigkeitszonen der Lexik. Die Schichtung der Lexik nach Häufigkeit hat nicht nur große Bedeutung für wissenschaftliche Untersuchungen im Bereich der Linguistik (der Lexikologie und der Lexikostatistik), sondern auch für die Pädagogik und die Psychologie (z.B. für das

Zusammenstellen von Minimalvokabularen, die Messung der Textschwierigkeit von Lehrbüchern, die Durchführung psychologischer Versuche), und auch für Aufgaben der automatischen Textverarbeitung (vor allem für die Effektivierung des automatischen Zugriffs auf maschinenlesbare Wörterbüchern). Wichtig ist anzumerken, daß die quantitativen Eigenschaften von Wörtern in Abhängigkeit ihrer Zugehörigkeit zu einer der Häufigkeitszonen stark mit qualitativen Eigenschaften dieser Wörter korreliert, z.B. mit ihrer Neutralität (Gebräuchlichkeit), Thematizität (Schlüsselwörter), Informativität und anderen.

Es gibt zahlreiche Versuche einer formalen statistischen Zerlegung der Lexik in diskrete Frequenzzonen. Als Kriterium für die Bestimmung der Zonen werden am häufigsten Besonderheiten der Frequenzverteilung der Wörter herangezogen. So werden z.B. Frequenzzonen in Abhängigkeit von der Schwankung der Parameter des Zipfschen Gesetzes bei der Rangverteilung von Wörtern bestimmt (Gor'kova, 1969), in Abhängigkeit vom maximalen Krümmungspunkt der Integralverteilung von Weibull (Petrenko, 1974) oder anhand der Veränderung des (Rang-)Verteilungstyps der Wörter (Martynenko, 1978). Daneben gibt es noch eine Reihe weiterer Methoden zur formalen Zerlegung von Häufigkeitswörterbüchern in Zonen (siehe z.B. Maršakova, 1974; Malachovskij, 1980; Billmeier, 1968; Pao, 1978). Man kann offensichtlich die Bestimmung von Frequenzzonen auf verschiedene Weisen begründen und sie in Abhängigkeit von den Zielen und Aufgaben der jeweiligen Untersuchung wählen.

2.2. Die Häufigkeitsstruktur des Textes

Der Begriff der Häufigkeitsstruktur

Wenn man von den konkreten lexikalischen Einheiten abstrahiert, die ein Häufigkeitswörterbuch konstituieren, und lediglich die Häufigkeiten F_i und Ränge i der lexikalischen Einheiten betrachtet, dann erhält man die sogenannte Rangverteilung der Häufigkeiten oder kurz Rang-Frequenz-Verteilung. Eine andere Möglichkeit der formalen Analyse von Häufigkeitswörterbüchern ist die Gegenüberstellung der Häufigkeiten F_i mit der Anzahl von Einheiten, die die jeweilige Frequenz besitzen - $m(F_i)$ -, woraus sich die spektrale Frequenzverteilung oder kurz das Frequenzspektrum der Lexik ergibt. Die gemeinsame Betrachtung der Rangverteilung und des Spektrums der lexikalischen Einheiten eröffnet uns die Häufigkeitsstruktur des Textvokabulars, die sich als ein bestimmter Teilaspekt der allgemeinen statistischen Textorganisation darstellt (also der Gesamtproblematik der quantitativen Struktur- und Funktionsanalyse eines Textes und des entsprechenden Vokabulars).

Die so bestimmte Häufigkeitsstruktur eines Textes kann man anschaulich in folgender kompakter Form darstellen:

i	F_i	$m(F_i)$
1	F_1	$m(F_1)$
2	F_2	$m(F_2)$
...	...	...
k-n	$F_{k\text{-}n}$	$m(F_{k\text{-}n})$
...	...	...

Tabelle 2.4 zeigt integriert die Rang-, Häufigkeits- und Klassenbelegungsverteilung nach dem eben beschriebenen Schema auf der Grundlage der Daten eines Häufigkeitswörterbuchs der Lexeme der Autorensprache zeitgenössischer estnischer künstlerischer Prosa (Kaasik et al., 1977). Neben diesen rangierten Größen gehört zu den Ausgangsdaten bei der Betrachtung der Häufigkeitsstruktur eines Texts auch der Textumfang (N) und der Umfang des entsprechenden Vokabulars (L). Im vorliegenden Fall beträgt $N = 99898$ (Wortformen-Tokens) und $L = i_{max} = 14654$ (Lexeme).

Verteilungen des Rangs und des Spektrums können differentielle (nicht-kumulative) Form besitzen oder in integrierender Form (kumulativ, d.h. mit klassenweise aufsummierten Häufigkeiten bzw. Klassengrößen) gegeben sein.[2] Die Integralform der Verteilung drückt den Abdeckungsgrad aus, also mit welcher Größe eines Textfragments es möglich ist, den Grad der Konzentration von lexikalischen Einheiten in bestimmten Teilstücken des Häufigkeitswörterbuchs zu beurteilen. Natürlich können die Häufigkeiten und die Klassengrößen in absoluten Zahlen angegeben werden oder auch als relative Größen (z.B. als Prozentzahlen). Als Beispiel bringen wir hier Daten zur Rangverteilung der Häufigkeiten von Wortformen in abgekürzter Form zusammen mit Daten zur Textfragmentgröße in absoluten und relativen Zahlen (Tabelle 2.5).

Zunächst betrachten wir die eine der Seiten der Häufigkeitsstruktur - die Rangverteilung - und die Möglichkeiten ihrer analytischen Beschreibung mit Hilfe verschiedener Versionen des Zipfschen Gesetzes.

Rangverteilung und das Zipfsche Gesetz

Eine der wichtigsten Gesetzmäßigkeiten, die man bei der quantitativen Analyse von Texten feststellt, ist der statistische Zusammenhang zwischen der Frequenz und dem Rang von Wörtern. Dabei zeigt sich, daß, obwohl die Wörter in verschie-

[2] Entsprechend unterscheidet man die differentielle Funktion (Dichte) und die Integralfunktion der Verteilung (vgl. Mitropol'skij, 1971:209ff). Ähnlich wird zwischen einer Differentialgleichung und einer Integralgleichung (d.h. der Lösung der Differentialgleichung) unterschieden.

Tabelle 2.4. Häufigkeitsstruktur eines Textes (Häufigkeitswörterbuch der Lexeme der Autorenrede in estnischer künstlerischer Prosa (i = Rang, F_i = Häufigkeit, $m(F_i)$ = Klassengröße). Textumfang $N = 99\,898$; Vokabularumfang $L = 14.654$; $F_{max} = 4.237$.

i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$
1	4237	1	41	264	1	80	153	1	132	96	1	239-241	57	3	587-612	22	26
2	3493	1	42	261	1	81	151	1	133-135	94	3	242-244	56	3	613-650	21	38
3	2598	1	43	260	1	82	150	1	136-139	93	4	245-247	55	3	651-676	20	26
4	1981	1	44	254	1	83	149	1	140-141	92	2	248-252	54	5	677-707	19	31
5	1395	1	45	248	1	84	148	1	142-144	91	3	253-256	53	4	708-747	18	40
6	1300	1	46	238	1	85	146	1	145-147	90	3	257-263	52	7	748-793	17	46
7	1047	1	47	237	1	86	144	1	148-149	89	2	264-268	51	5	794-856	16	63
8	879	1	48	234	1	87	139	1	150	88	1	269-270	50	2	857-893	15	37
9	845	1	49	230	1	88	138	1	151	87	1	271-274	49	4	894-940	14	47
10	827	1	50	223	1	89	136	1	152-154	86	3	275-280	48	6	941-1017	13	77
11	724	1	51	218	1	90	135	1	155-157	85	3	281-285	47	5	1018-1080	12	63
12	634	1	52	209	1	91-92	133	2	158-159	84	2	286-289	46	4	1081-1169	11	89
13	613	1	53	207	1	93-94	131	2	160	83	1	290-299	45	10	1170-1293	10	124
14	581	1	54	206	1	95-96	130	2	161-162	82	2	300-305	44	6	1294-1414	9	121
15	568	1	55	200	1	97	126	1	163-164	81	2	306-312	43	7	1415-1567	8	153
16	499	1	56	198	1	98-99	125	2	165-169	79	5	313-318	42	6	1568-1779	7	212
17	496	1	57	195	1	100	124	1	170-172	78	3	319-329	41	11	1780-2044	6	265
18	493	1	58	194	1	101	123	1	173-179	76	7	330-339	40	10	2045-2389	5	345
19	465	1	59	192	1	102-103	122	2	180-183	75	4	340-346	39	7	2390-2980	4	591
20	448	1	60	190	1	104	120	1	184-187	74	4	347-349	38	3	2981-3918	3	938
21	436	1	61	189	1	105	116	1	188-191	73	4	350-360	37	11	3919-5972	2	2054
22	434	1	62-63	185	1	106	113	1	192-194	72	3	361-369	36	9	5973-14654	1	8682
23	428	1	64	181	1	107-108	112	2	195-202	71	8	370-382	35	13			
24	382	1	65	180	1	109-110	110	2	203-204	70	2	383-389	34	7			
25	373	1	66	179	1	111	109	1	205-206	69	2	390-404	33	15			
26	366	1	67	176	1	112	108	1	207	68	1	405-412	32	8			
27	350	1	68	175	1	113-115	107	3	208-209	67	2	413-428	31	16			
28	339	1	69-70	174	2	116	106	1	210-211	66	2	429-445	30	17			
29	327	1	71	171	1	117-119	105	3	212-217	64	6	446-466	29	21			
30	309	1	72	167	1	120	104	1	218-220	63	3	467-479	28	13			
31-33	304	3	73	166	1	121-124	101	4	221	62	1	480-509	27	30			
34	297	1	74	165	1	125-126	100	2	222-223	61	2	510-529	26	20			
35-37	285	3	75-76	164	2	127	99	1	224-229	60	6	530-543	25	14			
38-39	272	2	77-78	163	2	128	98	1	230-235	59	6	544-564	24	21			
40	268	1	79	159	1	129-131	97	3	236-238	58	3	565-586	23	22			

Tabelle 2.5
Rangverteilung der Wortformen-Häufigkeiten in der Autorensprache estnischer künstlerischer Prosa (i = Rang, F_i = Häufigkeit, F^*_i = kumulative Häufigkeit, P^*_i (%) = Textfragmentgröße); N = 99898; V = 30733; F_{max} = 3221

i	F_i	F_i^*	$P_i^*(\%)$	i	F_i	F_i^*	$P_i^*(\%)$
1	3221	3221	3,22	200	48	35066	35,10
2	1602	4823	4,83	300	34	38592	38,63
3	1439	6262	6,27	400	25	41484	41,53
4	1375	7637	7,64	500	21	43755	43,80
5	1264	8901	8,91	600	17	45634	45,68
6	1116	10017	10,03	700	15	47264	47,31
7	995	11012	11,02	800	13	48703	48,75
8	713	11725	11,74	900	12	49966	50,02
9	592	12317	12,33	1000	11	51109	51,16
10	542	12859	12,87	1001-1034	11	51483	51,54
20	329	16801	16,82	1035-1168	10	52823	52,88
30	224	19344	19,36	1169-1299	9	54002	54,06
40	189	21384	21,41	1300-1500	8	55610	55,67
50	165	23130	23,15	1501-1753	7	57381	57,44
60	141	24635	24,66	1754-2131	6	59649	59,71
70	120	25920	25,95	2132-2650	5	62244	62,31
80	108	27045	27,07	2651-3460	4	65484	65,55
90	89	28017	28,05	3461-5088	3	70368	70,44
100	83	28871	28,90	5089-8973	2	78138	78,22
				8974-30733	1	99898	100,00

denen Texten unterschiedliche Ränge besitzen können, die Form der Verteilung selbst konstant ist, d.h. das Aussehen der Gesetzmäßigkeit im Ganzen. Hier spricht man vom "topologischen" Prinzip, nach dem "wichtig nicht 'die Metrik' [...], sondern die Bewahrung des 'Musters' ist, dessen Äußeres variieren kann, das aber mit sich selbst identisch bleibt und mit verschiedenem Inhalt gefüllt werden kann" (Bernštejn, 1966:65). In allen Fällen, in denen wir es mit Texten einer natürlichen Sprache zu tun haben, zeigt sich der sogenannte Effekt der Konzentration und der Dispersion, der darin besteht, daß es eine kleine Gruppe von sehr häufigen Wörtern gibt (den "Kern" des Häufigkeitswörterbuchs) und eine große Gruppe von seltenen Wörtern (den "Schweif" des Häufigkeitswörterbuchs); zwischen diesen beobachtet man einen fließenden Übergang ("den Bereich der mittelhäufigen Wörter"). In graphischer Darstellung erinnert dies an eine Hyperbel (siehe z.B. Abbildung 2.1). Diese Ungleichverteilung zeigt sich nicht nur in bezug auf Wörter,

sondern auch in bezug auf andere sprachliche Einheiten (Buchstaben, Phoneme, Morpheme, Syntagmen usw.). Darüber hinaus wird eine analoge Verteilungsform auch in vielen weiteren Bereichen der menschlichen Tätigkeit angetroffen (Informatik, Ökologie, Demographie u.a.), was diesen Verteilungstyp als universelles semiologisches "Präferenzgesetz" (Perebijnis, 1970) zu betrachten verlangt oder als "Verteilungsgesetz der Einheiten nach Wichtigkeit" (Martynenko, 1978).

Für den analytischen Ausdruck der Abhängigkeit zwischen der Häufigkeit und dem Rang eines Wortes bieten sich eine Reihe von Formeln an, die Spielarten des Zipfschen Gesetzes darstellen (Zipf, 1935, 1949; Mandelbrot, 1954; Orlov, 1976; Alekseev 1978; Krylov, 1982; und andere). Die Grundform des Zipfschen Gesetzes wird mit der folgenden Formel dargestellt, die eine Potenzfunktion mit negativem Exponenten ist:

(2.8) $$F_i = Ci^{-\gamma} \qquad \text{oder} \qquad p_i = ki^{-\gamma},$$

wo F_i die absolute Häufigkeit, P_i die relative Häufigkeit (Wahrscheinlichkeit), i der Rang, C, k und γ Parameter der Verteilung sind (wobei $P_i = F_i/N$ und $C = kN$, mit N als Textumfang). In dem häufig auftretenden Fall, daß $\gamma = 1$, nimmt die Formel das Aussehen der "klassischen einparametrigen Zipfschen Verteilung" an:

(2.9) $$F_i = Ci^{-1} \qquad \text{oder} \qquad p_i = ki^{-1}.$$

Was ist nun die substantielle Bedeutung des Zipfschen Gesetzes? Um den Mechanismus der Variablenveränderung aufzudecken, betrachten wir die den Funktionen (2.8) entsprechende Differentialgleichung:

(2.10) $$\frac{dF_i/F_i}{di/i} = -\gamma.$$

Aus der Gleichung ist ersichtlich, daß die relativen Veränderungen von F_i und i im Verhältnis konstant sind. Eben dies ist das Charakteristikum auch des Gesetzes des konstanten relativen Wachstums (siehe Formel 2.1), nur mit dem Unterschied, daß im gegebenen Fall die Konstante einen negativen Wert hat und das Gesetz die "konstante relative Abnahme" der Funktion ausdrückt. Dieses Gesetz ist in vielen Bereichen der Wissenschaft bekannt. Es zeigt sich, daß das Zipfsche Gesetz formal mit einem bestimmten universellen Gesetz zusammenfällt, das einen weiten Bereich von Erscheinungen der materiellen Welt umfaßt. Im vorliegenden Fall nimmt die Häufigkeit F_i mit einer dem Wachstum des Vokabulars proportionalen Geschwindigkeit ab, die auf spezielle Weise - nämlich anhand des Rangs i - gemessen wird.

Die Rangverteilung, die den Funktionen (2.8) und (2.9) entspricht, besitzt in graphischer Darstellung die Form einer Hyperbel, während sich in bilogarithmi-

scher Darstellung eine Gerade ergibt; d.h. zwischen *ln* F_i und *ln i* wird eine lineare Abhängigkeit sichtbar. Dies bestätigt sich zum Beispiel anhand des Häufigkeitswörterbuchs der Wortformen des Estnischen auf der Grundlage von Texten mit einem Gesamtumfang von ungefähr 100000 laufenden Wörtern (siehe Abbildung 2.2 nach Daten der Tabelle 2.5).

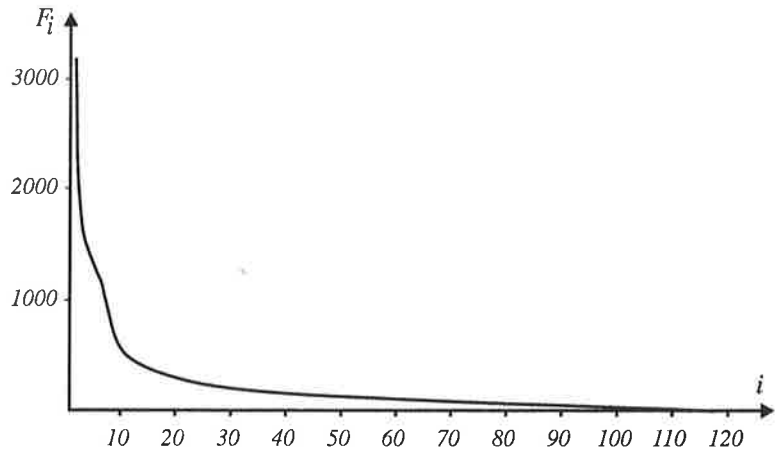


Abbildung 2.1. Der Zusammenhang zwischen der Häufigkeit F_i und dem Rang *i* nach Daten des Häufigkeitswörterbuchs von Wortformen im Estnischen

Bekanntlich zeigen sich typologische Unterschiede zwischen Sprachen besonders klar beim Vergleich der Rangverteilungen der Wortformen. Vergleichen wir z.B. die Daten des Häufigkeitswörterbuchs der Wortformen des Estnischen und des Englischen (bei gleichen Textumfängen), können wir feststellen, daß der Wert des Parameters γ sich wesentlich unterscheidet: bei der flektierend-synthetischen estnischen Sprache ist $\gamma = 0{,}86$, während er für die flektierend-analytische englische Sprache $\gamma = 1.002$ beträgt (Kučera, Francis, 1967:357), das heißt, daß der Steigungswinkel der Geraden in der Graphik (oder entsprechend der Tangens des Winkels γ) für das Englische signifikant größer ist als für das Estnische (siehe Abbildung 2.3). Das bedeutet, daß die hochfrequenten Wortformen (insbesondere die Funktionswörter) in der englischen Sprache einen größeren Teil des Textes abdecken als im Estnischen und die Sättigung des Vokabulars mit Wortformen im englischen Text schneller erfolgt. Als Resultat dieser linguistisch-typologischen Differenzen gibt es im estnischen Text mit 100000 laufenden Wörtern etwa 30000 Wortformen, während im englischen Text des gleichen Umfangs ihre Anzahl 15000 nicht überschreitet.

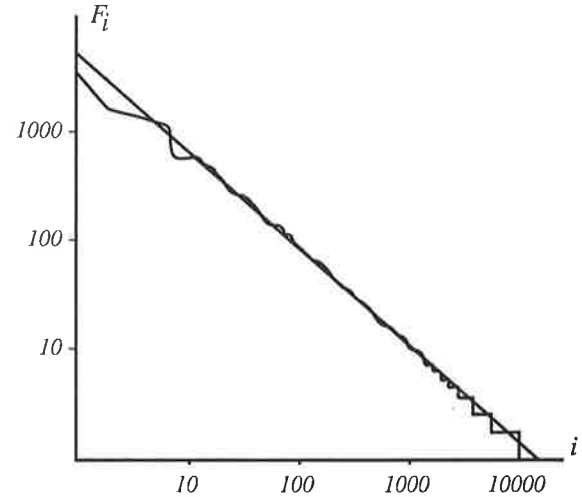


Abbildung 2.2. Der Zusammenhang zwischen F_i und *i* nach dem Häufigkeitswörterbuch von Wortformen im Estnischen. Bilogarithmische Skalierung.

Eine eingehendere Analyse zeigt, daß die exakte lineare Abhängigkeit zwischen der Häufigkeit und dem Rang (in bilogarithmischen Koordinaten) nicht auf der ganzen Ausdehnung der Rangverteilung eingehalten wird. In vielen Sprachen erscheint eine Abweichung von der linearen Abhängigkeit in der Rangverteilung der sprachlichen Einheiten im Bereich der großen Häufigkeiten (in der Kernzone des Häufigkeitswörterbuchs). Um eine bessere Übereinstimmung zwischen den empirischen und den theoretischen Daten zu erreichen, verwendet man gewöhnlich eine Version des Zipfschen Gesetzes, die die sogenannte Mandelbrotsche Korrektur enthält (Mandelbrot, 1954). Diese Version heißt Zipfs "Kanonisches Gesetz" oder "das Zipf-Mandelbrotsche Gesetz":

$$(2.11) \qquad F_i = C(i + B)^{-\gamma} \qquad \text{oder} \qquad p_i = k(i + B)^{-\gamma},$$

wo *B* die Mandelbrotsche Korrektur ist. In vielen Fällen, wenn $\gamma = 1$, hat die Formel folgende Gestalt:

$$(2.12) \qquad F_i = C(i + B)^{-1} \qquad \text{oder} \qquad p_i = k(i + B)^{-1}.$$

Die Formel (2.11) ergibt eine gute Übereinstimmung zwischen theoretischen und empirischen Daten im ersten Teil des Häufigkeitswörterbuchs, wobei die Anwendung der Mandelbrotschen Korrektur zu einer leichten Erhöhung der theoretischen Werte der Konstanten *C* (oder *k*) und γ führt (siehe Tabelle 2.6).

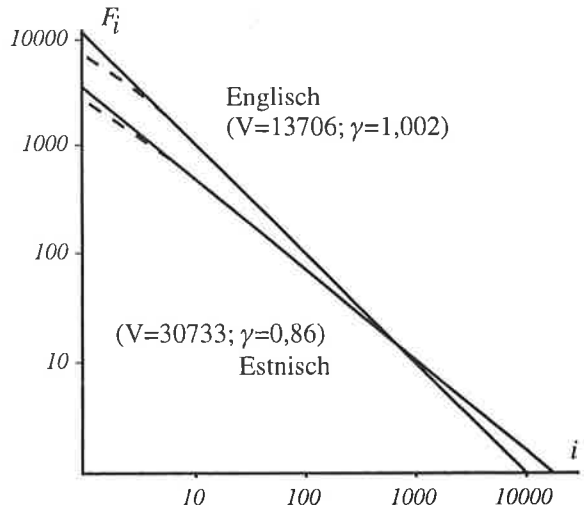


Abbildung 2.3. Rangverteilung: der Zusammenhang zwischen der Frequenz F_i und Rang i nach Daten des Häufigkeitswörterbuchs der Wortformen des Estnischen und des Englischen (Stichproben mit N = 100000 laufenden Wörtern; V: Vokabularumfang). Bilogarithmische Darstellung.

Außer der Abweichung von der (in bilogarithmischer Darstellung) linearen Abhängigkeit im Kopfteil des Häufigkeitswörterbuchs beobachtet man in vielen Fällen eine mehr oder weniger merkliche Abweichung vom linearen Zusammenhang im Bereich der geringen Häufigkeiten (im "Schweif" der Verteilung). In bezug auf die Daten des Wörterbuchs des estnischen Gesamtkorpus ist diese Abweichung gering (im Häufigkeitswörterbuch der Lexeme ist bei i = 30 ... 2500 der Tangens des Steigungswinkels γ = 0,98, während er bei i größer als 2500 γ = 1,03 beträgt). Die Abweichung im Bereich der kleinen Häufigkeiten ist deutlicher bei den Daten der Gesamtkorpus-Häufigkeitswörterbücher des Russischen (Tabelle 2.7), und ebenso bei den Daten von Einzeltext-Häufigkeitswörterbüchern des Estnischen (Tabelle 2.8). Zur Illustration bringen wir die Rangverteilung der Häufigkeiten von Lexemen zu den Daten des Häufigkeitswörterbuchs des Russischen (Tabelle 2.9) und eine entsprechende Graphik (Abbildung 2.5), auf der die Abweichungen im Anfangs- und im Endteil der Verteilung klar zu sehen sind.

Die Abweichungen von der (in bilogarithmischer Darstellung) linearen Abhängigkeit im ersten und letzten Teil der Rangverteilung verringert sich in der Regel mit wachsendem Umfang der Stichprobe (des Textes). Es läßt sich mit vollem Recht behaupten, daß die Häufigkeitsstruktur des Textes dynamisch und gesetzmäßig mit dem Textumfang variiert. Diese Dynamik der Häufigkeitsstruktur ist so-

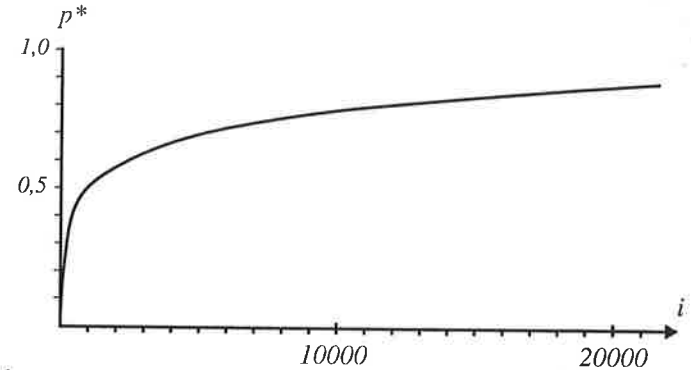


Abbildung 2.4. Der Zusammenhang zwischen der relativen kumulativen Häufigkeit P^*_i und dem Rang i nach Daten des Häufigkeitswörterbuchs von Wortformen des Estnischen.

Tabelle 2.6
Werte der Parameter C, k, γ, B der Zipf-Mandelbrot-Verteilung zu den Daten der Häufigkeitswörterbücher verschiedener Sprachen (eigenes Material; Bektaev, 1978, Kalinina, 1968, Kučera, Francis, 1967).

Sprache und Subsprache	Typ der Einheiten	Textumfang	Vokabularumfang	Parameter			
				C	k	γ	B
Estnisch (künstl. Prosa; Textkorpus	Lexeme	99898	14654	6793	0,068	0,92	-
				14785	0,148	1,04	2,3
	Wortformen	99898	30733	4095	0,041	0,86	-
				4595	0,046	0,87	0,5
Kasachisch (künstl. Prosa; Einzeltext)	Wortformen	105494	22642	6224	0,059	0,87	-
				8334	0,079	0,92	4,96
Russisch (Elektronik)	Wortformen	100000	14062	4200	0,042	0,81	-
				5400	0,052	0,84	1,9
Englisch (vermischte Texte)	Wortformen	101586	13706	13100	0,129	1,002	-
				14000	0,138	1,05	1,0

wohl in Bezug auf Gesamtkorpus- als auch auf Einzeltext-Häufigkeitswörterbücher zu beobachten.

Daher kann man die Abweichungen vom Zipfschen Gesetz in seiner Grundform als gesetzmäßige Erscheinung betrachten, wenn man den dynamischen Charakter der statistischen Textorganisation in Rechnung stellt. Es wird vorgeschlagen, als allgemeinste Bedingung für die Erfüllung des Zipfschen Gesetzes in seiner Rangform (jedenfalls für Texte nicht zu großen Umfangs) den linearen Zusammenhang zwischen $ln\ F_i$ und $ln\ i$ im mittleren Teil bei gleichzeitiger Abweichung im Anfangs- und Endteil der Verteilung anzusehen, und diese sowie die konkreten Werte des Parameters γ mit verschiedenen linguistischen Ursachen zu begründen (Sprachtyp, strukturelle Eigenheiten der Lexik, Wahl der Einheiten etc.; siehe Boroda, Polikarpov, 1984). Außerdem ist es möglich, anhand des Wendepunkts eine Segmentierung der Lexik in objektiv determinierte Frequenzzonen vorzunehmen.

Auf einer speziellen Sichtweise beharrt Martynenko (1978), nach dessen Ansicht die erwähnte Verteilung als Verteilung nichthomogener Elemente prinzipiell nicht mit einer einzelnen Funktion approximiert werden kann und insbesondere Kern- bzw. Peripherieelemente mit verschiedenen Gesetzen beschrieben werden müssen. Einen speziellen integrativen Ansatz schlägt Nešitoj (1984, 1986) vor, der das Problem der analytischen Beschreibung der verschiedenen Seiten der Häufigkeitsstruktur des Textes auf der Grundlage eines eigenen verallgemeinerten Systems stetiger Verteilungen löst.

Tabelle 2.7

Die Werte des Parameters γ in den verschiedenen Häufigkeitszonen. Daten der Häufigkeitswörterbücher des Estnischen und des Russischen (I Zasorina. 1966; II Häufigkeitswörterbuch der russischen Sprache, Zasorina 1977)

Frequenzzone (i)	Estnisch		Russisch (Lexeme)	
	Lexeme	Wortformen	I	II
1...30 (γ_1)	0,83	0,77	0,70	0,71
30...2500 (γ_2)	0,98	0,89	0,94	0,95
> 2500 (γ_3)	1,03	0,90	1,24	1,51
Gesamtvokabular (γ)	0,92	0,86	0,93	1,0
Vokabularumfang (L)	14654	30733	10830	39268
Textumfang (N)	99898	99898	120474	1056382

Tabelle 2.8

Werte des Parameters γ in verschiedenen Häufigkeitszonen. Daten der Häufigkeitswörterbücher der Lexeme des ersten Bandes des Romans 'Wahrheit und Recht' von A. H. Tammsaare (ausgewertet auf der Grundlage der Daten von: Villup, 1978). Estnisch.

Frequenzzone (i)	Gesamter Roman (1.Band)	Davon	
		Autorensprache	wörtliche Rede
1..30 (γ_1)	0,7	0,68	0,59
30...1500 (γ_2)	1,1	1,11	1,21
> 1500 (γ_3)	1,4	1,43	1,47
Gesamtvokabular (γ)	1,0	1,0	1,01
Vokabularumfang (L)	8228	7348	3135
Textumfang (N)	160356	114124	46232

Anmerkung: Die Werte der Parameter in den Tabellen 2.7 und 2.8 wurden aufgrund gemittelter Ränge nach der Methode der kleinsten Quadrate berechnet. Die gemittelten Ränge berechnen sich folgendermaßen: Alle aufeinanderfolgenden Ränge mit gleicher Frequenz werden durch den Mittelwert eben dieser Ränge ersetzt. Beispiel: im Häufigkeitswörterbuch der russischen Sprache (Tabelle 2.9) entsprechen der Häufigkeit $F = 1$ die Ränge 25890 .. 39268; der entsprechende gemittelte Rang ist 32600. Über den Vorteil der Berechnung der Parameter auf der Grundlage gemittelter Ränge siehe Kalinin, 1964: 125).

Neue Arbeiten zum Zipfschen Gesetz

Im Zusammenhang mit den von der (in bilogarithmischer Darstellung) linearen Abhängigkeit vorzufindenden Abweichungen zwischen den Variablen ist die These von Alekseev (1978) über den "nichtlinearen Charakter" der Rangverteilung lexikalischer Einheiten in langen Texten interessant. Darin wird vorgeschlagen, daß bei starker Vergrößerung des gegebenen Textmaterials viele seltene Wörter in die mittlere Schicht des Häufigkeitswörterbuchs geraten und eine Veränderung des Charakters der Abhängigkeit Rang - Häufigkeit entsteht: die Stufen in den niedrigeren Schichten verkleinern sich, wohingegen sie sich im mittleren Teil des Häufigkeitswörterbuchs verbreitern, wobei sie diesen Teil der theoretischen Grafik nach rechts verschieben und seine Schiefe vergrößern. Alekseev schlägt folgende Verallgemeinerung des Zipfschen Gesetzes vor:

Tabelle 2.9
Rangverteilung der Häufigkeiten der Lexeme. Häufigkeitswörterbuch der russischen Sprache (Zasorina, 1977); $N = 1056382$; $L = 39268$; $F_{max} = 42845$.

i	F_i	i	F_i	i (Mittelwert)	F_i
1	42854	80	1415	5000	22
2	36266	90	1210	6000	17
3	19228	100	1093	7000	14
4	17261	200	557	7300	13
5	13839	300	332	7700	12
6	13307	400	312	8200	11
7	13185	500	256	8750 (8451-9045)	10
8	13143	600	217	9400 (9046-9758)	9
9	12975	700	187	10200 (9759-10599)	8
10	10719	800	164	11000 (10600-11576)	7
15	6246	900	147	12200 (11577-12855)	6
20	5179	1000	134	13700 (12856-14536)	5
30	4156	1500	90	15650 (14537-16779)	4
40	2830	2000	67	18500 (18780-20143)	3
50	2136	2500	52	23000 (20144-25889)	2
60	1887	3000	42	32600 (25890-39268)	1
70	1655	4000	30		

(2.13) $$F_i = Ci^{-(\gamma + \varphi \ln i)} \quad \text{oder} \quad p_i = ki^{-(\gamma + \varphi \ln i)},$$

wo C (oder k), γ und φ Parameter sind. Diese Formel entspricht in logarithmischer Notation der Gleichung der parabolischen Regression. Alekseev nennt den Ausdruck 2.13 die "vierte Approximation" des Zipfschen Gesetzes, wobei er im Sinn hat, daß die erste Approximation die klassische einparametrige Form (Formel 2.9) ist, die zweite und dritte Approximation die Formeln 2.8 und 2.11 darstellen.

Die Formel 2.13 besitzt den Vorzug, daß sie die Grundformen der Zipfschen Verteilung als Spezialfälle enthält: bei $\varphi = 0$ verwandelt sich Formel 2.13 in Formel 2.8, aber wenn $\gamma = 1$, erhält man die einparametrige Formel 2.9. Eine Überprüfung ergibt, daß die Formel 2.13 die empirischen Daten gut approximiert, wenn Abweichungen von der linearen Abhängigkeit im Anfangs- und Endbereich der Verteilung gegeben sind. Die Frage ist nur, ob Formel 2.13 die wirkliche Textstruktur (die parabolische Abhängigkeit zwischen F_i und i in bilogarithmischen Koordinaten) reflektiert oder ob sie lediglich tatsächlich gesetzmäßige Abweichungen im Anfangs- und Endbereich der Verteilung verschleiert und nicht berücksich-

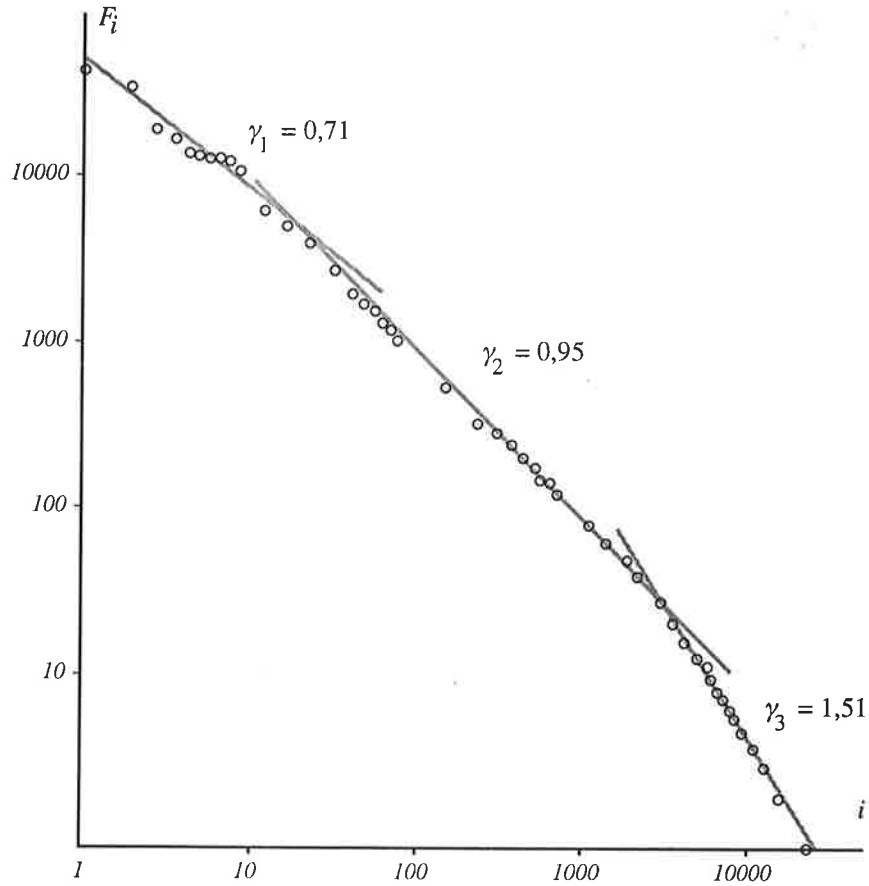


Abb. 2.5. Rangverteilung: der Zusammenhang zwischen der Frequenz F_i und Rang i in den Daten des Häufigkeitswörterbuchs der Lexeme des Russischen (Zasorina 1977). Bei kleinen Häufigkeiten werden gemittelte Ränge verwendet (siehe Tab. 2.9). Bilogarithmische Darstellung.

tigt, daß im mittleren Teil der Verteilung in der Realität die Linearität des Zusammenhangs (in der Grafik die gerade Linie in bilogarithmischen Koordinaten) erhalten bleibt. Die endgültige Entscheidung über diese Frage erfordert neue Untersuchungen auf der Grundlage von sehr großen Textkorpora.

Eine weitere theoretische Begründung einer "nichtlinearen" Konzeption des Zipfschen Gesetzes ist in der Arbeit von Byčkov (1984) gegeben, der die These eines in zwei Komponenten gespaltenen Parameters entwickelt - das eigentlich

konstante γ_0 und eine "gleitende" Variable γ_i. Die Formel 2.13 kann man in der Form

(2.14) $$F_i = Ci^{-\gamma_i}$$

schreiben, wo γ_i (d.h. der Wert des γ-Parameters in jedem einzelnen Punkt der Rang-Frequenz-Kurve) sich aus:

(2.15) $$\gamma_i = \gamma_0 e^{di}$$

bestimmt, wo d der Koeffizient des Zuwachses des γ-Parameters im Übergangsintervall von $i = 1$ bis $i_{max} = V$ ist.

Auf der Grundlage kombinatorischer und wahrscheinlichkeitstheoretischer Überlegungen kommt Krylov (1982) zu der folgenden Formel für den Zusammenhang zwischen der Häufigkeit und dem Rang eines Wortes (hier in unserer Notationsweise):

(2.16) $$F_i = \frac{C}{i + B_1} - B_2$$

mit C, B_1 und B_2 als Parameter. Es ist leicht zu sehen, daß diese Formel sich von der Zipf-Mandelbrotschen Formel bei $\gamma = 1$ (2.12) lediglich durch den Summanden B_2 unterscheidet. Die Parameter B_1 und B_2 sorgen für die Möglichkeit einer achsenparallelen Verschiebung, dank derer der Abweichung vom linearen Zusammenhang sowohl im vorderen als auch im hinteren Teil der Verteilung Rechnung getragen wird. Es versteht sich, daß der Zipfsche Parameter γ, der die Steigung der Geraden in bilogarithmischer Darstellung repräsentiert, 1 beträgt. In den Fällen, in denen der Parameter γ auf Grund der empirischen Daten nahe an 1 ist, beschreibt die Formel den Zusammenhang zwischen F_i und i hinreichend gut.

Die Abweichung vom linearen Zusammenhang der Rangfrequenzverteilung veranlaßte auch den polnischen Forscher Woronczak (1967), eine eigene Variante des Zipfschen Gesetzes aufzustellen. Auch seine Formel trägt - ebenso wie die von Krylov - der Abweichung sowohl im vorderen wie im hinteren Bereich der Verteilung Rechnung, aber im Gegensatz zu letzterem schlägt Woronczak eine Variante vor, bei der der Parameter γ (der Steigungskoeffizient) sich von 1 unterscheiden kann ($\gamma \neq 1$). In unserer Notation hat Woronczaks Formel folgende Gestalt:

(2.17) $$F_i = N(i + B)^{-\gamma} Z^i \Phi^{-1}$$

mit γ, B und Z als Parameter und N als Textumfang und

(2.18) $$\Phi = \sum_{j=0}^{\infty} (j + B)^{-\gamma} Z^j.$$

Der Parameter B spielt die Rolle der Mandelbrotschen Korrektur, die der Abweichung im vorderen Teil der Verteilung Rechnung trägt, der Parameter Z (genauer der Ausdruck Z^i bei $|Z| < 1$) sorgt für ein im Vergleich zu Mandelbrots Formel (2.11) schnelleres Fallen der Funktion; eben dadurch wird die Abweichung der Verteilung im hinteren Teil berücksichtigt. Im Wesentlichen hat der Faktor Z^i die gleiche Wirkung wie das Wachsen von γ mit dem Rang i. Dieses Prinzip macht im gewissen Maße die Methode von Woronczak mit dem nichtlinearen Ansatz von Alekseev und Byčkov vergleichbar, und ähnlich wie sie ignoriert Woronczak im Wesentlichen den linearen Charakter des Zusammenhangs im mittleren Bereich der Verteilung.

Die Formel von Mandelbrot in ihrer zweiparametrigen Form (2.12) ist die Grundlage für das "verallgemeinerte Zipf-Mandelbrotsche Gesetz" von Orlov (1970, 1976, 1980). Dieser Autor entwickelt einen originellen Ansatz, dessen zentrales Konzept der sogenannte "Zipfsche Umfang" ist und den er mit dem Symbol Z bezeichnet und der als Ausgangspunkt für die Berechnung von Parametern der Frequenzstruktur des Textes dient. Aus seiner theoretischen Ableitung des Wertes von Z (zur Berechnung siehe Orlov, 1982) folgt die Formel

(2.19) $$p_i = k(i + B)^{-1}$$

wo p_i die relative Häufigkeit des Wortes mit dem Rang i ist, $k = [\ln(Z_{p_{max}})]^{-1}$, $B = kp_{max}^{-1} - 1$ und p_{max} die relative Häufigkeit des häufigsten Worts.

Anzumerken ist, daß der Parameter γ, der die Steigung der Geraden anzeigt, in dieser Formel den festen Wert 1 erhält (zu einer Variante mit $\gamma \neq 1$ siehe Orlov, 1976: 184 ff).

Die Formel von Orlov ergibt hinreichend gute Resultate in den Fällen, in denen der Parameter γ auf Grund der empirischen Daten nahe an 1 liegt und die Berechnung des "Zipfschen Umfangs" (Z) für den gegebenen Text gelingt. Der Zipfsche Umfang repräsentiert nach der Intention von Orlov den einzigen Umfang eines gegebenen Textes, bei dem er theoretisch dem Zipf-Mandelbrotschen Gesetz genügen kann, wobei dieser Umfang mit dem Konzept der "Abgeschlossenheit" eines Textes in Zusammenhang gebracht wird.

Alle oben betrachteten Varianten von Formeln für das Zipfsche Gesetz (in der Form von Rangverteilungen) stellen sich als kontinuierliche Funktionen dar, obwohl wir in der Realität diskrete Verteilungen linguistischer Objekte vorliegen haben. Gewöhnlich wird diese Tatsache ignoriert, weil angenommen wird, daß eine stetige Funktion alle Eigenschaften der "sprunghaften" Veränderung der empirischen Funktion der Rangfrequenzverteilung vollständig adäquat ausdrückt. Um dennoch eine präzisere Fassung der Verteilung zu erhalten, schlagen Arapov und

andere (1975) ein spezielles "diskretes Analogon zum Zipfschen Gesetz" vor, das der Ganzzahligkeit von Frequenz oder Rang als hinreichender Bedingung genügt. Die entsprechende Formel hat folgende Gestalt:

$$(2.20) \qquad F_i = \frac{\beta(L + 1)^{\gamma}}{1 - \gamma}[(i + 1)^{1-\gamma} - i^{1-\gamma}],$$

bei dem sich der Parameter β in Abhängigkeit vom Wert des Zipfschen Parameters γ bestimmt (siehe Arapov, Efimova, 1975: 6) und L der Vokabularumfang ist (beim Autor als N bezeichnet).

In der letzten Zeit festigte sich - hauptsächlich dank der Untersuchungen von Jablonski (1977) und Haitun (1983) - die Ansicht, daß das Zipfsche Gesetz in seinen Gültigkeitsbereichen (insbesondere bei der Beschreibung soziologischer Erscheinungen) dieselbe universelle Grenzprozeßrolle spielt wie die Gaußverteilung (Normalverteilung) in nichtorganischen und anderen Prozessen. In diesem Zusammenhang spricht man von der Nichtnormalität des Zipfschen Gesetzes, die sich in Konzentration und Streuung sowie formal im Fehlen von Varianz (sie ist unendlich) ausdrückt. Es bestätigt sich, daß die Zipfsche Verteilung nicht nur eine von vielen empirischen Verteilungen ist, sondern auch ein theoretisches Gesetz, das eine zuverlässige mathematische Basis in Form der Theorie der stabilen "nichtnormalen" Verteilungen hat (Jablonski, 1977). Bekannt ist auch eine Hypothese, nach der die Zipfsche Verteilung ein universelles Gesetz darstellt, dessen Wirkungssphäre die "natürlich entstandenen komplexen Systeme" (Arapov, Šrejder, 1978: 75) sind. Dem kann man hinzufügen, daß nach Meinung einiger Forscher die Verteilungen vom Typ von einer "hyperbolischen Treppe" vor allem Systemeigenschaften wiederspiegeln: Ganzheit, Organisiertheit, Hierarchisierung (Kozačkov, 1978: 15).

Die Stabilität der Zipfschen Verteilung in bezug auf soziale Phänomene verbindet den Vorschlag, Erscheinungen, die dem Zipfschen Gesetz gehorchen, als sich im Gleichgewicht befindliche Systeme zu betrachten, in dem für das System günstigsten (optimalen) Zustand. Unter Berücksichtigung der Dynamik der Verteilung und der Gesetzmäßigkeit der Frequenzfluktuationen läßt sich die Zipfsche Struktur des Textes insgesamt als *Fließgleichgewicht* des Systems charakterisieren.

Auf diese Weise und durch wahrscheinlichkeitstheoretische Überlegungen (Nichtnormalität, Fließgleichgewicht) kann man zu der Folgerung kommen, daß das Zipfsche Gesetz einen optimierenden Charakter besitzt. Bei diesen Überlegungen sollte man sich jedoch nicht zu übermäßiger Begeisterung hinreißen lassen, etwa in dem Sinne, daß die Übereinstimmung mit dem Zipfschen Gesetz für einen Text mit "literarischer Vollkommenheit", "künstlerischer Hochwertigkeit" usw. zusammenhängt. Man muß bedenken, daß sich mit Hilfe des Zipfschen Gesetzes lediglich die abstrakte, formale Häufigkeitsstruktur (die Anordnung konstruktiver Elemente) des Textes messen läßt. Ohne sprachliche Konkretisierung kann der

Zusammenhang zwischen der Häufigkeitsstruktur und der inhaltlichen Seite des Textes wohl nur sehr indirekt, mittelbar, sein. Zu seiner Bestimmung ist noch viel linguistische Forschung erforderlich.

Besondere Aufmerksamkeit verdienen die Versuche, das Prinzip der Zipfschen Verteilung in der *Parole* (Konzentration und Dispersion) mit der Gehirntätigkeit in Verbindung zu bringen. So versucht Lebedev (1983) die quantitativen Eigenschaften der Sprachgenerierung mit der raum-zeitlichen Organisation der periodischen Prozesse des Großhirns zu erklären. Ausgehend von der Vorstellung der Kodierung von Wortbildern in "Wellenbündel neuronaler Aktivität" stellt er zunächst *per analogiam* eine Formel zur Bestimmung "des Gesamtbereichs der Wortrangschwankungen" (q) auf und gelangt dann (in unserer Notation) zu dem Ausdruck

$$F = CQ,$$

wo $C = F_{max}$ und $Q = (1/q)\ ln\ (1+q/i)$ und i der mittlere Wortrang sind. Diese Formel und ihre Parameter haben wie von ihrem Autor intendiert eine klare psychologische und physiologische Bedeutung. Ein erstes Experiment (Lebedev, 1983: 16) bestätigt die Übereinstimmung von empirischen und theoretischen Daten.

Die vergleichende Analyse verschiedener Varianten des Zipfschen Gesetzes zeigt, daß es möglich ist, in Abhängigkeit von den Eigenheiten des jeweiligen Textes erfolgreich mit Hilfe der einen oder anderen Variante die Rangverteilung der Wörter zu beschreiben. So folgt z.B. das Häufigkeitswörterbuch der Wortformen im Estnischen (Tabelle 2.5) gut der Grundformel des Zipfschen Gesetzes (2.8) mit geringen Abweichungen im Anfangs- und im Endteil der Verteilung; für das Häufigkeitswörterbuch der Lexeme (Tabelle 2.4) mit einer Abweichung im Anfangsteil ist die Formel mit der Korrektur von Mandelbrot (2.11) besser geeignet usw. (s. Tuldava, 1985 für nähere Angaben zu den Resultaten der vergleichenden Analyse). Die festgestellten Abweichungen von den empirischen Daten lassen sich mit natürlichen Fluktuationen der Parameterwerte der linguistischen Verteilungen erklären, die man im Rahmen des quantitativ-systemischen Ansatzes als stochastische Systeme mit ihren Stabilitäts- und Variabilitätseigenschaften betrachten muß. Aber auf alle Fälle bleibt die allgemeine Form der Verteilung gewahrt, die Konzentration und Dispersion der Objekte aufweist.

Schließlich muß noch erwähnt werden, daß die Rangverteilung der Häufigkeit lexikalischer Einheiten auch in Integralform dargestellt werden kann, d.h. in Form der Abhängigkeit zwischen dem Rang i und der kumulativen Häufigkeit F_i^* oder p_i^*. Graphisch ergibt dies eine Kurve, die die kontinuierliche Annäherung an das Maximum, den Gesamtumfang des Vokabulars, zeigt (s. Abbildung 2.4). Diese Integralform der Verteilung läßt sich analytisch mit Hilfe der dementsprechenden Formeln des Zipfschen Gesetzes ausdrücken. Die Formeln 2.8 und 2.9 nehmen in der entsprechenden Integralform die folgende Gestalt an: (vergleiche Arapov,

1981; Haitun, 1983):

$$(2.21) \quad F_i^* = \frac{C}{\gamma - 1}[1 - (i + 1)^{1-\gamma}] \quad (\text{bei } \gamma \neq 1)$$

$$(2.22) \quad F_i^* = a - b \ln i \quad (\text{bei } \gamma = 1)$$

Die Werte der Parameter C und γ sind mit denen der Parameter in den Formeln 2.8 bzw. 2.9 identisch. Die Parameter $a \approx C \ln (1+B)$ und $b \approx C / \log e$ (s. Haitun, 1983: 71).

Bei Berücksichtigung der Korrektur von Mandelbrot stellt sich die integrale Form der Verteilung folgendermaßen dar:

$$(2.23) \quad F_i^* = \frac{C}{\gamma - 1}[(1 + B)^{1-\gamma} - (i + 1 + B)^{1-\gamma}] \quad (\text{bei } \gamma \neq 1)$$

$$(2.24) \quad F_i^* = C \ln \frac{i + 1 + B}{1 + B} \quad (\text{bei } \gamma = 1).$$

Bekannt ist auch die Möglichkeit der Approximation der Integralverteilung mit Hilfe der Weibullfunktion (Belonogov, Novoselov, 1971). Die Weibullfunktion hat in Anwendung auf die Rangverteilung von Wörtern die Gestalt:

$$(2.25) \quad F_i^* = N(1 - e^{-ci^k}),$$

wo N der Textumfang, i der Wortrang, c und k Parameter und e die Basis des natürlichen Logarithmus sind.

Die Integralform der Rangverteilung erlaubt, die Abdeckung eines Textes durch eine gegebene Anzahl lexikalischer Einheiten zu bestimmen. Sie hängt vom Typ der lexikalischen Einheiten ab. Bei der Untersuchung des textuellen Abdekkungsgrades von Lexemen wurde festgestellt, daß die entsprechenden Daten für verschiedene Sprachen sich einander insbesondere im Bereich der mittelfrequenten Wörter ähneln. Man kann der Aussage von Frumkina (1961), daß für die Mehrzahl der Sprachen der textuelle Abdeckungsgrad (abgesehen von der Abhängigkeit vom Sprachtyp) der 1500 häufigsten Lexeme ungefähr 80 ± 10% beträgt, zustimmen. Dabei ist die Dynamik der Veränderung des Abdeckungsgrades in Abhängigkeit von der Textlänge zu beachten: je länger der Text, desto geringer ist (im Durchschnitt) die relative Abdeckung im Bereich der mittel- und niederfrequenten Wörter.

Angaben zum textuellen Abdeckungsgrad von *Lexemen* werfen ein Licht auf wichtige Aspekte der Texterzeugung. Es zeigt sich z.B., daß die 10 häufigsten Lexeme, die ja nur einen unbedeutenden Anteil des Vokabulars ausmachen, in einem Text vom Umfang von $N \approx 100000$ (Tabelle 2.4) ungefähr 20% abdecken; die

1000 häufigsten Lexeme (7% des Vokabulars) entsprechen ungefähr 70% des Textes usw. Man kann auch zeigen (anhand der Daten von Tabelle 2.4), daß ein 25prozentiger Abdeckungsgrad des Textes durch 22 Lexeme erreicht wird, eine 50prozentige Abdeckung durch 196 Lexeme und eine 75 prozentige Abdeckung durch 1213 Lexeme. Solche Angaben haben Bedeutung in der Fremdsprachendidaktik und auch für einige weitere Anwendungsbereiche.

Der textuelle Abdeckungsgrad von *Wortformen* wird als eine der wichtigsten Charakteristiken der quantitativen Sprachtypologie angesehen. Er hängt eng mit dem morphologischen Wortaufbau in der untersuchten Sprache zusammen (Bektaev, 1978). Untersuchungen haben z.B. gezeigt, daß in den indoeuropäischen Sprachen für die Abdeckung von 50% eines Textes im Mittel 80 ... 200 Wortformen erforderlich sind, während in den agglutinierenden türkischen Sprachen hierzu 700 ... 800 Wortformen erforderlich sind. K.B. Bektaev stellte folgendes Schema zur typologischen Charakterisierung von Sprachen auf der Grundlage des textuellen Abdeckungsgrades der 100 häufigsten Wortformen vor: agglutinierende Sprachen 20...28%, flektierend-synthetische Sprachen 24...42%, flektierend-analytische Sprachen 43...54% und flektierend-analytische Sprachen mit amorphen Elementen 48...60%. Die estnische Sprache, die eine Abdeckung von 28,9% (Tabelle 2.5) aufweist, fällt in die Gruppe der flektierend-synthetischen Sprachen, wo sie sich jedoch dicht an der Grenze zu den agglutinierenden Sprachen befindet. Im Englischen beträgt der Abdeckungsgrad (bei einem Textumfang von 100000) 47,6% (Kučera, Francis, 1967: 313). Für das Russische ist die entsprechende Prozentzahl bei dem gleichen Textumfang 32,5 (Kalinina, 1968: 104-105).

Das Frequenzspektrum der Lexik

Gibt man für alle vorkommenden Wortfrequenzen die Anzahl der Wörter der jeweiligen Frequenz in rangierter Reihenfolge an, so erhält man die Spektralverteilung der Häufigkeiten oder das Frequenzspektrum der Lexik. Das Frequenzspektrum kann sowohl in bezug auf das Vokabular (auf das Wörterbuch zum konkreten Text) als auch in bezug auf den Text betrachtet werden, in beiden Fällen kann es in differentieller (nicht kumulativer) oder integral-(kumulativer) Form dargestellt werden (ein Beispiel eines verkürzten Frequenzspektrums auf der Grundlage von Daten eines abgeschlossenen Einzeltexts siehe Tabelle 2.10).

Das Frequenzspektrum reflektiert genau dasselbe Grundprinzip der Konzentration und Dispersion lexikalischer Einheiten, das wir bei der Betrachtung der Rangverteilung der Lexik festgestellt haben.

Die Konzentration der Einheiten erscheint hier im Bereich der geringen Häufigkeiten: im Wörterbuch und im Text bilden die Wörter mit der Häufigkeit $F = 1$ die größte Gruppe, dann folgen die Gruppen der Wörter mit den Häufigkeiten $F =$

Tabelle 2.10
Das Frequenzspektrum der Lexik nach Daten des Häufigkeitswörterbuchs der Lexeme im Roman von A.H. Tammsaare (estnisch): F = Wortfrequenz, m = Anzahl der Wörter mit gegebener Häufigkeit, p = relative Häufigkeit, m^*, mF^*, p^* = kumulierte Häufigkeit.

F	Wörterbuch				Text			
	m	m^*	p	p^*	mF	mF^*	p	p^*
1	3637	3637	0,442	0,442	3637	3637	0,023	0,023
2	1216	4853	0,148	0,590	2432	6059	0,015	0,038
3	613	5466	0,074	0,664	1839	7908	0,011	0,049
4	441	5907	0,054	0,718	1764	9672	0,011	0,060
5	297	6204	0,036	0,754	1485	11157	0,010	0,070
6	219	6423	0,027	0,781	1314	12471	0,008	0,078
7	175	6598	0,021	0,802	1225	13696	0,007	0,085
8	124	6722	0,015	0,817	992	14688	0,006	0,092
9	110	6832	0,013	0,830	990	15678	0,006	0,098
10	85	6917	0,011	0,841	850	16528	0,005	0,103
11-20	446	7363	0,054	0,895	6609	23137	0,041	0,144
21-50	421	7784	0,051	0,946	13826	36963	0,086	0,230
51-100	193	7977	0,024	0,970	13153	50116	0,082	0,312
101-500	202	8179	0,024	0,994	42576	92692	0,266	0,578
501-1000	29	8208	0,0035	0,9975	20841	113533	0,130	0,708
> 1000	20	8228	0,0025	1,0	46823	160356	0,292	1,0
Σ	8228 (L)	-	1,0	-	160356 (N)	-	1,0	-

2, $F = 3$ usw. Das Frequenzspektrum hat jedoch dynamischen Charakter. Bei Vergrößerung des Textes verringert sich der Anteil der *hapax legomena* sowohl im Wörterbuch wie auch im entsprechenden Text. In den Daten des englischen Häufigkeitswörterbuches (Kučera, Francis, 1967) z.B. beträgt der Anteil der *hapax legomena* in Stichproben verschiedener Länge:

	$N = 2000$	$N = 100000$	$N = 1000000$
im Wörterbuch	69,9%	51,6%	44,7%
im Text	28,3%	7,0%	2,2%

Entsprechend vergrößert sich der gemeinsame Anteil häufigerer Wörter. Einige Forscher nehmen an, daß bei Vergrößerung des Textes ein Moment eintritt, in dem das Häufigkeitsspektrum, insbesondere die Menge der *hapax legomena*, sich stabilisiert (siehe z.B. Williams, 1970: 103). Die experimentellen Daten besagen jedoch, daß bei Textvergrößerung das anteilige Gewicht der seltenen Wörter ständig sinkt. Theoretisch kann man sogar vermuten, daß in einem sehr großen Text ("in einem Globalkorpus") *hapax* oder *dis legomena* lediglich ausnahmsweise angetroffen werden können (Piotrowski, 1984). Die Wirkungsweise des Mechanismus, der das Frequenzspektrum hervorbringt, kann auf folgende Weise erklärt werden: Bei fortlaufender Vergrößerung des Textmaterials sinkt die Zahl der "0-mal vorkommenden" Wörter immer mehr, d.h. der Wörter, die tatsächlich in der Lexik der untersuchten Sprache existieren, aber in der Stichprobe nicht erfaßt sind, in der Korpusstichprobe und manifestiert sich im Häufigkeitswörterbuch. Gleichzeitig vergrößert sich die Anzahl einiger seltener Wörter, die in die mittlere Zone des Häufigkeitswörterbuchs hinüberwandern. Einfluß auf den Charakter der Strukturänderung des Frequenzspektrums kann auch der Übergang von Wörtern aus der mittleren Zone in die Zone der hochfrequenten Wörter haben und umgekehrt, obwohl diese Übergänge bei der Vergrößerung des Umfangs eines homogenen Textes weniger verbreitet sind (Piotrowski, *ibidem*). Im allgemeinen ist die Gesetzmäßigkeit des "ungleichmäßigen Übergangs" (Muller, 1976: 144) zu beobachten: Jede Teilmenge m_i, d.h. die Anzahl von Wörtern mit der Häufigkeit $i = 1,2,...$ hat die Tendenz, bei Vergrößerung der Textgrundlage eher zu "gewinnen" als zu "verlieren". Das bedeutet z.B., daß Übergänge aus der Gruppe m_i nach m_{i+1} häufiger sind als Übergänge aus der Gruppe m_{i+1} nach m_{i+2}. Was nun den analytischen Ausdruck der spektralen Verteilung der Wortfrequenzen angeht, also des Frequenzspektrums der Lexik, so muß man bedenken, daß diese Verteilung "organisch" mit der Rangverteilung zusammenhängt (sie bilden zwei interdependente Hälften der allgemeinen Frequenzstruktur des Textes). Folglich muß sich die Wirkungssphäre des Zipfschen Gesetzes auch auf die Spektralverteilung erstrecken. Geht man davon aus, daß die Rangverteilung einer gegebenen Frequenzverteilung mit Hilfe der Zipfschen Formel und Mandelbrots Korrektur (siehe Formel 2.11) beschrieben wird, dann ist

das spektrale Analogon der Rangverteilung (siehe Haitun 1983: 161)

(2.26) $m(F) = cF^{-(1+\alpha)}$,

wo $m(F)$ die Anzahl der Wörter mit der Häufigkeit F, c und α Parameter sind, wobei $\alpha = 1/\gamma$ (γ aus Formel 2.11) und $c = \alpha(L-1)/(1-F_{max}^{-\alpha})$ wobei L der Vokabularumfang und F_{max} die Häufigkeit des häufigsten Wortes sind. Weil in Worthäufigkeitsverteilungen $F_{max}^{-\alpha}$ gewöhnlich extrem klein und $L \gg 1$ ist, kann man praktisch annehmen, daß c ungefähr $= \alpha L$.

Auf Grund theoretischer Überlegungen kann man annehmen, daß die Entsprechungen zwischen dem Parameter γ der Rangverteilung und dem Parameter α aus der Formel 2.26 am besten erreicht wird, wenn Parameterwerte von γ zugrunde gelegt werden, die für den mittleren oder den Endteil der Rangverteilung festgestellt wurden. Diese Teile der Rangverteilung entsprechen dem mittleren bzw. Anfangsteil der spektralen Verteilung der Worthäufigkeiten. Bekanntlich verändert der Parameter γ (der die Steigung der Geraden der Rangfrequenzverteilung in doppelt logarithmischem Maßstab ausdrückt) gewöhnlich einen Wert entsprechend den Abweichungen im Anfangs- und im Endteil der Rangverteilung. Grob kann man von drei "Stadien" der Verteilung sprechen, denen die Parameter γ_1, γ_2 und γ_3 (für den Anfangs-, den mittleren und den Endteil) entsprechen.

Bei $\gamma = 1$ nimmt die Formel die folgende Gestalt an:

(2.27) $m(F) = cF^{-2}$.

So stellte sich diese Abhängigkeit zu allererst Zipf selbst vor, obwohl er später eine präzisere Variante vorschlug:

(2.28) $m(F) = c(F^2 - 0{,}25)^{-1}$.

In Tuldava (1986a, 1996) wird eine Formel mit einem zusätzlichen Parameter d des Typs wie in der Mandelbrotschen Korrektur vorgeschlagen:

(2.29) $m(F) = c(F+d)^{-\beta}$,

wo c, d und β Parameter sind. Die Formel ist formal identisch mit dem Zipf-Mandelbrotschen Gesetz für Rangverteilungen.

Es gab noch weitere Versuche, die Spektralverteilung durch Hinzufügung ergänzender Parameter genauer zu approximieren, z.B. nach der folgenden Formel (Krallmann 1966: 88)

(2.30) $m(F) = cF^{-k}e^{bF}$,

wo c, b und k Parameter sind.

Auf der Grundlage einer der Varianten des Zipfschen Gesetzes haben noch weitere Autoren Formeln als analytischen Ausdruck des Frequenzspektrums entwickelt (Orlov, 1976; Arapov u.a., 1975; Krylov, 1982; Brookes, 1982; usw.). Sehr interessant ist ein neuer Versuch von Krylov (1987), eine Formel für das Frequenzspektrum theoretisch abzuleiten, wobei er von Variationsprinzipien ausgeht. Seine Formel ist im Prinzip ähnlich der Formel 2.30, hat aber eine andere Begründung:

(2.31) $m(F) = cF^{-2}e^{b/F}$

mit c und b als Parametern, deren Werte theoretisch auf der Grundlage der beobachteten Größen L (Vokabularumfang), N (Textumfang) und F_{max} (Häufigkeit des häufigsten Wortes im gegebenen Text) bestimmt werden.

Alle erwähnten Formeln beschreiben mehr oder weniger genau die Spektralverteilung der Worthäufigkeiten im Wörterbuch und im Text. Wir illustrieren die Anwendung der zwei Formeln (2.26 und 2.29) anhand von Daten von Lexemen verschiedener Sprachen (Tabelle 2.11). Es zeigt sich, daß bei der Abweichung von der "Linearität" im Anfangsteil der Verteilung Formel (2.29) mit ihrem Korrekturkoeffizienten bessere Resultate erbringt.

Die Basisformel 2.26, die dem Zipfschen Gesetz in seiner Rangform entspricht, stellt sich als dieselbe Potenzfunktion mit negativem Exponenten dar wie die Funktion, die die Rangverteilung ausdrückt. In der Differentialdarstellung, die das Wesen der Veränderung der Variablen offensichtlich macht, ist festzustellen, daß sowohl die Rang- wie auch die Spektralform der Zipfschen Verteilung unter das Gesetz der "konstanten relativen Zu- bzw. Abnahme" zu subsumieren ist, bei dem ein konstantes Verhältnis zwischen dem relativen Zuwachs der abhängigen und dem der unabhängigen Variable zu beobachten ist (Proportionalität der relativen Zuwächse). Für das Frequenzspektrum haben wir:

(2.32) $\dfrac{dm/m}{dF/F} = -(1+\alpha)$.

Die partiellen Abweichungen von dieser einfachen und natürlichen Abhängigkeit erklärt man mit verschiedenen, vor allem linguistischen Ursachen.

Das Modell des Häufigkeitsspektrums von Waring-Herdan (vgl. Herdan 1964) stellt eine andere Möglichkeit für die analytische Beschreibung der Spektralverteilung der Häufigkeit von Wörtern dar. Hier geht man von der Vorstellung aus, daß die Spektralverteilung der Häufigkeiten von Wörtern eine monoton fallende Folge $m_1, m_2, \ldots, m_n$ bildet (d.h. die Zahl der Wörter mit der Häufigkeit 1, mit der Häufigkeit 2 usw.), die durch zwei Parameter a und x determiniert ist:

Tabelle 2.11

Frequenzspektren: beobachtete und erwartete Anzahl von Wörtern $m(F)$ mit der Häufigkeit F in den Daten 1. des Häufigkeitswörterbuchs der Lexeme des Romans von A.H. Tammsaare (Estnisch; $N = 160356$, $L = 8228$; 2. Häufigkeitswörterbuch der Lexeme des Russischen (1977); $N = 1056382$; $L = 39268$. Berechnung nach den Formeln: I.(2.26): $m(F) = cF^{-(1+\alpha)}$; II.(2.29): $m(F) = c(F+d)^{-\beta}$; III.(2.33): $m_{i+1} = m_i(a+i-1)/(x+i)$

F	Häufigkeitswörterbuch der Lexeme des Romans von A.H. Tammsaare (Estnisch)				Häufigkeitswörterbuch der Lexeme der russischen Sprache (1977)			
	$m(F)$ beobachtet	$m(F)$ erwartet I	II	III	$m(F)$ beobachtet	$m(F)$ erwartet I	II	III
1	3637	5700	3617	(3637)	13379	17000	13068	(13379)
2	1216	1754	1295	1436	5746	6010	5771	6690
3	613	881	676	763	3364	3272	3368	39 83
4	441	540	420	472	2243	2125	2253	2638
5	297	370	288	320	1681	1521	1634	1872
6	219	271	211	231	1279	1157	1251	1395
7	175	208	162	174	977	918	995	1078
8	124	166	129	136	841	751	815	857
9	110	136	105	109	713	630	682	698
10	85	114	87	89	595	538	580	578
15	53	57	43	41	286	293	311	275
20	34	35	26	23	200	190	199	159
30	15	18	13	10	109	104	105	70
40	7	10	7	7	60	67	67	40
50	4	7	5	5	45	48	47	30
60					30	37	35	20
70					30	29	28	
80	3	3	2	3	26	24	22	
90					15	20	19	
100	2	2	2	2	14	17	16	
150					7	9	8	
200	1	1	1	1	4	6	5	
300					3	3	3	
400					2	2	2	
500					2	2	1	
600					1	1	1	
Parameter		c = 5700 α = 0,7	c = 5800 β = 1,8 d = 0,3	a = 1,35 x = 2,42	-	c = 17000 α = 0,5	c = 25000 β = 1,6 d = 0,5	a = 2,08 x = 3,16

Anmerkung: Für die Häufigkeiten $F \geq 100$ wurden Mittelwerte von $m(F)$ genommen; z.B. für $F = 100$ das Mittel der Werte $m(F)$ von $F = 98$ bis 102.

$$(2.33) \qquad m_{i+1} = m_i \frac{a + i - 1}{x + i},$$

wo i die Worthäufigkeit ($i = 1,2,...$) ist. Dieses Modell entspricht der Gesetzmäßigkeit des "ungleichmäßigen Übergangs", von dem oben schon die Rede war. Praktisch bedeutet dies eine konstante Verringerung des Verhältnisses m_i/m_{i+1}. Für die

Anwendung dieses Modells ist es erforderlich, den Textumfang (N), die Vokabulargröße (L oder V) und die Zahl der hapax legomena (m_1) zu kennen. Die Parameter a und x berechnen sich auf folgende Weise: $a = (Q - M - 1)^{-1}$; $x = aQ$ wo $Q = (1-m_1/L)^{-1}$; $M = L/N$.

Das Modell von Waring-Herdan gilt als gut bewährt bei Stichproben mäßigen Umfangs ($N < 200000$). Dies bestätigt sich am Material eines Einzeltexts (N ungefähr = 160000), während in einem größeren Textkorpus (mit $N \approx 1.000.000$ laufenden Wörtern) die Entsprechung nur im Anfangsbereich des Frequenzspektrums gut ist - ungefähr bis m_{15} - und das Modell im Weiteren nur eine starke Unterschätzung des Spektrums liefert (vgl. Tabelle 2.11).

Das Modell von Waring-Herdan ist bemerkenswert, da es auf den Übergängen von m_1 zu m_2, von m_2 zu m_3, usw. beruht. Auf dieser Grundlage hat die französische Forscherin Dolphin (zitiert nach: Muller 1976) die Hypothese aufgestellt, daß man mit diesem Modell auch einen Rückwärtsübergang bestimmen könnte, insbesondere den von m_1 nach m_0, wobei man unter m_0 die Wörter mit der Häufigkeit 0 versteht, d.h. Wörter, die (definitionsgemäß) zum Lexikon des Autors bzw. der Autoren gehören, im gegebenen Text aber nicht verwendet werden. Dolphin schlägt für diesen Fall eine spezielle Methode der Berechnung der Parameter a und x vor (Muller, 1976: 143). Aber auch auf die herkömmliche Weise läßt sich m_0 folgendermaßen berechnen: gemäß Formel 2.33 ergibt sich $m_1 = m_0(a-1)/x$ und folglich $m_0 = m_1/(a-1)/x$. Vorläufige Versuche zeigen, daß diese Methode erfolgreich bei der stilometrischen Analyse von Stichproben ungefähr gleich großen Umfangs verwendet werden kann, wobei der Index m_0 als spezielle Stil-differenzierende Maßzahl zu interpretieren ist.

Die spektrale Verteilung von Worthäufigkeiten kann auch in integraler (kumulativer) Form dargestellt werden. Auf der Grundlage von Daten zur Rangverteilung gemäß der Formel von Zipf (2.8) ergibt sich folgende Formel für die integrale spektrale Verteilung (Haitun, 1983: 161):

$$(2.34) \qquad m^*(F) = L(1 - F^{-\alpha}),$$

wo $m^*(F)$ die kumulierten Wortanzahlen, F die Worthäufgkeit, L der Vokabularumfang und der Parameter $\alpha = 1/\gamma$ sind. Hier ist es ebenso wie in der differentiellen Verteilungsform erforderlich, einen Korrekturparameter (c) zur Verbesserung der Entsprechung zwischen empirischen und theoretischen Daten einzuführen:

$$(2.35) \qquad m^*(F) = L[1 - (F + c)^{-\alpha}].$$

Das integrale Häufigkeitssprektrum gibt eine Vorstellung über die Abdeckung des Textes durch Wörter verschiedener Häufigkeiten. Solche Angaben haben Bedeutung für die stilometrische Analyse von Texten und für den typologischen

Vergleich von Sprachen.

Besonderes Interesse gebührt dem Umstand, daß das integrale Häufigkeitssprektrum der Wörter gut durch die Log-Normalverteilung (siehe z.B. Carroll, 1967) beschrieben wird (vgl. Abb. 2.6), und zwar in ihrer gestutzten Variante (siehe zu Einzelheiten Tuldava, 1986a; Manasian, 1987). Bekanntlich wird die Log-Normalverteilung zur Beschreibung einer bestimmten Art von stochastischen Prozessen verwendet. In diesem Sinne kann die Verteilung der Worthäufigkeiten (das Häufigkeitssprektrum) als Resultat eines speziellen stochastischen Prozesses interpretiert werden, der sich bei der Spracherzeugung auswirkt. Viele Forscher vertreten eben diese Ansicht, wobei zu bedenken ist, daß die Log-Normalität einer Verteilung das "der natürlichen Sprache eigentümliche Prinzip der optimalen Informationskodierung" widerspiegelt (Herdan, 1964: 61-62). Es gibt jedoch auch Einwände gegen die Verwendung der log-normalen Verteilung für die Approximation der spektralen Verteilung von Worthäufigkeiten. Die Approximation sei nämlich deswegen nicht korrekt, weil beim gegebenen Typ linguistischer Verteilungen eine Abhängigkeit der Momente vom Stichprobenumfang beobachtet wird, und dies widerspricht der Natur der Gaußschen Verteilungen, zu denen die Log-Normalverteilung gehört (siehe Haitun, 1983: 81,184-185).

Als Fazit kann man festhalten, daß viele der betrachteten Methoden es erlauben, die Spektralverteilung von Worthäufigkeiten ganz und gar zufriedenstellend analytisch zu beschreiben. Diese Vielfalt von Lösungen sollte nicht verwundern. In der Wissenschaft ist schon lange bekannt, daß als mathematische Modelle der untersuchten Erscheinungen verschiedene Funktionen dienen können, nämlich Exponential-, hyperbolische, logistische Gleichungen usw.; es kommt aber vor allem darauf an, ob das System der Ausgangspostulate im Hinblick auf die konkrete Erscheinung berechtigt ist. Welches aus mehreren möglichen Modellen vorzuziehen ist, hängt von der inhaltlichen Analyse des jeweiligen Problems und von der Möglichkeit ab, das Modell angemessen zu interpretieren. In unserem konkreten Fall ist es z.B. möglich, dem Modell von Zipf aus dem Grund den Vorzug zu geben, daß es eine sehr einfache und natürliche Abhängigkeit zwischen den Variablen ausdrückt (nach dem Gesetz der "konstanten relativen Zu- und Abnahme") oder weil es in gewissem Sinne auf einen Zusammenhang mit der Gehirntätigkeit hinweist (Hypothese von Lebedev). Als Alternative (oder Ergänzung) kann man das Modell von Waring/Herdan deshalb nehmen, weil hier der Zusammenhang mit der Gesetzmäßigkeit des "ungleichmäßigen Übergangs" betrachtet wird, der in der Praxis beobachtet wird; dabei besteht die Möglichkeit, das Vorhandensein von Wörtern mit sogenannter Nullhäufigkeit in Betracht zu ziehen. Die Verwendung der Log-Normalverteilung erlaubt es, die Spracherzeugung als stochastischen Prozeß zu betrachten, was als Grundlage für bestimmte Schlußfolgerungen über die Natur der Sprache dienen kann usw.

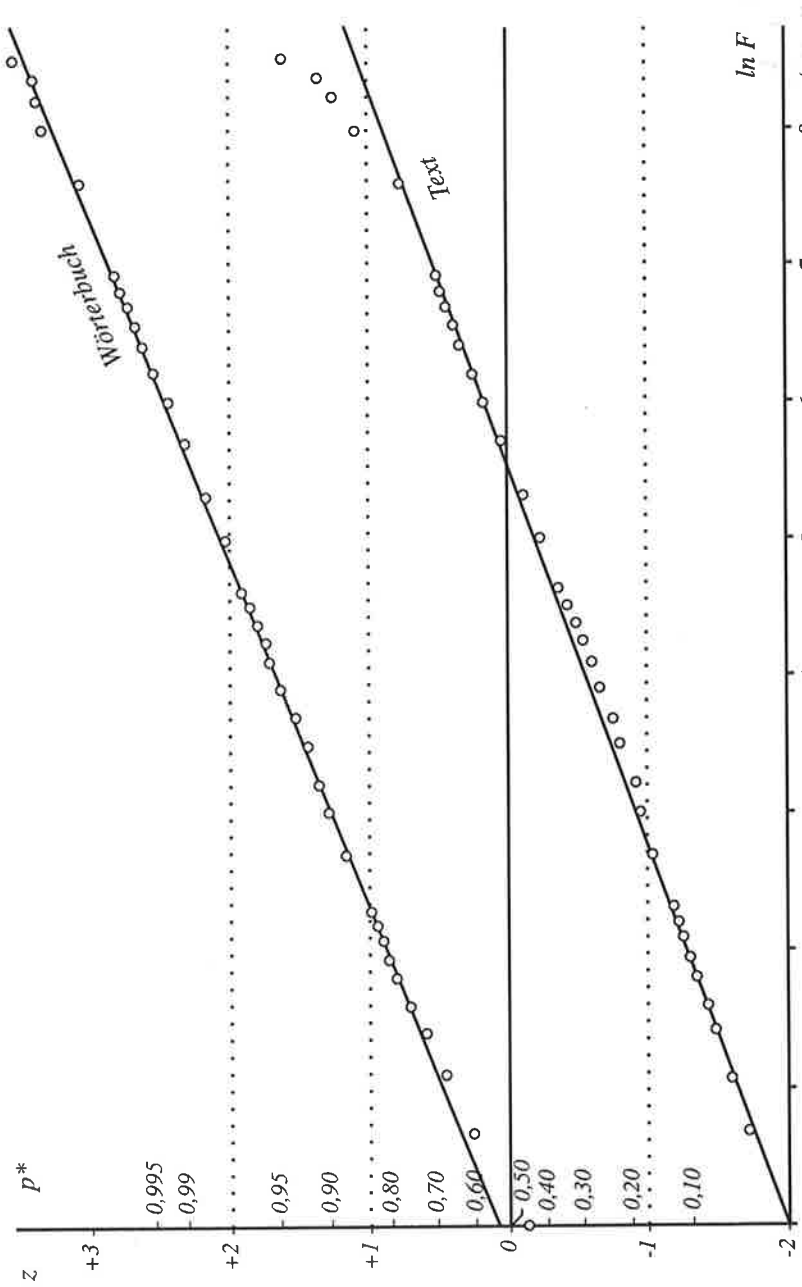


Abbildung 2.6. Überprüfung der Log-Normalität des spektralen Verteilung der Worthäufigkeiten im Tammsaares Roman. Logarithmischer Maßstab auf der Abszisse, Quantile der Normalverteilung (*z*) und die Abdeckung des Textes (*p**) auf der Ordinate

2.3. Die Abhängigkeit Vokabular - Text

Fragestellung

Die Frage nach der quantitativen Abhängigkeit zwischen Vokabularumfang und Textumfang in der Dynamik der Spracherzeugung hat sowohl theoretische wie auch praktische Bedeutung in der quantitativen Linguistik. Die Modellierung des Vokabularzuwachses in Abhängigkeit von der Vergrößerung des Textumfangs als Funktion, die eine bestimmte inhaltliche Interpretation besitzt, kann nicht nur unser Wissen über sehr allgemeine quantitative Gesetzmäßigkeiten der Spracherzeugung erweitern, sondern erlaubt auch die Lösung einer Reihe interessanter und aktueller Probleme angewandter Natur. Auf der Grundlage einer Funktion, die die Abhängigkeit des Vokabularumfangs (L)[2] vom Textumfang (N) ausdrückt, kann man z.B. bei gegebenem N das unbekannte L finden oder umgekehrt den Sättigungsgrad oder die Adäquatheit eines Stichprobenumfangs für die Planung von automatischen Informationssystemen bestimmen. Diese Bestimmung der Form des Zusammenhangs zwischen Vokabularumfang und Textgröße gestattet es auch, die stilistischen Eigenarten einzelner Texte oder Textsorten zu untersuchen, und trägt zur Lösung von pädagogischen und psychologischen Aufgaben bei (Textschwierigkeitsmessung, Bestimmung des Vokabularreichtums eines Textes, Autorenbestimmung und anderes).

Es gibt zahlreiche Versuche, eine solche empirische Formel aufzustellen (z.B. Kuraszkiewicz, 1958; Guiraud, 1959; Somers, 1959, Müller, 1971; Zacharova, 1967). Neben der einfachen Verwendung empirischer Formeln gab es auch Versuche, den Prozeß des Vokabularwachstums auf der Grundlage bestimmter theoretischer Voraussetzungen zu modellieren, z. B. auf der Basis der Vermutung einer log-normalen Verteilung der Wörter im Text (Carroll, 1967) oder einer Annahme über die Wirkung des Zipfschen Gesetzes (Kalinin, 1964; Cherc, 1969). Als Ausgangspunkt für die Aufstellung einer Wachstumsformel für das Vokabular sind auch andere bekannte Verteilungen verwendet worden, z.B. die Weibull-Verteilung (Nešitoj, 1975; 1984). Basierend auf einem neurophysiologischen Auswahlmechanismus stellte Lebedev (1986) eine Theorie über den Zusammenhang der Spracherzeugung mit Eigenschaften des menschlichen Gedächtnisses auf und schlug ein Modell zur Bestimmung des Vokabularumfangs anhand der Textgröße vor, quantitativ, unter Verwendung einheitlicher neurophysiologischer Parameter. Auf Grund einer Analyse bestimmter Eigenschaften der Integralfunktionen der Verteilung erstellte Krylov (1985) ein Modell der Abhängigkeit Vokabular - Text, das mit der Veränderung der Anzahl von *hapax legomena* und der Rangverteilung von Worthäufigkeiten im Text zusammenhängt.

[2] In Abhängigkeit von den konkreten Untersuchungsbedingungen kann der Vokabularumfang als Zahl der Lexeme (L) oder Wortformen (V) ausgedrückt werden.

Man geht davon aus, daß die theoretisch am besten anwendbaren diejenigen Modelle sind, die es erlauben, den Zusammenhang zwischen Vokabularumfang und Textumfang im Zusammenhang mit anderen Seiten der statistischen Textorganisation zu betrachten. Daher war ein wichtiges Ereignis in der quantitativen Linguistik die Veröffentlichung der Arbeiten von Kalinin (1964 und 1965), in denen erstmals das Problem gelöst wurde, ein komplexes Modell aufzustellen, das die Häufigkeitsstruktur des Textes selbst (die Rang- und Spektralverteilung der Häufigkeiten) und die Abhängigkeit "Text - Vokabular" erfaßte. Aber in Anbetracht der starken Idealisierung der Textstruktur (durch die Annahme von Zufälligkeit und Unabhängigkeit des Vorkommens von Wörtern in Texten und die strikte Interdependenz verschiedener quantitativer Aspekte des Textes) gelang es selten, dieses Modell in der Praxis zur linguistischen Textanalyse anzuwenden. Etwas bessere Resultate erzielte Orlov (1978; 1982), der eine verbesserte Variante des Modells auf der Grundlage des sogenannten "verallgemeinerten Zipf-Mandelbrotschen Gesetzes" vorschlug (zur kritischen Analyse dieses Ansatzes zur Untersuchung der Abhängigkeit "Text - Vokabular" siehe Tuldava 1980: 116-117).

Trotz der theoretischen Bedeutung "komplexer" (oder "struktureller") Modelle ergibt ihre praktische Anwendung nicht immer zuverlässige Ergebnisse, vor allem in den Fällen, wo höhere Genauigkeit gefordert ist - speziell beim stilometrischen Vergleich, bei der Prognose usw. Bekanntlich ist man bestrebt, in mathematischen Modellen möglichst viele Prozesse in ihren Zusammenhängen und Interdependenzen abzubilden. Doch einer der hochragenden Spezialisten für mathematische Methodologie schreibt: "Es ist jedoch noch nicht gelungen, ein Modell aufzustellen, das in harmonischer Weise die Gesamtheit der gleichzeitig stattfindenden Prozesse wiedergäbe. Fast immer wird nur einer der Prozesse zur näheren Reflexion separiert, für die übrigen benutzt man nur vereinfachte Angaben" (Rybnikov, 1979: 110). Daher kann man versuchen, die Abhängigkeit "Text - Vokabular" vom Gesichtspunkt der "immanenten" Eigenschaft dieser Erscheinung mit dem Ziel zu untersuchen, ein genaueres Modell zur Anwendung für praktische Probleme der Prognose oder der stilometrischen Analyse aufzustellen.

Das Modell der sukzessiven Auswahl

Der Zusammenhang von Vokabular und Text erscheint in diesem Modell als der primäre Aspekt der statistischen Organisation des Textes. Die quantitative Abhängigkeit zwischen Vokabular und Text muß man bei diesem Ansatz im Zusammenhang mit einigen besonders allgemeinen Faktoren der Texterzeugung untersuchen. Wir stützen uns dabei auf die Voraussetzungen der Theorie stochastischer Systeme (siehe Kapitel 1.1), nach der das untersuchte Objekt nicht nur durch den stochasti-

schen Charakter der Parameter (das niedrigste Organisationsniveau) bestimmt ist, sondern auch durch eine gewisse Stabilität und Regularität in der Masse der zufälligen Ereignisse (das höchste Organisationsniveau).

Im vorliegenden Fall kann man sich das Anwachsen des Vokabularumfangs im Verlauf der Textgenerierung als stochastischen Prozeß vorstellen, bei dem mit jedem Schritt eine zufällige Auswahl zwischen "neuen", vorher noch nicht vorgekommenen, und "alten", im gegebenen Text bereits verwendeten Wörtern stattfindet. Es ist eine empirische Tatsache, daß bei Vergrößerung des Textes die Wahrscheinlichkeit, ein "neues" Wort zu wählen, ständig sinkt. Daraus kann man folgern, daß der betrachtete stochastische Prozeß einer tieferen Gesetzmäßigkeit der Textgenerierung gehorcht. Inhaltlich wird der Prozeß des Vokabularwachstums durch eine komplizierte Wechselwirkung zweier entgegengesetzter Tendenzen bestimmt: das Streben nach Erweiterung und das Streben nach Begrenzung des lexikalischen Inventars des jeweiligen Textes. Einerseits werden bei der Texterzeugung Sprecher/Schreiber von dem Wunsch beherrscht, der assoziativen Kraft der Gedanken zu folgen und das gewählte Thema zu entwickeln und in die Breite zu gehen. Andererseits ist die freie Entwicklung durch Eigenschaften des menschlichen Gedächtnisses und auch z.B. durch die Notwendigkeit, im Rahmen einer bestimmten Thematik zu bleiben, begrenzt, was zur Wiederholung schon verwendeter Inhaltswörter führt; dazu kommt noch die aus grammatischen Gründen erforderliche ständige Wiederholung von Funktionswörtern. Diese beiden Tendenzen bestimmen in globaler Weise den stochastischen Prozeß des Vokabularwachstums; im Rahmen unserer Analyse jedoch ist es erforderlich, ein genaueres, allgemeines strukturell-funktionales Prinzip zu formulieren, das linguistisch sinnvoll bleibt und gleichzeitig einer mathematischen Analyse unterzogen werden kann. Als Regulierungsinstanz bei der Textgenerierung kann man ein Prinzip der *Begrenzung der lexikalischen Vielfalt* im Text ansehen, dessen letzte Ursachen offensichtlich auf phylogenetischen Faktoren beruhen (als Analogie zu diesem Prinzip kann man das Prinzip der Entropieverringerung in offenen selbstorganisierenden Systemen betrachten).

Lexikalische Vielfalt im Text ist ein etwas abstrakter Begriff, aber er wird präzise und anschaulich, wenn er mit Hilfe eines in der quantitativen Linguistik bekannten Maßes ausgedrückt wird - der "Type-Token-Ratio", das heißt, des Verhältnisses zwischen Vokabular- und Textumfang (*L/N*), oder durch das inverse Verhältnis (*N/L*), das die mittlere Worthäufigkeit im Text ausdrückt. Es ist bekannt, daß der Grad der lexikalischen Vielfalt sich mit der Vergrößerung des Textumfangs verändert: die Type-Token-Ratio (kurz TTR-Index) fällt monoton, während die mittlere Worthäufigkeit entsprechend ansteigt. Das nachstehende Beispiel zeigt die Veränderung der TTR mit wachsendem Textumfang anhand der Daten von Puschkins "Kapitanskaja Dočka" (siehe Frumkina, 1960):

N	*L*	*L/N*	*N/L*
5000	1568	0,31	3,19
10000	2432	0,24	4,11
29345	4900	0,17	5,99

Der TTR-Index, der das Verhältnis zwischen *L* und *N* mißt, ist auch insofern in bezug auf wesentliche quantitative Texteigenschaften erhellend, als er mit dem stochastischen Prozeß, der bei der Texterzeugung in jedem Schritt zwischen der Wahl eines "neuen" oder "alten" Worts entscheidet, eng korreliert ist. Das Verhältnis *L/N* spiegelt in gewissem Sinn die Wahrscheinlichkeit des Auftretens eines neuen Wortes und damit der Inventarvergrößerung wider. Tatsächlich ist nach Erreichen eines Textumfangs von *N* laufenden Wörtern und einem angesammelten Vokabular von *L* verschiedenen Wörtern die Wahrscheinlichkeit dafür, daß ein neues Wort zum Vokabular hinzukommt, proportional zum Verhältnis *L/N*.

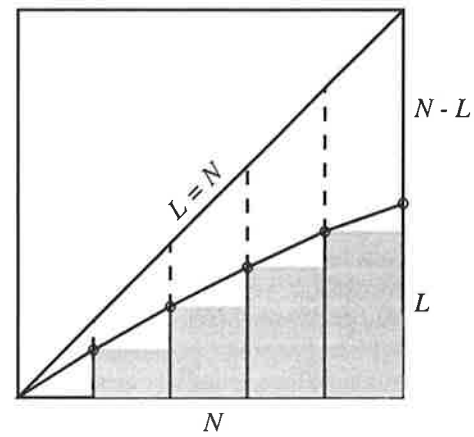


Abbildung 2.7. Das Wachstum der Größe des Vokabulars (*L*) in Abhängigkeit von der Zunahme des Textumfangs (*N*).

Zur Klärung einiger weiterer struktureller Eigenschaften der untersuchten Erscheinung kann man den Prozeß des Vokabularwachstums als ständige Veränderung des Anteils des verwendeten Vokabulars in einem Raum maximaler Möglichkeiten darstellen. Wenn man z.B. den Verlauf des Vokabularwachstums graphisch darstellt (Abbildung 2.7), kann man beobachten, wie sich das Anwachsen des Vokabularumfangs im Vergleich mit einer "Basislinie", der Geraden, der denjenigen idealen Vokabularumfang anzeigt, bei dem die Textwörter sich über-

haupt nicht wiederholen würden ($L = N$), stetig verlangsamt. Ein solcher Fall ist tatsächlich ganz am Anfang der Textgenerierung anzutreffen, während der folgende Verlauf eine beständige Abdrift vom Anfangszustand zeigt, wo der TTR-Index seinen maximalen Wert besitzt ($L/N = N/L = 1$). Wie bereits angemerkt, ist der Faktor, der diese Abdrift steuert, der durch das Sprachsystem bedingte Prozeß der Begrenzung der lexikalischen Vielfalt im Text.

In diesem Zusammenhang führen wir den Begriff des Rekurrenzdrucks (des Drangs zur Wortwiederholung) ein, der sich mit wachsendem Textumfang ständig verstärkt (das Verhältnis $(N - L)/N$, siehe Abbildung 2.7) und den Grad der lexikalischen Vielfalt (das Verhältnis L/N) begrenzt. Bezeichnet man L/N mit p, läßt sich der Rekurrenzdruck (ein Analogon zur Redundanz) folgendermaßen ausdrücken: $q = (N - L)/N = 1 - L/N$; $p + q = 1$. Wenn man den Vokabularumfang L als einen Teil eines hypothetischen Gesamtvokabulars ansieht, läßt sich, ausgehend von dem oben Gesagten, die Berechnung des Vokabularumfangs auf Grund beider Modelle durchführen:

$$L = Np \quad \text{oder} \quad L = N(1-q).$$

Aufstellen und Testen von Formeln

Um die verschiedenen Lösungsmöglichkeiten für den analytischen Ausdrucks des Zusammenhangs zwischen Vokabularumfang und Textlänge anschaulich darzustellen, werden wir die induktive Methode des schrittweisen Testens der Hypothesen und der graduellen Annäherung an die ursprüngliche Form der Abhängigkeit anhand des allgemeinen Modells $L = Np$ zeigen. Der Vorteil dieses Vorgehens liegt darin, daß wir unsere Ausgangsvoraussetzungen über die Rolle des Heterogenitätsfaktors im Prozeß der Texterzeugung beibehalten und dadurch entprechende Formeln aus einem systemischen Hintergrund konstruieren können.

In der ersten Approximation kann man von der Annahme ausgehen, daß die Veränderung der lexikalischen Vielfalt (des Heterogenitätsgrades der Lexik) ein kontinuierlicher Prozeß ist, der gleichmäßig und linear verläuft. Die empirische Überprüfung zeigt aber, daß die Abhängigkeit zwischen $p = L/N$ und N nur für einzelne Abschnitte des Textes linear approximiert werden kann. Nichtsdestoweniger kann man auf dieser Basis Formeln konstruieren, die für kleine Stichproben geeignet sind, z.B. (vgl. Tuldava 1974):

$$L = Na(N + b)^{-1}, \tag{2.36}$$

wo a und b Parameter sind (a zeigt die Grenze des Vokabularumfangs). Im gegebenen Fall geht es um den linearen Zusammenhang zwischen N/L (mittlere Worthäufigkeit) und N (Textumfang)[3].

Theoretisch besser begründet scheint die Annahme eines *hyperbolischen* Zusammenhangs zwischen $p = L/N$ und N. In diesem Falle besteht der lineare Zusammenhang zwischen $ln(L/N)$ und $ln\ N$, und der Zusammenhang zwischen dem Heterogenitätsgrad und dem Textumfang ergibt sich als

$$L/N = aN^b, \tag{2.37}$$

wobei a und b Parameter sind (im gegebenen Fall, wenn L/N mit zunehmenden N abnimmt, ist $b < 0$). Diese Formel ist formal analog dem Zipfschen Gesetz (Potenzfunktion mit negativem Exponenten). Aufgrund dieser Abhängigkeit erhält man

$$L = N(aN^b) = aN^B, \tag{2.38}$$

wobei $B = b+1$. Dieser Funktion entspricht in Differentialdarstellung $(dL/L)/(dN/N) = B$, d.h., es geht um das bekannte "allometrische" Gesetz oder das "Gesetz des stetigen relativen Zuwachses" (vgl. Formel 2.1).

Bekanntlich hat Herdan (1966) Funktion (2.38) als universal für den Ausdruck des Zusammenhangs zwischen Vokabular- und Textumfang betrachtet. Die Überprüfung zeigt, daß diese Funktion den Zusammenhang zwischen L und N in der Anfangsphase der Texterzeugung gut beschreibt (vom Anfang bis zu 5...10 tausend laufenden Wörtern, d.h. im Ausmaß einer Kurzgeschichte). Das weitere Tempo des Zuwachses des Vokabularumfangs in realen Texten verlangsamt sich aber, und die Voraussagen aufgrund der allometrischen Funktion ergeben überhöhte Schätzungen (im Detail s. Tuldava 1980).

Berücksichtigt man die Tatsache, daß bei Vergrößerung des Texts sich das Tempo des Vokabularzuwachses im Vergleich mit dem allometrischen Gesetz allmählich verlangsamt, so kann man die Variablen logarithmisch transformieren. Die Analyse zeigt, daß es zweckmäßig ist, zu dem ursprünglichen Prinzip zurückzukehren und die Gesetzmäßigkeit des Vokabularzuwachses durch den "Grad der lexikalischen Heterogenität" zu bestimmen, d.h. mit Hilfe des Logarithmus der Variablen in Formel (2.37). Daraus erhält man

$$ln(L/N) = a(ln\ N)b. \tag{2.39}$$

Das ergibt $L/N = e^{a(ln\ N)^b}$ $(a < 0)$ und die Formel für das Anwachsen des Vokabulars

[3] Formel (2.36) kann man als $N/L = (N+b)/a$ schreiben, woraus $N/L = N/a + b/a$. Setzt man $1/a = \alpha$ und $b/a = \beta$, dann erhält man die lineare Gleichung $N/L = \alpha N + \beta$.

(2.40) $$L = Ne^{-a(\ln N)^b}.$$

Wichtig ist der Umstand, daß Formel (2.40) den "regulären" Randbedingungen am Anfang des Textes unabhängig von dem Wert der Parameter a und b entspricht, d.h. bei $N = 1$ ist der Vokabularumfang $L = 1$:

$$L(N = 1) = 1e^{-a(\ln 1)^b} = 1e^0 = 1.$$

Bei der praktischen Anwendung der Formel kann man die Werte der Parameter a und b mit Hilfe der folgenden Linearisierung berechnen:

(2.41) $$\ln|\ln(L/N)| = A + b \ln \ln N,$$

wobei $A = ln|a|$. Der lineare Zusammenhang (2.41) erweist sich als adäquat bei der Analyse kurzer und langer, sowohl individueller als auch gemischter Texte (vgl. z.B. Abb. 2.8, vgl. Tabelle 2.12).

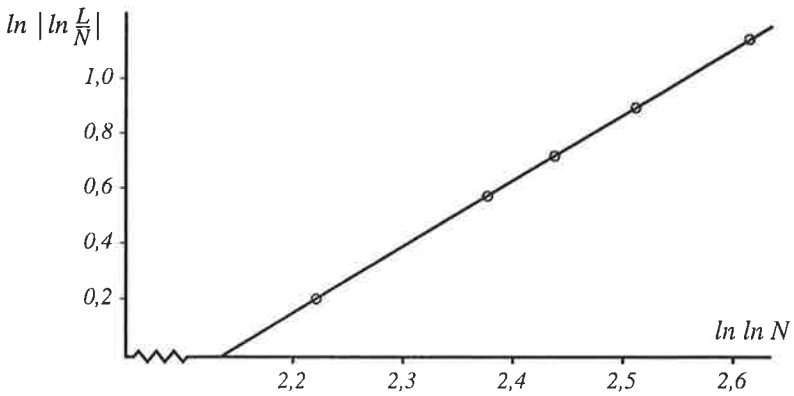


Abbildung 2.8. Linearer Zusammenhang zwischen $ln|ln\ (L/N)| = ln\ ln\ N$ für das Häufigkeitswörterbuch des Englischen nach Kučera und Francis.

Die Möglichkeiten der Extrapolation

Die gute deskriptive Fähigkeit einer Funktion für gegebene empirische Daten und vielfache Prognose sowohl vorwärts als auch rückwärts legt es nahe, sie auch zur Extrapolation für entferntere Textteile zu verwenden. Das ist ohne Zweifel ein riskantes Vorgehen, insbesondere wenn man bedenkt, daß wir im Augenblick keine

Kontrolle für eine derartige Prognose haben. Nichtsdestoweniger kann uns die automatische Textverarbeitung vor die Aufgabe stellen, den Vokabularumfang großer Textkorpora zumindest approximativ vorherzusagen. In einem solchen Fall ist es korrekt, eine Formel zu benutzen, die sich in der Praxis als äußerst stabil und zuverlässig erwiesen hat. Eine derartige Formel scheint aufgrund unserer Untersuchungen (2.40) zu sein, auch wenn sie eine obere Grenze hat (für natürlich sprachliche Texte bei $N = 10^9 ... 10^{10}$)[4].

Extrapolation oder *Prognose* beruht auf der Voraussetzung, daß die Interrelationen der prognostizierten Erscheinung mit anderen Erscheinungen im Wesentlichen konstant bleiben, und kann nur unter der Bedingung der Homogenität der Erscheinung durchgeführt werden. Unter Homogenität ist im vorliegenden Fall die qualitative Homogenität zu verstehen, d.h. die Gleichheit der Zusammensetzung, beispielsweise gleiche Thematik der Texte und der entsprechenden Vokabulare. Bei der Analyse zusammengesetzter (gemischter) Texte bedeutet die Forderung nach Homogenität die Aufrechterhaltung bestimmter Ausgangsproportionen verschiedener Genres oder Subsprachen und Einhaltung der übrigen Bedingungen des Experiments.

Zur Illustration zeigen wir Daten zur Wachstumsdynamik des Lexemvokabulars im Werk von A.S. Puschkin (Frumkina 1960):

	N	L
(1)	5000	1568
(2)	10000	2432
(3)	29345	4900
(4)	544777	21197

Auf der Grundlage der Daten der ersten drei Punkte (aus dem Roman "Kapitanskaja dočka") berechnen wir die Werte der Parameter a und b mit Hilfe der Methode der kleinsten Fehlerquadrate nach Formel (2.40) und bekommen $a = 0{,}0087$, $b = 2.29$. Extrapoliert man auf $N = 544777$ (Gesamtwerk des Autors), so erhält man $L = 22028$, d.h. ein Resultat, das die Realität sehr gut approximiert.

Wir überprüfen die Effektivität der Formel (2.40) auch am Material des zusammengesetzten Häufigkeitswörterbuches des Englischen (Kučera, Francis, 1967). Die Parameter schätzen wir mit Hilfe der Zweipunktmethode aus $N = 101566$ und $N = 253538$. Progose und Retrognose erweisen sich als gut (vgl. Tabelle 2.12).

Mit Hilfe von (2.40) unternahmen wir die Vorhersage des Vokabularumfangs für verschiedene Texte aus unterschiedlichen Sprachen bis zum Umfang von $N =$

[4] Die Analyse zeigte, daß Formel (2.40) auch für die Beschreibung des Zusammenhangs zwischen der Zahl der *hapax legomena* und der Textlänge geeignet ist (vgl. Tuldava 1977b).

Tabelle 2.12
Prognose des Vokabularwachstums aufgrund von (2.40) für das Häufigkeitswörterbuch des Englischen

Beobachtet	N	V	
Empirische Punkte	101566	13706	
	253538	23655	
Prognose	1014232	50617	(beob. 50406)
	10000000	148000	-
Retrognose	50721	8853	(8749)
	10051	2968	(3009)
	2000	902	(700-1000)
	1000	525	(-)
	500	300	(-)
	100	77	(-)
	10	9	(-)

$a = 0{,}00891$, $b = 2{,}215$, V = Vokabularumfang in Wortformen

10^7 (s. Tabelle 2.13). Beim Vergleich der Daten müssen der Unterschied zwischen den Genres (Subsprachen) und der Unterschied in der Berechnung des Vokabularumfangs (in Wortformen oder Lexemen) berücksichtigt werden. Für wissenschaftlich-technische Texte des Englischen beispielsweise (Tabelle 2.13, f,g) bei $N = 10^7$ prognostiziert man einen Textumfang von 36000 ... 38000 Wortformen, während man in gemischten englischen Texten dieser Länge ein Vokabular von 148000 Wortformen erwartet (Tabelle 2.12). Beim Vergleich von Daten, die man aus unterschiedlichen Sprachen erhebt, muß man auch den Unterschied im Analytismus berücksichtigen. So sind beispielsweise in englischen Texten über Elektronik bei $N = 10^7$ 38000 Wortformen, dagegen in russischen Texten über Elektronik bei gleicher Textlänge 94000 unterschiedliche Wortformen zu erwarten. In der agglutinierenden kasachischen Sprache (Zeitungstexte) erwartet man einen Vokabularumfang von 230000 Wortformen bei $N = 10^7$.

Zusammenfassend kann man folgendes feststellen:

Parallel zu dem rein empirischen Herangehen an das Problem der analytischen Erfassung des Zusammenhangs zwischen Text- und Vokabularumfang gibt es auch einige Richtungen der theoretischen Analyse in der zeitgenössichen quantitativen Linguistik. Man kann dabei sowohl hypothetisch-theoretisch als auch hypothetisch-empirisch verfahren (z.B. bei der Ableitung der Formeln im Abschnitt 2.3). Genauso wie bei der Analyse der Häufigkeitsstruktur des Textes (Abschnitt 2.2) kann man auch in diesem Fall verschiedene Modelle für die Beschreibung der untersuchten Erscheinungen sowie unterschiedliche Formeln für die Approximation empirischer Daten (Verteilungen) in Abhängigkeit von den angenommenen Ausgangs-

postulaten verwenden (vgl. auch Tuldava 1996).

Tabelle 2.13
Beobachteter und erwarteter Umfang (L oder V) des Vokabulars in Abhängigkeit von der Textlänge (N) in verschiedenen Sprachen nach Formel (2.40)

(a) Lettisch: Zeitungen, Lexeme (Latviešu val. 1969)			(b) Tschechisch: Technische Texte, Wortformen (Bečka 1972)			(c) Kasachisch: Zeitungen, Wortformen (Achabaev 1971)		
N	L	L'	N	V	V'	N	V	V'
50000	7065	7025	25000	4829	4827	25000	9088	9161
100000	9834	9919	75000	9603	9626	50000	15047	14875
200000	13389	13510	125000	13056	13050	100000	23895	23523
300000	16103	15912	175000	15858	15853	150000	29785	30378
500000	-	19200	500000	-	28200	500000	-	61000
10^6	-	24000	10^6	-	40000	10^6	-	87000
10^7	-	37000	10^7	-	114000	10^7	-	230000
$a = 0{,}003736$, $b = 2{,}6304$			$a = 0{,}01123$, $b = 2{,}1539$			$a = 0{,}001372$, $b = 2{,}8488$		
(d) Polnisch: Mickiewicz, Wortformen (Sambor 1970)			(e) Ukrainisch: Dovženko, Wortformen (Darčuk 1975)			(f) Englisch: Mechanismen, Wortformen (Lukjanenkov, Nešitoj 1975)		
N	V	V'	N	V	V'	N	V	V'
12172	3434	3458	5000	1629	1629	50495	4871	4849
29787	6146	6044	10000	2637	2646	100970	6858	6882
48255	8026	7998	15000	3504	3482	201966	9470	9520
64510	9250	9398	20000	4195	4214	302156	11314	11360
100000	-	11800	100000	-	11500	403966	12975	12832
500000	-	25000	500000	-	28000	500000	-	14000
10^6	-	33000	10^6	-	40000	10^6	-	18000
10^7	-	60000	10^7	-	110000	10^7	-	36000
$a = 0{,}00364$, $b = 2{,}6081$			$a = 0{,}01055$, $b = 2{,}1783$			$a = 0{,}01235$, $b = 2{,}2019$		
(g) Englisch: Elektrotechnik, Wortformen (Alekseev 1968)			(h) Rumänisch: Elektronik, Wortformen (Ešan 1966)			(i) Russisch: Elektronik, Wortformen (Kalinina 1963)		
N	V	V'	N	V	V'	N	V	V'
50000	5399	5437	50000	6785	6841	50000	9464	9388
100000	7853	7728	100000	10281	10070	100000	14062	14168
150000	9361	9371	150000	12477	12479	150000	17263	17803
200000	10582	10682	200000	14292	14454	200000	21468	20818
500000	-	15600	500000	-	22400	500000	-	33000
10^6	-	20000	10^6	-	30000	10^6	-	45000
10^7	-	38000	10^7	-	68000	10^7	-	94000
$a = 0{,}009152$, $b = 2{,}3057$			$a = 0{,}008148$, $b = 2{,}3086$			$a = 0{,}004284$, $b = 2{,}5058$		

3. Phonetische, grammatische und semantische Aspekte der Erforschung der Lexik

Entsprechend den in dieser Arbeit etablierten theoretisch-methodologischen Prinzipien (vgl. Abschnitt 1.3) kann man die Lexik nach verschiedenen linguistischen Gesichtspunkten (auf unterschiedlichen Unterebenen der Lexik) untersuchen, insbesondere unter dem Gesichtspunkt der Wechselbeziehung und der gegenseitigen Durchdringung der Lexik im Hinblick auf die phonetische, grammatische und semantische Ebene der Sprache. In Übereinstimmung mit den Prinzipien des quantitativ-systemischen Ansatzes zur Erforschung der Lexik sind die grundlegenden Methoden der Analyse die Taxonomie (Klassifikation) und die Modellierung mit Hilfe von Verteilungen.

3.1. Der phonetische Aspekt

Phonetische Klassifikation von Wörtern

Jedes Wort als Element des lexikalischen Systems der Sprache hat eine konkrete lautliche (phonemische, graphemische) *Form*, aufgrund derer man das Wort identifizieren und einer bestimmten formalen, hier lexikalisch-phonetischen, Gruppe zuordnen kann. Die phonetische Klassifikation der Wörter ist nach Smirnickij (1956:140) "absolut fundamental für die Charakterisierung der äußeren Gestalt der Lexik der gegebenen Sprache". Sie ist auch für die typologische Erforschung der Sprachen und Stile fundamental. Ungeachtet der bekannten Unabhängigkeit der Ausdrucksseite von der Inhaltsseite trägt die phonetische Struktur des Wortes das Abbild höherer Sprachebenen in sich. Folglich kann die Erforschung der Wortphonetik für die Erkennung allgemeinerer Gesetzmäßigkeiten in Struktur und Funktion der Wörter in der Sprache eine wichtige Rolle spielen.

Bei der quantitativ-systemischen Untersuchung lexikalisch-phonetischer Gruppen stellt sich die Frage nach der Bestimmung der Eigentümlichkeit der quantitativen Verteilung der Elemente (Wörter) nach ihrer phonetischen Struktur (nach Wortanfang und -ende, nach Phono- und Graphotaktik, nach den Modellen der Laut-, Phonem- oder Graphemdistribution, nach der Wortlänge) unter Berücksichtigung der Wechselbeziehungen mit anderen Teilebenen der Lexik.

Zur Illustration bringen wir Beispiele aus verschiedenen Sprachen, wobei wir im Grunde von Daten über lineare Buchstabenfolgen ausgehen, da wir es bei der automatischen Textanalyse mit Buchstaben-(Graphem-)Ketten zu tun haben. In vielen Fällen kann man aus den Buchstabenfolgen begründete Schlüsse auf die Distribution der Laute oder Phoneme ziehen.

Die Kenntnis der Besonderheiten der *Wort-initialen* und *Wort-finalen Struktur* ist nicht nur für die Sprachtypologie wichtig, sondern auch für die Lösung von Problemen der automatischen Textverarbeitung, besonders für die Ausarbeitung von Algorithmen der Textsegmentierung in relevante Einheiten und für die morphologische Analyse dieser Einheiten.

Der Vergleich der Daten aus verschiedenen Sprachen zeigt Ähnlichkeiten und Unterschiede in der phonetischen Struktur des Wortan- und -auslauts. Beispielsweise sind die fünf häufigsten Buchstaben am Wortanfang im Estnischen und im Russischen aufgrund der Daten des großen orthologischen Wörterbuches des Estnischen (Õigekeelsussônaraamat 1976) und des Akademischen Wörterbuches des Russischen in 17 Bänden (vgl. Andreev, 1967: 280) folgendermaßen verteilt (in %):

Estnisch	k (16,1)	p (10,2)	s (9,2)	t (8,2)	v (6,3)
Russisch	п (19,1)	с (9,1)	о (8,1)	н (7,1)	в (6,5).

Außer den Unterschieden in der Verteilung von Anfangsbuchstaben kann man auch allgemeine Merkmale beobachten: In beiden Fällen gibt es die gleiche Konzentration der Wörter, die mit den fünf häufigsten Buchstaben anfangen: Im estnischen Wörterbuch bilden sie 50,0%, im russischen 49,9%. In beiden Sprachen gibt es ein großes Übergewicht konsonantischer Elemente im Anlaut: Im Estnischen 84,5%, im Russischen 83,7% Konsonanten.

Wir verfügen auch über Daten von Auslautbuchstaben in verschiedenen Sprachen. In Auslaut des estnischen Wortes z.B. kommen hauptsächlich die Vokale *e* und *a*, sowie die Konsonanten *s, k, t* vor. Diese charakterisieren jedoch das Vokabular der Lexeme in ihrer "Grundform" (Nomina u.a. im Nominativ Singular, bzw. Verben im Infinitiv). Um die Besonderheiten der Wortverwendung in lebendiger Sprache (im Text) zu ermitteln, muß man Texte untersuchen.

Die Verteilung der Wörter im *Text* in unterschiedlichen Sprachen (Setälä, 1972; Lesochin u.a., 1982) aufgrund der fünf häufigsten Anfangsbuchstaben:

Estnisch	k (14,1)	t (9,5)	s (8,8)	m (8,3)	p (7,4)
Finnisch	j (13,0)	s (10,9)	k (9,9)	h (9,5)	t (9,3)
Russisch	в (11,7)	п (10,9)	и (9,6)	с (8,1)	о (7,9).

Hier erkennt man einige Unterschiede zwischen den nahe verwandten Sprachen Estnisch und Finnisch. In allen untersuchten Sprachen gehört das /s/ zu den

häufigsten Anlautbuchstaben (Lauten), im Estnischen und Finnischen kommen dazu noch /k/ und /t/.

Die entsprechenden Daten für den Auslaut in Texten:

Estnisch	a (19,1)	e (17,7)	s (13,0)	d (11,6)	i (11,5)
Finnisch	n (28,1)	a (23,7)	ä (13,1)	i (11,0)	e (10,1)
Russisch	й (22,1)	ь (22,0)	e (12,3)	a (9,7)	я (9,5).

Die aufgeführten Daten zur Verteilung von Anlaut- und Auslautbuchstaben wurden aus gemischten Texten ermittelt. Es ist bekannt, daß eine detailliertere Analyse bestimmte Unterschiede zwischen konkreten Verteilungen von phonetischen Einheiten in Texten unterschiedlicher Funktionalstile zutage bringt.

Die Rangverteilung der Wortformen aufgrund von Anlaut- oder Auslautbuchstaben im Text folgt approximativ dem *logarithmischen* Gesetz (Tuldava, 1986). Hier, sowie bei der Untersuchung der nichtbedingten Buchstaben- (Laut-, Phonem-) Häufigkeiten und auch in einigen anderen Fällen tritt eine Gesetzmäßigkeit zutage, laut derer die rangierten Häufigkeiten "elementarer" Einheiten (z.B. Laute oder Lautverbindungen) in der Regel dem Gesetz der logarithmischen (oder exponentiellen) Häufigkeitsabnahme folgen (vgl. auch Orlov, 1976: 185). Die empirische Häufigkeitsverteilung solcher Einheiten kann auch durch die Kreisgleichung approximiert werden (Piotrowski, 1984).

Phonotaktische Worttypen

Noch wichtigere Merkmale bei der Beschreibung der Worttypen liefert die Phono- oder Graphotaktik, d.h. die Bedingungen der Laut-(Phonem-) oder Buchstabenverbindungen in den einzelnen Wortpositionen. Detaillierte statistische Angaben über die Häufigkeitsverteilung der Buchstaben (Laute) und ihrer Verbindungen an der Wortgrenze und die Statistik ihrer Positionen im Wort in verschiedenen Sprachen findet man im Buch von Andreev (1967); für das Russische in den Arbeiten von Belonogov, Frolov (1963), Denisov u.a. (1978), für das Estnische in Kaasik, Tuldava (1980), Hint (1988).

Es ist bekannt, daß Buchstaben (Laute) in Wort-finaler Position, die mit den grammatischen Merkmalen der Wörter korrelieren, helfen, die Wortklassen zu bestimmen und letzten Endes (zusammen mit anderen Merkmalen) die automatische Textanalyse unterstützen (vgl. z.B. Belonogov u.a., 1983). Sehr hilfreich sind dabei rückläufige Wörterbücher, besonders Häufigkeitswörterbücher von Wortformen.

Das Wort weist verschiedene Typen von phonischen (phonemischen, graphischen) Strukturen auf, die aus Sequenzen von Konsonanten (C) und Vokalen (V) bestehen. Diese verallgemeinerten phonetisch-strukturellen Typen, die man mit den

Klassen C und V erfassen kann (weiter als CV-Typen bezeichnet), kann man vom quantitativen Standpunkt aus im Rahmen der Analyse der Verteilung formaler Wortgruppen in Sprachen untersuchen. In typologischen Untersuchungen der phonetischen Wortstruktur spricht man oft von sogenannten *kanonischen Formen*, mit denen man auf einer allgemeineren Ebene eine vergleichende Analyse der Wortstruktur in verschiedenen Sprachen durchführen kann. In kanonischen Formen berücksichtigt man das Merkmal der Dauer (Kürze oder Länge) der Laute nicht, und Diphtonge werden als einfache vokalische Elemente behandelt. Bei solch einer Analyse kann man beispielsweise die häufigsten CV-Strukturen einsilbiger Wörter in verschiedenen Sprachen vergleichen (aufgrund von Wörterbuchdaten vgl. Krámský 1966; Pankrac 1981):

Russisch	CVC, CCVC, CVCC, CCVCC, CV
Deutsch	CVC, CVCC, CCVC, CCVCC, CV
Englisch	CVC, CVCC, CCVC, CV, CCVCC
Kasachisch	CVC, VC, CVCC, CV, VCC
Türkisch	CVC, CVCC, VC
Ungarisch	CVC, CVCC, VC
Estnisch	CVC, CVCC, CV, CCVC, CVCC

Auf der einen Seite kann man die Ähnlichkeit indoeuropäischer Sprachen feststellen, auf der anderen Seite die der türkischen und finno-ugrischen Sprachen. Die häufigste Struktur einsilbiger Wörter in allen untersuchten Sprachen ist offensichtlich CVC.

Die Häufigkeitsverteilungen der CV-Strukturen im Text und im Wörterbuch können sich stark voneinander unterscheiden, zum Beispiel sind in estnischen Texten die häufigsten Strukturen (nach abnehmender Häufigkeit) CV, CVC, VC, CVCC, V, die zusammen 86.7% des Textes ausmachen (die Gesamtzahl der einsilbigen CV-Strukturen im Estnischen ist 15; detaillierter s. Tuldava, 1978).

Im englischen Wörterbuch sind außer den oben erwähnten Strukturen die zweisilbigen CVCVC und CVCCVC (Slipčenko, 1973) die häufigsten, und im Text CVCVC (Roberts 1965).

Interessant kann auch die quantitative Darstellung der Symmetrie des Systems der phonetischen Wortstrukturen sein. Ein System ist symmetrisch, wenn die Spiegelformen (CV und VC, CCV und VCC usw.) gleich häufig vorkommen. Die Symmetrie erhöht sich automatisch, wenn man die in sich symmetrischen Formen einbezieht (CVC, CCVCC, CVCVC usw.). Solche "autosymmetrischen" Strukturen bilden im Deutschen etwa 40% des gesamten Inventars einsilbiger Wörter, im Estnischen 43% des Inventars und 35% des Textes. Die Produktivität autosymmetrischer Strukturen findet man auch in anderen Sprachen, z.B. im Ukrainischen (Perebijnis, 1970). Gleichzeitig sieht man, daß zu den produktiven Wortmodellen

Tabelle 3.1
Verteilung der kanonischen Strukturen einsilbiger Wörter nach An- und Auslaut (konsonantische Umgebung des vokalischen Kerns)

(a) Deutsches Wörterbuch

Zahl der Konsonanten		Am Wortende				Σ(%)
		0	1	2	>2	
Am Wortanfang	0	0,6	2,1	2,5	0,6	5,8
	1	4,4	28,7	22,1	4,9	60,1
	2	1,9	18,1	10,2	1,6	31,8
	>2	0,4	1,4	0,5	0,04	2,3
Σ (%)		7,3	50,3	35,3	7,1	100,0

(b) Estnisches Wörterbuch

Zahl der Konsonanten		Am Wortende				Σ (%)
		0	1	2	3	
Am Wortanfang	0	0,4	2,9	2,7	0,03	6,0
	1	5,3	40,8	37,4	3,0	86,5
	2	0,3	4,2	2,1	0,5	7,1
	3	0,03	0,2	0,15	0,0	0,4
Σ (%)		6,0	48,1	42,4	3,5	100,0

(b) Estnischer Text

Zahl der Konsonanten		Am Wortende				Σ (%)
		0	1	2	3	
Am Wortanfang	0	5,6	11,9	2,8	0,0	20,3
	1	39,5	29,7	10,0	0,06	79,3
	2	0,06	0,2	0,06	0,03	0,4
	3	0,0	0,03	0,0	0,0	0,03
Σ (%)		45,2	41,8	12,9	0,1	100,0

auch asymmetrische Lautketten gehören. Ein ausgeglichenes Verhältnis zwischen Symmetrie und Asymmetrie wird mit Recht als ein wichtiges Merkmal betrachtet. Es gehört zur Grundlage der Struktur und der Funktion der Elemente des Sprachsystems als eines expressiven Kommunikationsmittels und ist ebenso Grundlage

seiner Entwicklung (Muravickaja, Slipčenko, 1982: 77; mehr über Symmetrie und Asymmetrie in der Sprache s. Tuldava, 1990).

Den Zusammenhang zwischen phonetischen Strukturen des Wortan- und auslauts kann man in Form von sogenannten Menzerathschen Parallelogrammen illustrieren (Menzerath, 1954) oder in Form einer Matrix, einer *Kontingenztabelle*. Weiter unten findet man solche Tabellen für die Verteilung einsilbiger Wörter im Estnischen und im Deutschen (für slawische Sprachen s. Zaplatkina, 1975, 1982). Die Tabellen enthalten verschiedene Informationen über quantitative Eigenschaften der phonetischen (kanonischen) Konstruktion einsilbiger Wörter mit einem vokalischen Kern (s. Tabelle 3.1). Beispielsweise endet im deutschen Wörterbuch der Einsilbler am häufigsten mit einem Konsonanten (50,3%), ebenso im estnischen Wörterbuch (48,1%), aber im estnischen Text überwiegt die Form mit einem Vokal (45,2%). Man kann feststellen, daß die Symmetrie im Deutschen (Wörterbuch) etwas höher ist als im Estnischen. Es gibt z.B. keinen großen Unterschied zwischen den Häufigkeiten der Spiegelstrukturen CVCC und CCVC (22,1% und 18,1%; vgl. im Estnischen 37,4% und 4,2%). Im Deutschen sind auch die Randverteilungen hinreichend ähnlich. Die Unterschiede zwischen den Sprachen erklärt man aufgrund typologischer Eigentümlichkeiten dieser Sprachen und bis zu einem gewissen Grad auch mit dem Unterschied der Häufigkeiten von Einheiten[1].

Die mathematische Schätzung des Zusammenhangs zwischen zwei Merkmalen, den Lautstrukturen im An- und Auslaut, kann man mit Hilfe des Kontingenzkoeffizienten

$$\Phi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1 \qquad (3.1)$$

ausdrücken, wobei n_{ij} die Häufigkeiten in einzelnen Zellen, $n_{i.}$ und $n_{.j}$ die Zeilen- und Spaltensummen, r die Zahl der Zeilen, k die Zahl der Spalten ist. Als Index der Stärke des Zusammenhangs zwischen den Merkmalen benutzt man eher die Wurzel aus Φ^2, d.h. Φ. Für das deutsche Wörterbuch ist $\Phi = 0{,}141$, für das estnische $\Phi = 0{,}102$. Es ist interessant, daß der Zusammenhang zwischen An- und Auslaut in Einsilblern im Text stärker ist als im Wörterbuch (im estnischen Text $\Phi = 0{,}239$). Dies hängt offensichtlich mit den Gesetzmäßigkeiten der rhythmischen Gliederung des Sprechflusses zusammen.

[1] Die deutschen Daten wurden nach Menzeraths Angaben zusammengestellt, wobei ihnen einsilbige Wörterbucheinheiten zugrundeliegen (insgesamt 2225 Einheiten). Im Estnischen wurden auch einsilbige flektierte Wortformen (Gesamtumfang des Inventars ist 3295 Einheiten) berücksichtigt.

Ein wichtiges Charakteristikum der Textstruktur und des Textvokabulars ist die Wortlänge mit ihrer Verteilung im Text und in seinem Vokabular. Die Wortlänge ist ein wichtiges quantitativ-typologisches Kriterium, das nicht nur strukturelle Merkmale der Sprache, sondern auch individuelle und funktionale Besonderheiten von Texten und Wörterbüchern prognostiziert (vgl. Alekseev, 1986). Heutzutage benutzt man Angaben über Wortlänge und Wortlängenverteilung auch bei der automatischen Textverarbeitung (Vertel', V.A., Vertel', E.V., 1970). Für die Praxis wichtig ist die Frage nach der analytischen Beschreibung der Wortlängenverteilung und des Zusammenhangs der Wortlänge mit anderen strukturellen Charakteristika des Textes und des Wörterbuches (Häufigkeit, wortbildende Aktivität, Wortwachstum usw.). Unter dem Gesichtspunkt des quantitativ-systemischen Ansatzes zur Erforschung der Lexik stellt die Wortlänge einen systembildenden Faktor dar: Sie bestimmt die Aufteilung der Lexik in Wortgruppen unterschiedlicher formaler Strukturen, und die Wortlängenverteilung weist auf die statistische Wechselbeziehung zwischen Wörtern verschiedener Länge im Prozeß der Redeerzeugung hin.

Als Grundlage für die Untersuchung der Wortlänge im Text dient die Wortverwendung in Form von Wortformen. Die Wortlänge im Wörterbuch hängt vom Wörterbuchumfang und von der gewählten Zähleinheit ab, d.h. von der Einzelform oder dem Lexem (Wort in Grundform) ab. Sehr informativ ist der Vergleich der Wortlängenformen im Text und im dazugehörenden Textvokabular.

Die Wortlänge (Wortformlänge) kann man numerisch in Buchstaben, Lauten, Phonemen, Silben oder Morphemen messen, also in Abhängigkeit von den Möglichkeiten und den gestellten Aufgaben.

Die Messung der Wortlänge in Buchstaben (Graphemen) ist insofern günstig, als die Prozedur leicht automatisiert werden kann. Wichtig ist, daß in vielen Sprachen die Buchstaben- und Lautinventare (oder Phoneminventare) und folglich auch die Wortlängenmessungen stark korrelieren.

Experimente zeigen, daß die mittlere Länge der Wortformen und die Verteilung der Wortformenlängen zwischen *Text* und *Wörterbuch* signifikant unterschiedlich sind. So findet man z.B. im Häufigkeitswörterbuch des Englischen (Kučera, Francis, 1967: 365-366), das aus ungefähr einer Million laufender Wörter kompiliert wurde, eine mittlere Wortformenlänge von 4,74 Buchstaben und in dem dazugehörigen Textvokabular (von etwa hunderttausend Wortformen) eine mittlere Länge von 8,13 Buchstaben. Dabei decken die kurzen Wortformen (1...4 Buchstaben) 34,5% des Textes ab, im Vokabular aber nur 4%. In einigen Sprachen, hauptsächlich in flektierend-synthetischen und agglutinierenden, findet man zweigipflige (bimodale) Verteilungen von Wortformenlängen mit Gipfeln z.B. auf zwei- und vierbuchstabigen Wortformen. Dies erklärt sich dadurch, daß die Verteilung der Wortformen nicht homogen ist: Sie setzt sich aus kurzen Hilfswörtern und langen

Autosemantika zusammen. Die Verteilung der Wortformenlänge im dazugehörigen Wörterbuch ist in der Regel "regulärer", die Bimodalität fehlt (für das Estnische s. Abbildung 3.1; ausführlicher s. Tuldava, 1986).

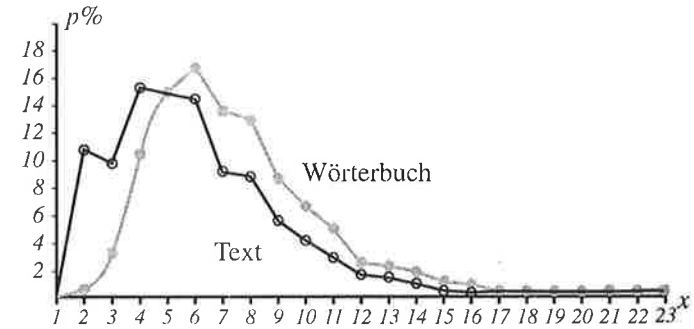


Abbildung 3.1. Die Verteilung der Wortformenlänge im Text und in dem dazugehörigen Vokabular im Estnischen (*Buchstabenlänge*)

Es ist zu bemerken, daß sowohl die mittlere Länge der Wortformen als auch die Verteilung der Wortformenlänge in Texten (und Textvokabularen) in verschiedenen Subsprachen beträchtlich variieren. So variiert beispielsweise die mittlere Länge der Wortformen im estnischen Text zwischen 4,8 (Dialogsprache in künstlerischen Texten) und 7,1 Buchstaben (in wissenschaftlich-technischen Texten). Der Anteil kurzer Wortformen (2...3 Buchstaben) in Dialogen beträgt 31%, aber in wissenschaftlich-technischen Texten nur 15%. (Ähnliche Angaben für andere Sprachen findet man in Andreev, 1967: 247; Nikonov, 1978: 107; Papp, 1980: 22, u.a.).

Die Wortformenlänge variiert auch je nach der Wortart und der Häufigkeitszone (vgl. Alekseev, 1986).

Empirisch wurde festgestellt, daß die Verteilung der Wortformenlängen (gemessen in Buchstaben oder Lauten) im Wörterbuch durch die Log-Normalverteilung gut approximiert werden kann (Herdan, 1966; vgl. auch Tuldava, 1986: 151).

Es wird angenommen, daß die Log-Normalität der Verteilung der Wortformenlängen dem Prinzip der optimalen Kodierung der Information entspricht und "die Tendenz zur klaren und fehlerfreien Differenzierung der Wörter ausdrückt" (Herdan 1966:205). Im Grunde genommen deutet die Log-Normalverteilung darauf hin, daß die Wahl eines Wortes einer bestimmten Länge gewissermaßen von der Länge des vorangehenden Wortes abhängt, wodurch sich im Text kurze, lange und mittlere Wörter abwechseln (vgl. Piotrovskij u.a., 1977:204). Für die analytische Beschreibung der Verteilung der Wortformenlängen im Text braucht man jedoch - insbesondere bei bimodalen Verteilungen - einen anderen Ansatz. Martynenko (1965) schlägt in solchen Fällen vor, Hilfswörter und Autosemantika separat zu

behandeln, und zeigt, daß man bei dieser Trennung die empirischen Daten mit einer modifizierten Potenzfunktion gut approximieren kann. Man kann die Verteilung auch kumulativ betrachten; in diesem Fall ist die Weibull-Verteilung geeignet.

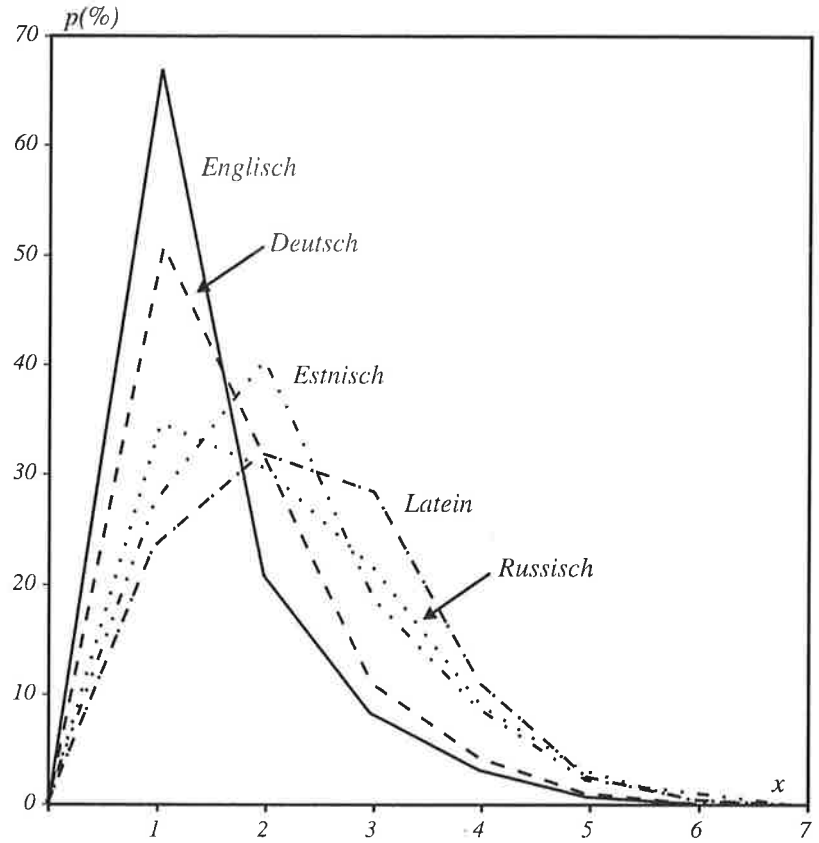


Abbildung 3.2. Verteilung der Wortformenlängen in Texten verschiedener Sprachen (x - mittlere Länge in Zahl der *Silben*, p (%) - Abdeckung des Textes in %)

Die Messung der Wortlänge in *Silben* (oder Morphemen) hat eine besondere Bedeutung für die Ermittlung der sogenannten Worttiefe. Solche Daten braucht man in Pädagogik, Psycholinguistik, Stilistik und auch in typologischen Untersuchungen. Besonders klar zeigt sich die Differenzierung der Sprachen beim Vergleich der Verteilungen der Wortformenlängen im Text (in identischen Funktionalstilen, vgl. Tabelle 3.2). Es zeigt sich beispielsweise, daß einsilbige Wortformen im englischen künstlerischen Text 66,8% ausmachen, im türkischen Text des gleichen Stils nur 18,8%. Im estnischen Text decken einsilbige Wortformen 25...40% des Textes ab, jeweils in Abhängigkeit von der Subsprache und vom Stil (in der künstlerischen Prosa ist es durchschnittlich 28,8%). Im Russischen machen einsilbige Wortformen ungefähr 33% des künstlerischen Textes aus. Die Ähnlichkeiten und Differenzen der Sprachen sieht man besonders klar, wenn man sie graphisch vergleicht (Abb. 3.2).

Die Verteilung der Wortformenlängen im Text, gemessen in Anzahl der Buchstaben oder Silben, kann man z.B. mit der Log-Normalverteilung beschreiben. Es gibt jedoch auch andere Möglichkeiten. Die Verteilung der Wortformenlängen in Silben approximiert man oft mit der Čebanov-Fucks-Funktion (Čebanov, 1947; Fucks, 1956, 1957). Es handelt sich um eine modifizierte Poisson-Verteilung, bei der man annimmt, daß der Wortbildungsprozeß zufällig ist. Untersuchungen haben gezeigt, daß die empirischen Daten noch besser mit der verallgemeinerten Gačečiladze-Cilosani Formel (verallgemeinerte Čebanov-Fucks Formel) erfaßbar sind, die auf spezielle Weise das Zusammenwirken von zufälligen und deterministischen Prozessen in der Rede berücksichtigt (Gačečiladze, Cilosani, 1971).

Bei den Untersuchungen der Wortlänge wurde beobachtet, daß zwischen der Zahl der Silben (oder Morpheme) und der Zahl der Buchstaben (Phoneme) im Wort eine gesetzesartige Verbindung besteht. Verlängert sich das Wort (Wortform), dann verringert sich das Verhältnis der Buchstabenzahl zur Silbenzahl. Anders gesagt, je länger das Wort, desto kürzer seine Silben, oder allgemeiner: "je länger ein linguistisches Konstrukt, desto kürzer seine Komponenten" (das *Menzerathsche Gesetz;* vgl. Altmann, 1980).

Diese Tendenz kann man aufgrund des empirischen Materials bestätigen. Beispielsweise sieht man bei den Daten aus der estnischen künstlerischen Prosa (s. Tabelle 3.3), daß die mittlere Silbenlänge mit der Wortlänge abnimmt: Die mittlere Silbenlänge der Einsilbler ist 2,75 Buchstaben (letzte Spalte der Tabelle 3.3), der Zweisilbler 2,40, der Dreisilbler 2,37 usw. bis zu 2,17 für Sechssilbler (die Abweichungen der Fünfsilbler von der allgemeinen Tendenz erklärt man dabei durch zufällige Faktoren). Die Abhängigkeit der Silbenlänge von der Wortlänge beschreibt man approximativ nach logarithmischer Transformation durch

$$y = a + b \ln x, \tag{3.2}$$

wobei y die Silbenlänge in Buchstaben, x die Wortlänge in Silben, a und b die Parameter (im gegebenen Fall $a = 2{,}75$ und $b = -0{,}33$)[2] sind. Dieser Gesetzmäßigkeit

[2] Wenn x die Wortlänge in Buchstaben ist (vgl. Tabelle 3.2), dann verändern sich nur die Parameter a und b.

Tabelle 3.2
Die Verteilung der Wortformenlängen in Texten der künstlerischen Prosa verschiedener Sprachen
(aus Fucks, 1956; Herdan, 1956; Zsilka, 1974; Jakubajtis, 1963)

Zahl der Silben	Englisch	Französisch	Deutsch	Rumänisch	Ungarisch	Russisch	Lettisch	Estnisch	Latein	Türkisch
1	66,8	55,8	51,7	45,5	34,4	33,9	33,8	28,8	24,2	18,8
2	21,1	27,9	31,6	28,4	30,4	30,3	38,4	40,2	32,1	37,8
3	8,2	12,9	11,1	18,6	20,7	21,4	19,4	18,5	28,7	37,0
4	3,3	2,9	4,3	6,7	9,9	9,7	7,0	8,8	11,6	12,1
5	0,5	0,5	1,0	0,7	3,4	3,5	1,2	2,5	2,8	3,6
>5	0,1	0,0	0,3	0,1	1,2	1,2	0,2	1,2	0,6	0,7
Mittlere Länge	1,50	1,64	1,82	1,89	2,22	2,23	2,04	2,20	2,39	2,46

Tabelle 3.3
Komplexe Verteilung der Wortformenlängen gemessen in Buchstaben und Silben in estnischer künstlerischer Prosa

Laute / Silben	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Zahl der Wortformen	Zahl der Silben	Zahl der Laute	Mittlere Länge in Lauten	
																		WF	Silben
1	118	86	53	4											261	261	717	2,75	2,75
2		46	108	150	77	19	2								402	804	1931	4,80	2,40
3				10	42	81	44	18	1						196	588	1393	7,11	2,37
4						4	24	41	20	9	3				101	404	924	9,15	2,29
5						1		1	4	5	10	2	4	1	28	140	328	11,71	2,34
6										2	2	4	2	2	12	72	156	13,00	2,17
Gesamt	118	132	161	164	119	105	70	60	25	16	15	6	6	3	1000	2269	5449	5,45	2,40

liegt "die proportionale Abnahme des relativen Wachstums" zugrunde (Land, 1977: 388-389), die man mit Hilfe der Differentialgleichung

$$\frac{dy/y}{dx/x} = \frac{b}{y} \tag{3.3}$$

ausdrücken kann.

Das bedeutet, daß das Verhältnis relativer Zuwächse von y und x (oder der Silbe und des Wortes) dem Wert von y (Silbenlänge) invers proportional ist. Dadurch erreicht man die für die logarithmische Kurve charakteristische verlangsamte Veränderung (Zunahme oder Abnahme) von y bezüglich der Zunahme von x.

Ein verallgemeinertes und inhaltlich interpretiertes Modell des Menzerathschen Gesetzes, d.h. das Modell für die Beschreibung des Zusammenhangs zwischen dem Konstrukt und seinen Komponenten, stammt von Altmann (1980). Das Modell wird als ein System von Gleichungen in der Form

$$y = ax^b e^{-cx} \tag{3.4}$$

dargestellt, wobei y die Komponentenlänge, x die Konstruktlänge und a, b, und c die Parameter sind. Wenn $b = 0$, dann erhält man die Exponentialfunktion $y = ae^{-cx}$, wenn $c = 0$, dann erhält man die Potenzfunktion $y = ax^b$. Das Modell wird mit Hilfe inhaltlicher Postulate und Differentialgleichungen interpretiert.

Es ist bekannt, daß die Wortlänge direkt oder indirekt mit vielen strukturellen Charakteristika des Vokabulars und des Textes verbunden ist. Betrachten wir nun den Zusammenhang zwischen der *Wortlänge* und der *Häufigkeit* seiner Verwendung im Text. Die Tatsache, daß kurze Wörter im Durchschnitt öfter vorkommen als lange, zeugt offensichtlich von der Wirkung des Ökonomieprinzips in der Kommunikation, von dem bereits Zipf (1935) und Martinet (1963) gesprochen haben. Aufgrund dieser Annahme hat der französische Forscher Guiraud (1954) eine Gesetzmäßigkeit formuliert, nach der man für jede Sprache eine Konstante (C) bestimmen kann, die den Zusammenhang zwischen Wortlänge (x) und seinem Rang (i) in einer Häufigkeitsliste erfaßt:

$$C = \frac{log\ 2i}{x}. \tag{3.5}$$

Diese Formel beschreibt mit hinreichender Genauigkeit den erwähnten Zusammenhang, wobei für das Englische $C \approx 0{,}59$ ist.

Im Grunde genommen drückt (3.5) das logarithmische Gesetz des Zusammenhangs zwischen Wortlänge und Worthäufigkeit aus. Diesen Zusammenhang kann man aus dem Zipf-Mandelbrotschen Gesetz ableiten. Die exakte Ableitung aus dem Zipf-Mandelbrotschen Gesetz zeigt in der Tat, daß die Wortlänge (x) dem Logarithmus ihrer Wahrscheinlichkeit proportional sein muß, d.h., sie muß vom

Logarithmus des Ranges (i) linear abhängen (Kalinin, 1964):

(3.6) $$x_i = a + b \ln i,$$

wobei a und b Konstanten sind. Diese exaktere (und theoretisch besser begründete) Formel beschreibt den Zusammenhang zwischen der Länge und Häufigkeit der Wortformen bei vielen Daten noch besser.

Eine andere Möglichkeit bietet die analytische Beschreibung des Zusammenhangs zwischen der kumulativen mittleren Wortlänge (X_i)

$$X_i = \frac{1}{i}\sum_{j=1}^{i} x_j$$

und ihrem Rang (i). Setzt man hier auch die logarithmische Abhängigkeit zwischen X_i und i voraus, dann kann man X_i mit der Formel

(3.7) $$X_i = \alpha + \beta \ln i$$

ausdrücken, wobei α und β Parameter sind. Die Übereinstimmung der empirischen mit den theoretischen Daten ist zufriedenstellend (Tuldava, 1986: 156). Üblicherweise nimmt man an, daß diese Abhängigkeit am besten mit der *Weibull-Verteilung* approximierbar ist, die für die Lösung dieses Problems zum ersten Mal von G.G. Belonogov (1962) benutzt wurde. Die Formel lautet

(3.8) $$X_i = X_n\left(1 - e^{-ci^k}\right)$$

wobei X_n das Maximum der mittleren Wortlänge (in der gegebenen Stichprobe), i der Rang des Wortes und c, k Parameter sind.

Die Abhängigkeit zwischen der Wortlänge und der Häufigkeit ist gegenseitig, auch wenn man sagen kann, daß die Verwendungshäufigkeit die entscheidende Rolle spielt (vgl. Tuldava 1995, Kap. 2). Schon seit langem ist bekannt, daß die Wortform einer Reduktion unterliegt, wenn die Häufigkeit des Wortes zunimmt (vgl. z.B. Zipf 1935, 1949; Martinet 1963). Damit erklärt man auch viele Kürzungen und Stutzungen in modernen Sprachen, wenn man z.B. entweder den Anfang oder das Ende des Wortes stutzt, Russisch: *kino*(*teatr*), *retro*(*spektivnyj stil'*); Estnisch: *kopter* < *helikopter;* Englisch: (*omni*)*bus.*

Über ein neues Projekt zur Erforschung der Wortlänge in verschiedenen Sprachen s. Best, Altmann (1996).

3.2. Der grammatische Aspekt

Lexik und Grammatik

Der grammatische Aspekt der Analyse ist mit der lexikalischen Erforschung des Wortes eng verbunden, da "in der realen Geschichte der Sprache die grammatischen und lexikalischen Formen und Bedeutungen organisch miteinander verbunden sind" und "jedes Wort schon dadurch geformt ist, daß es bekannte grammatische Funktionen trägt, eine bestimmte Stelle im grammatischen System der Sprache einnimmt" (Vinogradov, 1947: 7f). Im Rahmen der quantitativ-systemischen Untersuchung der Lexik interessiert uns die Möglichkeit der Klassifikation der Lexik aufgrund grammatischer Merkmale und die Bestimmung der Verteilungsgesetzmäßigkeiten der aufgestellten Klassen (lexikalischen Gruppen) im Wörterbuch und im Text. In der vorliegenden Arbeit beschränken wir uns auf einige Aspekte der *Wortbildung*, die uns ermöglicht, lexikalisch-formale Gruppen aufgrund der morphologischen Struktur zu bestimmen, und auf die Untersuchung lexikalisch-grammatischer Wortklassen, *Wortarten* genannt, die man im Rahmen der Grammatik mit Hilfe syntaktisch-morphologischer Kriterien bestimmen kann. Die Bestimmung grundlegender Wortbildungsgruppen und lexikalisch-grammatischer Wortklassen und die quantitative Analyse ihrer Strukturen und der Bedingungen ihrer Verteilung in konkreten Sprachen (oder Subsprachen) haben eine erstrangige Bedeutung für die typologische Charakterisierung der Sprachen (oder Subsprachen) und fördern die Lösung vieler aktueller Probleme der angewandten Linguistik.

Die erwähnten Probleme und Aufgaben gehören zu dem "Grenzgebiet" zwischen Lexikologie und Grammatik. Auch wenn einige Forscher die Wortbildung als zu der Lexikologie gehörend betrachten (Smirnickij 1956, Levkovskaja 1968), überwiegt heutzutage die Meinung, daß die Wortbildung zur Grammatik gehört oder eine selbständige Sparte der Linguistik bildet (vgl. Nemčenko, 1984:9). Dabei wird immer unterstrichen, daß Wortbildung eng mit der Lexikologie verbunden ist. Die Teile der Grammatik, wie Morphologie und Syntax, sind mit der Lexikologie hauptsächlich dadurch verbunden, daß sie als Basis für die Stratifikation der Lexik dienen. Wenn wir in der vorliegenden Arbeit die Wortarten als grundlegende Vertreter der syntaktisch-morphologischen Ebenen betrachten, dann schließen wir die Möglichkeit der Erforschung anderer lexikalisch-grammatischer Wortgruppen, die für die lexikalische Erforschung einer konkreten Sprache wichtig sind, nicht aus. In diesem Zusammenhang kann man die Untersuchungen in "klassifikatorischer Morphologie" erwähnen (Viks, 1980), die eine ergänzende Klassifikation der Lexik aufgrund der Formbildung (Flexionsmorphologie) liefert.

Die üblichste Klassifikation der Wörter nach ihrer morphologischen Struktur ist ihre Aufteilung in die grundlegenden Strukturtypen: *Stammwörter*, *Derivate* und *Komposita* (mit den Untertypen *abgeleitete Komposita, kontaminierte Wörter*, u.a.). Im normativen orthologischen Wörterbuch des Estnischen (Umfang etwa 115000 Wörter) sind die einzelnen Typen folgendermaßen verteilt:

Stammwörter	5%		
einfache Derivate	35%	Nicht-Komposita	40%
abgeleitete Komposita	25%		
einfache Komposita	35%	Komposita	60%

Es ist schwer, quantitative Charakteristika unterschiedlicher Sprachen zu vergleichen, wenn es keine exakten Angaben über die Verteilung der Subtypen und über den Stichprobenumfang gibt. Hakulinen (1979) macht beispielsweise folgende Angaben über das Finnische: Stammwörter 12%, Derivate 44%, Komposita 44%. Dabei weiß man nicht, wie der Anteil der einfachen und der zusammengesetzten Derivate ist und aus welchem Material die Stichprobe stammt. G. Papp ist etwas präziser: Er zeigt, daß es unter 31000 Substantiven des ungarischen akademischen Wörterbuches 55% Komposita (mit Suffix oder ohne) gibt; 42% der Substantive enthalten kein Suffix (wobei sie einfach oder zusammengesetzt sein können). Die Gesamtzahl ursprünglicher Stammwörter im ungarischen Wörterbuch ist 6000, d.h. 10,3% aus der Gesamtmenge von 58000 Lexemen (Papp, 1969).

Im Russischen etwa gibt es nach Angaben von Tichonov (1983) 13% einfache Stammwörter (im Material bestehend aus 145000 Wörtern aus mehreren Wörterbüchern), die restlichen 87% sind abgeleitet im weitesten Sinne (mit Affix oder zusammengesetzt). Die Anteile der Komposita im Russischen überschreiten 8% nicht (nach den Angaben des großen ukrainisch-russischen Wörterbuches mit 120000 Wörtern, vgl. Klimenko, 1974).

Im Wörterbuch der Pressetexte der modernen deutschen Sprache sind die Worttypen folgendermaßen verteilt: Stammwörter 5%, einfache Derivate 9%, Komposita und abgeleitete Komposita 83%, Abkürzungen 3% (Harlass, Vater 1974). Im Englischen (Subsprache nicht angegeben) gibt es unter den Substantiven 18% Stammwörter, 67% abgeleitete Wörter (affixale und nicht-affixale) und 15% Komposita (Ginzburg et al. 1966).

Daraus kann man schließen, daß in allen untersuchten (finno-ugrischen und indoeuropäischen) Sprachen Stammwörter nur einen kleinen Anteil ausmachen (5...18%), während die Sprachen sich vor allem durch den Anteil von Komposita unterscheiden: 45...60% in finno-ugrischen Sprachen, etwa 80% im Deutschen, 15% im Englischen und weniger als 10% im Russischen.

Bei der Verwendung im *Text* findet man ganz andere Verhältnisse. Im Deutschen und Englischen (Pressetexte) beispielsweise machen Stammwörter 70 bzw. 75%, abgeleitete Wörter 18 bzw. 23% und Komposita 12 bzw. 2% aus (Kubrjakova, 1970). Im Estnischen machen einfache Stammwörter 70...80%, abgeleitete und zusammengesetzte zusammen 20...30% des Textes aus (die Schwankungen hängen von der Textsorte ab). Wie bekannt, wurde aufgrund der Angaben über Verwendungshäufigkeit unterschiedlicher morphologischer Worttypen von Greenberg eine typologische Klassifikation von Sprachen aufgestellt (1960). Anstelle von Prozenten benutzte Greenberg besondere Indizes, die das Verhältnis spezieller Morpheme zur Wortanzahl im Text (Wortverwendungen) ausdrücken. Beispielsweise gibt der Kompositionsindex (compositional index) das Verhältnis von Wurzelmorphemen (R) zur Zahl der Wörter im Text (W) an. Laut präzisierter Daten von Kubrjakova (1970) ergibt der Index R/W für Pressetexte in germanischen Sprachen:

Englisch	1,02
Niederländisch und Dänisch	1,11
Deutsch	1,12
Schwedisch	1,13
Isländisch	1,14

Im estnischen Zeitungstext hat der Index den Wert von 1,20, im russischen Zeitungstext 1,04 (d.h. auf 100 Wörter kommen durchschnittlich 104 Wurzelmorpheme, anders ausgedrückt 4 Komposita auf 100 Wörter des Textes, wenn man berücksichtigt, daß ein russisches Kompositum üblicherweise aus zwei Wurzeln besteht).

Das Verhältnis der Morphemzahl (Wurzelmorpheme, Wortbildungsmorpheme oder Flexionsmorpheme) zur Zahl der laufenden Wörter (*M/W*) mißt das Ausmaß des Synthetismus (vgl. Abschnitt 2.1). Es ist offensichtlich, daß die Werte dieses Indexes für analytische Sprachen klein sein werden, für synthetische Sprachen groß. Für die germanischen Sprachen ergeben sich folgende Resultate (Kubrjakova, 1970:161):

Englisch	1,43
Niederländisch	1,81
Dänisch	1,98
Deutsch	2,02
Isländisch	2,09
Schwedisch	2,13

Im Estnischen (Zeitungsartikel) erreicht der Index den Wert 2,35. Noch höhere

Werte erreicht der Index in Sprachen wie Sanskrit, Suahili, Eskimo (vgl. Greenberg 1960).

Eine andere Klassifikation der Lexik erhält man, wenn man von den Wortbildungsklassen der Wörter ausgeht (Belonogov u.a., 1985). Eine Klasse wird bestimmt durch die Liste der Suffixe und Suffixkombinationen, die sich mit dem wortbildenden Stamm verbinden.[3] Im Russischen wurden etwa 1250 Wortbildungsklassen beschrieben, wobei die 10 häufigsten Klassen 60% aller russischen Wörter einschließen. Hier äußert sich das bekannte Prinzip der Konzentration und Dispersion der Einheiten, das zur Bildung des "Kerns" und der "Peripherie" in der Verteilung sprachlicher Objekte führt. Zu den häufigsten Klassen gehören (eine Klasse wird durch das repräsentierende Wort und die Liste der Suffixe bestimmt; "θ" bedeutet Nullsuffix):

masštab:	θ, -n-	(d.h. Wörter wie: *masštab, masštabnyj)*
slab-yj:	θ, -o-, -ost-	*(slabyj, slabo, slabost')*
port:	θ, -ov-	*(port, portovyj).*

Im Russischen gibt es insgesamt etwa zehntausend wortbildende Stämme, wobei die Zahl der Wortbildungssuffixe und ihrer Kombinationen größer als 1000 ist (Belonogov u.a.,1985). Auf der anderen Seite hat man im Russischen etwa zehntausend morphologische Wortbildungsmodelle bei einer Gesamtzahl von 5000 Morphen erfaßt, darunter 4500 Wurzelwörter, 425 unterschiedliche Suffixe und 75 Präfixe (Efremova, 1968).

Die Verteilung der Derivate nach ihren Wortbildungsformanten, d.h. Präfixen und Suffixen, erlaubt es, die produktivsten und häufigsten Typen zu bestimmen[4]. Nach den Daten aus dem *Obratnyj slovar' russkogo jazyka* (Bielfeldt, 1965) sind beispielsweise die häufigsten Suffixe *-nyj/noj* (etwa 9800 Wörter des Types *trudnyj, lesnoj*), *-nie/-ie* (3200: *penie, šestvie)*, *-ka/-očka* (3000: *rečka, lampočka*), *-skij/-skoj* (2700: *russkij, tverskoj)*, *- ost'* (2500), *-nik* (1200) u.a. (vgl. Arapov, Cherc, 1983). Im *Großen deutsch-russischen Wörterbuch* von O.I. Moskal'skaja (nach Bartkov, 1983) sind die häufigsten Präfixe des Deutschen: *ver-* (1100 Wörter), *un-* (885), *be-* (783), *er-* (308), *ent-* (270); die häufigsten Suffixe sind *-ung* (10000), *-ig* (3800), *-er* (3000), *-isch* (2100), *-keit* (2000). Im Text (Pressetexte) ordnen sich die Wörter nach der Häufigkeit der Formanten wie folgt: *be-, ver-, er-, ge-, ent-*, und *-ung, -er, -isch, -lich, -ig.*

[3] Der wortbildende Stamm des Wortes wird als der Anfangsteil des Wortes bestimmt, der die größte Anzahl von Suffixen erhält.

[4] Bartkov (1982) unterscheidet "diachronische Produktivität" (Klassenumfang, d.h. Häufigkeit in einem großen Wörterbuch) und "synchrone Produktivität" (aus dem Wörterbuch der Neologismen). Unter Verwendung verstehen wir die Häufigkeit im Text.

Im Unterschied zu lexikalisch-phonetischen Gruppen (vgl. Abschnitt 3.1), bei denen die logarithmische (oder exponentielle) Abhängigkeit in der Rangverteilung der Wörter erörtert wurde, unterliegt die Verteilung der Wörter nach Häufigkeit ihrer Wortbildungsformanten in der Regel einem Potenzgesetz, dem Zipfschen Gesetz. Arapov zeigte (1975), daß die Rangverteilung der Worthäufigkeiten mit einem gegebenen Suffix in einem hinreichend großen Wörterbuch annähernd nach dem Zipfschen Gesetz mit Mandelbrots Modifikation verläuft, d.h. die Verteilung kann mit der Funktion des Typs $p_i = k(i+B)^{-\gamma}$ (vgl. Abschnitt 2.2) approximiert werden, wobei p_i die relative Häufigkeit des Wortes mit gegebenem Formanten, i der Rang und k, γ, B Konstanten sind.

Das Verhältnis der Häufigkeit im Text (F_T) zu der Häufigkeit im entsprechenden Wörterbuch (F_W) drückt die funktionale Bedeutung oder (relative) *funktionale Belastung* des gegebenen Typs der Derivate aus. Im Englischen beispielsweise (aufgrund des Materials des Häufigkeitswörterbuches von Kučera und Francis bei $F \geq 5$; vgl. Pikver, 1973) haben im aufgeführten Fragment (Tabelle 3.4) Substantive mit Suffixen *-y, -ment* und *-ion* (z.B. *policy, government, action*) und die Adjektive mit den Suffixen *-ent* und *-ic* (*different, economic*) die größte funktionale Belastung. Das Verhältnis F_T/F_W drückt eigentlich die mittlere Häufigkeit des gegebenen Worttyps im Text aus: Je größer dieses Verhältnis, desto häufiger wiederholen sich Wörter mit dem gegebenen Formanten im Text, aber desto kleiner ist auch ihre relative Differenzierung. Substantive mit dem Suffix *-er* mit kleiner funktionaler Belastung im Text (FB = 6,5) haben eine große Menge von Derivaten. Das umgekehrte Verhältnis (F_W/F_T) drückt daher die Differenziertheit des gegebenen Typs der Derivate aus.

Die quantitative Untersuchung kann man auch auf die sogenannten wortbildenden Nester ausweiten, unter denen man Gruppen einstämmiger Wörter versteht, die aufgrund ihrer Wortbildungsproduktivität zusammengefaßt werden. Üblicherweise bezeichnet man das Zentrum des Nestes mit dem motivierenden Wort. Beispielsweise kann man im Englischen ein Nest mit dem Zentrum *TIME* bilden, das sich durch den größten derivationalen und lexikalischen Umfang auszeichnet: Nach Angaben aus mehreren englischen Wörterbüchern enthält es etwa 100 Derivate (Beljaeva, Vasil'eva, 1984). Die Fähigkeit eines Wortes, als produktive Basis zu dienen (sowohl für die Affixation als auch für die Zusammensetzung), heißt *Wortbildungspotential* des Wortes. Im Russischen haben Substantive wie z.B. *voda* ("Wasser") (mit 316 Derivaten und Komposita), *svet* ("Licht", "Welt") (306), *zemlja* ("Erde", "Land") (216); Verben wie *byt'* ("sein") (446), *brat'* ("nehmen") (393), *delat'* ("tun, machen") (318); Adjektive wie *belyj* ("weiß") (246), *černyj* ("schwarz") (236), *staryj* ("alt") (192) u.a. (Tichonov, 1983) großes Wortbildungspotential.

Aufgrund von Schätzungen des Wortbildungspotentials kann man quantitative Urteile über Wortbildungsnester fällen und Wörter klassifizieren, z.B. durch Auf-

Tabelle 3.4
Häufigkeit und funktionale Belastung der Wortformen mit gegebenem Suffix im Englischen

Rang	Suffix	Häufigkeit im Text (F_T)	Häufigkeit im Wörterbuch (F_W)	Funktionale Belastung (F_W/F_T)
1.	-ion	12772	794	16,1
2.	-ly (adv.)	10627	942	11,3
3.	-al (adj.)	7937	612	13,0
4.	-ate (verb)	7782	815	9,6
5.	-er (noun)	5221	803	6,5
6.	-ment	4991	296	16,9
7.	-ic	4691	396	15,9
8.	-y (noun)	3650	186	19,6
9.	-ent (adj.)	3486	192	18,2
10.	-ity	3196	350	9,1

teilung des Wörterbuchs in Nester mit hohem, mittlerem und niedrigem Potential aufteilen. Innerhalb des Nestes kann man aufgrund der Verwendungshäufigkeit einzelner Elemente des Nestes Kern und Peripherie bestimmen.

Aufgrund von empirischen Daten ist eine statistische Abhängigkeit zwischen den Schätzungen des wortbildenden Potentials und der Häufigkeit der erzeugenden Wörter festzustellen. Die häufigsten Wörter haben im Durchschnitt das größte Wortbildungspotential; die Korrelation zwischen der Häufigkeit des Zentrums und der Potenz des Wortbildungsnestes ist statistisch signifikant (Bartkov, 1983). Bei der analytischen Untersuchung des Zusammenhangs zwischen der Häufigkeit und der Schätzung des Potentials kommt eine komplizierte Abhängigkeit zu Tage, die eine logistische Form hat, d.h. bei hinreichend umfangreichem Material beobachtet man ein langsames Anwachsen des Potentials mit wachsender Häufigkeit, danach eine starke Wendung im Bereich der durchschnittlich häufigen Wörter und eine Stabilisierung im Bereich hochfrequenter Wörter (s. Andrukovič, Korolev, 1977). Die Resultate der konkreten Untersuchung der Lexik im Hinblick auf die Beziehung zwischen Häufigkeit und Potential wurden in der Praxis bei der Erstellung von Thesauri und bei der Optimierung der linguistischen Absicherung einer Reihe von automatischen Informationssystemen benutzt (Korolev u.a., 1984).

Die Gruppierung der Wörter in lexikalisch-grammatische Klassen, sogenannte Wortarten, geschieht auf der Grundlage einer gleichzeitigen Einbeziehung mehrerer Faktoren: der lexikalisch-grammatischen (kategorialen) Bedeutung, morphologischer Eigenschaften und syntaktischer Funktionen. Ungeachtet der Unsicherheiten bei ihrer Bestimmung stellen Wortarten prinzipiell stabile Kategorien dar, die man mit hinreichend objektiven Kenngrößen charakterisieren kann. Die Wichtigkeit der Untersuchung der Wortarten, darunter auch ihres quantitativen Aspekts, wird durch den Umstand unterstrichen, daß in der Form der Verteilung der Wortarten im Wörterbuch und im Text wichtige typologische Eigenschaften der gegebenen Sprache, Subsprache oder des Stils enthalten sind.

Die Verteilung der Wortarten im *Lexikon* hängt vom Umfang des Lexikons ab, wobei man einen stetigen Zuwachs des Anteils von Substantiven bei Vergrößerung des Lexikonumfangs beobachten kann. Wenn man beispielsweise die Wörterbücher der estnischen Lexeme nach der Größe der Stichproben vergleicht, dann stellt man fest, daß die Proportion der Substantive von 44,2 über 61,6 bis 75,0 anwächst (vgl. Tabelle 3.5). Eine umgekehrte Tendenz kann man bei allen anderen Wortarten beobachten, außer bei den Adjektiven, die im gegebenen Fall eine recht stabile Häufigkeit im Lexikon haben (etwa 11%). Die dominierende Stelle der Substantive im großen Lexikon ist teilweise dadurch bedingt, daß Substantive die Hauptquelle der Anreicherung des Lexikons mit neuen Wörtern darstellen.

Tabelle 3.5
Verteilung der Wortarten (in %) im estnischen Wörterbuch in Stichproben unterschiedlichen Umfangs

Wortart	I	II	III
Substantiv	44,2	61,6	75,0
Verb	26,6	14,0	8,5
Adjektiv	9,5	11,7	11,4
Adverb	10,7	9,6	4,2
Pronomen	4,4	0,4	0,07
Zahlwort	1,3	1,0	0,2
Prä- und Postposition	2,2	1,0	0,1
Konjunktion	0,9	0,2	0,03
Interjektion	0,2	0,5	0,5
Gesamt	100,0	100,0	100,0
Vokabularumfang	2200	14650	115000
Textumfang	5000	100000	-

I - Stichprobe, II - Häufigkeitswörterbuch der Lexeme, III - orthologisches Wörterbuch (Õigekeelsussônaraamat 1976).

Die Veränderungsdynamik der Lexikonstruktur kann man auch an der Verteilung der Wortarten nach Häufigkeitszonen zeigen (Tabelle 3.6). Beispielsweise machen die Substantive im estnischen Häufigkeitswörterbuch in der hochfrequenten Zone ($F \geq 10$) 35,2% und Verben 24,5% aus, aber in der Zone der einmalig vorkommenden Wörter ($F = 1$) beträgt ihre Häufigkeit 65,8% und die der Verben 11,1%. Substantive haben die Tendenz, in der Zone seltener Wörter vorzukommen, während Verben (auch Adverbien, Pronomina u.a.) meistens hoch- oder mittelfrequent sind. Die Adjektive weisen auch hier eine Stabilität auf. (Für ähnliche Angaben über das Russische s. Jiráková 1976; für das Lettische s. Jakubajtis 1981).

Die Häufigkeit der Wortarten im Text hängt nicht wesentlich von der Stichprobengröße ab, daher stellt ihre Verteilung im Text einen wichtigen stilunterscheidenden Faktor dar. Individuelle Unterschiede in den Verteilungen der Wortarten bewegen sich in bestimmten Grenzen, die vom Funktionalstil vorgegeben sind. Den größten Unterschied in der Wortartenverwendung findet man vor allem zwischen künstlerischen und nichtkünstlerischen Stilen. Beispielsweise gibt es nach den Angaben des russischen Häufigkeitswörterbuches (1977) folgende Substantivhäufigkeiten : Zeitungen 32,8%, wissenschaftlich-publizistische Texte 31,0%, künstlerische Prosa 23,4%, Dramen 20,4%. Die Häufigkeiten der Pronomina sind umgekehrt geordnet: Dramen 16,2%, künstlerische Prosa 14,9%, wissenschaftlich-publizistische Texte 11,6%, Zeitungen 10,0%. Die Häufigkeit der Verben: Dramen 20,9%, künstlerische Prosa 19,0%, Zeitungen 14,5%, wissenschaftlich-publizistische Texte 13,5%.

Tabelle 3.6
Veränderungsdynamik der Wortartenhäufigkeiten (in %) in Häufigkeitszonen im estnischen Häufigkeitswörterbuch der Lexeme in künstlerischer Prosa

Häufigkeit / Wortart	$F \geq 10$	9	8	7	6	5	4	3	2	1	Σ
Substantiv	35,2	40,6	43,3	37,9	48,5	50,6	45,5	54,1	57,3	65,8	61,6
Verb	24,5	22,6	27,0	29,3	22,7	23,8	20,7	20,5	17,8	11,1	14,0
Adjektiv	11,6	15,1	14,2	12,6	13,3	12,5	15,8	12,8	11,7	13,1	11,7
Adverb	16,8	17,0	12,0	14,7	11,2	10,9	15,2	10,4	11,5	9,0	9,6
Pronomen	3,4	0,0	0,0	0,0	0,0	0,3	0,7	0,6	0,3	0,1	0,4
Zahlwort	1,3	0,9	1,4	1,0	2,2	0,6	0,6	0,6	0,6	0,3	1,0
Prä- und /Postposition	5,6	2,9	1,4	3,0	1,7	1,3	0,8	0,9	0,4	0,1	1,0
Konjunktion	1,4	0,0	0,0	0,5	0,0	0,0	0,0	0,1	0,0	0,0	0,2
Interjektion	0,2	0,9	0,7	1,0	0,4	0,0	0,7	0,0	0,4	0,5	0,5

Innerhalb eines Genres kann man eine starke Wechselbeziehung und gegenseitige Bedingtheit bei der Verwendung der Wortarten im Text beobachten. So besteht z.B. eine signifikante negative Korrelation zwischen der Verwendung der Substantive und der Pronomina, d.h. diese Wortarten "konkurrieren" miteinander bei der Bezeichnung von Dingen und Erscheinungen. Ein negativer Zusammenhang besteht in der Regel auch zwischen den Häufigkeiten von Verben und Adjektiven: Dieser ist durch den stilistischen Unterschied zwischen "Aktivität" und "Deskriptivität" bedingt. Die Verhältnisse der Verwendungshäufigkeit der Wortarten (Substantive und Verben, Adjektive und Verben usw.) benutzt man in der Stilistik und in der Psychologie bei der Untersuchung des Individualstils als stilometrische Indizes für "Nominalität", "Verbalität" oder "Aktivität" usw. (vgl. z.B. Busemann, 1948; Antosch, 1969; Golovin, 1971; Tuldava, 1976; Jakubajtis, 1981).

Beim Vergleich von Daten aus verschiedenen Sprachen zeigt sich, daß bei Wahrung der Homogenität der Texte die Unterschiede nicht allzu groß sind. Man kann die dominierende Rolle der Substantive in der künstlerischen Prosa in finnougrischen Sprachen beobachten (Estnisch, Finnisch, Ungarisch), wo ihr Anteil 30...31% beträgt, dagegen im Lettischen, Russischen und Ukrainischen im Durchschnitt 28...29% (vgl. Tabelle 3.7).

Tabelle 3.7
Häufigkeitsverteilung der Wortarten (in %) in *Texten* der künstlerischen Prosa verschiedener Sprachen (nach Angaben von Saukkonen et al., 1979; Zsilka, 1973; Jakubajtis, 1981; Kločkova, 1968; Tiščenko, 1970)

Sprache / Wortart	Estnisch	Finnisch	Ungarisch	Lettisch	Russisch	Ukrainisch
Substantiv	31,7	29,7	30,0	28,1	28,7	29,2
Verb	22,5	27,6	22,4	23,1	18,3	19,7
Adjektiv	6,0	7,8	10,0	5,4	7,9	6,8
Adverb	15,8	10,9	8,0	10,0	6,0	6,2
Pronomen	11,4	11,2	5,7	14,6	10,2	9,0
Zahlwort	1,1	1,2	1,5	1,2	1,2	1,1
Prä-/Postposition	3,1	3,6	21,9	5,1	12,2	12,5
Konjunktion	8,2	7,3	-	7,3	8,5	9,1
Interjektion	0,2	0,7	0,5	0,6	0,0	0,0
Partikel	-	-	-	4,6	7,0	6,4
Entropie	2,62	2,60	2,47	2,79	2,85	2,82

Der Vergleich der Häufigkeitsverteilungen von Wortarten wird oft mit Hilfe der Entropie nach der Formel von Shannon (1948) durchgeführt:

(3.9) $$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

wobei H die Entropie, p_i die relativen Häufigkeiten und log_2 der Logarithmus zur Basis 2 sind. Das Entropiemaß zeigt den Grad der "Unbestimmtheit" bei der Wahl der Wörter, oder mit anderen Worten, große Entropie ist assoziiert mit einem großen Maß an Gleichverteilung der Einheiten. Estnisch und Finnisch stehen sich sehr nahe (H = 2,62 bzw. H = 2,60) in Unterschied z.B. zum Russischen (H = 2,85), wo die Verteilung der Wortarten im Text gleichmäßiger ist.

Was die Verteilungsform der Wortarten im Text betrifft (die sog. "Einzelobjektverteilung", vgl. Abschnitt 1.3) so kann man in Bezug auf die Untersuchung von Jakubajtis (1981) feststellen, daß sich die Verteilung der Wortarten (sowohl der sehr häufigen wie der Substantive, Verben als auch der selteneren wie der Präpositionen) aufgrund von 100 Stichproben von jeweils 1000 Wortverwendungen, unter den gegebenen experimentellen Bedingungen, in allen Subsprachen des Lettischen mit der Normalverteilung approximieren läßt.

3.3. Der semantische Aspekt

Lexikalisch-semantische Gruppen (LSG)

Eine der grundlegenden Aufgaben der lexikalischen Untersuchungen unter dem Gesichtspunkt ihrer systemischen Eigenschaften ist die Strukturierung in LSG. Dies ist eine allgemeine Bezeichnung für lexikalische Gruppen, die sich auf der lexikalisch-semantischen Ebene bilden, d.h. aufgrund des semantischen Zusammenhangs zwischen den Wörtern. Es gibt verschiedene Gruppen und Typen von LSG.

Aufgrund der Beziehung zwischen den Wörtern in den LSG unterscheiden wir Gruppen mit paradigmatischen Beziehungen (Übereinstimmung, Gegenteil, Subordination) und syntaktischen Beziehungen (gleichzeitiges Vorkommen, Kombinierbarkeit). Die paradigmatischen Beziehungen der Wörter konstituieren die Möglichkeit der Wahl zwischen Einheiten bei der Redeerzeugung, woraus die Wichtigkeit der Untersuchung der paradigmatischen LSG sowohl im qualitativen als auch im quantitativen Bereich hervorgeht (die Rolle des syntagmatischen Faktors, der mit dem paradigmatischen in Wechselwirkung steht, wird dadurch nicht geschmälert).

Die LSG kann man auf zwei Weisen untersuchen. Zum einen ist der synthetische Ansatz möglich, wenn man nämlich bei der Gruppenbildung vom einzelnen Element ausgeht (vom Speziellen zum Allgemeinen), zum anderen der analytische, wenn man von der Universalmenge zu Teilmengen vorstößt (vom Allgemeinen zum Speziellen).

Die grundlegende Einheit auf der lexikalisch-semantischen Ebene ist die lexikalisch-semantische Variante (LSV), von der ausgehend man beim synthetischen Ansatz die LSG bildet. Die LSV kann auch beim quantitativen Ansatz die grundlegende Zähleinheit sein, wenn es gelingt, die einzelnen Bedeutungen des Wortes exakt zu bestimmen.

Beim analytischen Ansatz erfolgt die Gliederung des lexikalisch-semantischen Systems der Sprache aufgrund der distinktiven semantischen Merkmale, die sich voneinander durch ihren Allgemeinheitsgrad unterscheiden. Bei der Organisierung der distinktiven semantischen Merkmale kann man das Prinzip der Dichotomie einsetzen, bei dem jedes Merkmal in Form einer binären privativen Opposition *A/non-A* dargestellt wird, wobei ein Glied durch Anwesenheit, das andere durch Abwesenheit des Merkmals charakterisiert ist (man kann dies auch als "+/-" oder "1/0" darstellen; der letzte Ausdruck ist quasi die Bestimmung der Wahrscheinlichkeit oder des Maßes der Zugehörigkeit des Elements zu der Menge[5]). Möglich ist auch ein System von Merkmalen, in dem man nicht nur binäre Oppositionen, sondern eine komplexere Skala verwendet. Das dichotomische Prinzip wählt man hauptsächlich deswegen, weil binäre Oppositionen kategorialer Merkmale wie *nominal - nicht nominal, konkret - nicht konkret* u.a. einen scharfen inhaltlichen (stilistischen) Sinn haben und fruchtbar und anschaulich bei der quantitativen Analyse der Verteilung der Wörter verwendet werden können. Da die angesetzten Merkmale nicht gleichwertig sind, ist es wichtig, im Merkmalssytem und folglich auch im System der LSG Hierarchien (Inklusionen) zu bilden. Das *Modell* eines derartigen Systems kann man schematisch als einen Entscheidungsbaum darstellen (s. Abb. 3.3). Es ist zu bemerken, daß man jeden Ast des dichotomischen Baumes in Abhängigkeit von der Aufgabe der Untersuchung verlängern oder abtrennen kann. Wie sich herausstellt, gehört eine gegebene Menge von distinktiven semantischen Merkmalen einer bestimmten Menge von Wörtern an, die man auf dieser Grundlage zu Gruppen (LSG) von verschieden Umfängen und auf verschiedenen Ebenen zusammenfaßt. Die Zugehörigkeit einzelner Wörter oder ganzer LSG kann man als eine Folge von distinktiven semantischen Merkmalen (z.B. *nominal - konkret - leblos usw.*) oder mit Hilfe des Binärkodes, d.h. als eine Folge von Einsen und Nullen in der Reihenfolge der Äste im binären Baum bestimmen. Die letzte Methode eignet sich für Prozeduren der automatischen Klassifikation in Verbindung mit quantitativer Zählung.

Als Beispiel bringen wir die quantitativen Angaben zu estnischen Substantiven in einem gekürzten Schema distinktiver semantischer Merkmale (s. Abb. 3.3). Das Material stammt aus dem Häufigkeitswörterbuch der Autorensprache in estnischer

[5] Theoretisch ist auch ein komplizierteres Modell möglich, in dem man die gesamte Skala der Wahrscheinlichkeiten von 0 bis 1 oder (in der Theorie der unscharfen Mengen) den Grad der Zugehörigkeit betrachtet.

Kunstprosa und aus dem Häufigkeitswörterbuch eines vollständigen künstlerischen Werkes (Roman "Wahrheit und Recht" von A.H. Tammsaare) getrennt nach Autorensprache und Sprache der Personen (detaillierter s. Tuldava, 1983a).

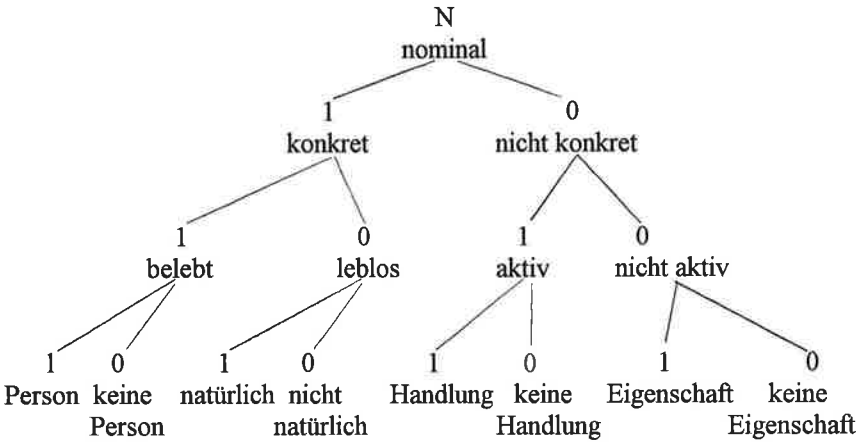


Abbildung 3.3. Der binäre Baum distinktiver semantischer Merkmale für Substantive

Nach dem binären kategorialen Merkmal *konkret - nicht konkret* kann man die Nomina in zwei große Gruppen aufteilen: konkrete und abstrakte. Nach den Angaben der aufgeführten Wörterbücher sind die Substantive folgendermaßen verteilt (in %):

	Künstlerische Sprache	Autorensprache in Tammsaare	Sprache der Personen in Tammsaare
Konkrete	60,9	55,2	65,6
Abstrakte	39,1	44,8	34,4

Konkrete Substantive überwiegen in allen drei Wörterbüchern, man kann aber einen statistisch signifikanten Unterschied zwischen der "Norm" (zusammengesetztes Wörterbuch der künstlerischen Prosa) und dem individuellen Wörterbuch der Autorensprache, sowie zwischen den Wörterbüchern der Autorensprache und der Sprache der Personen in dem gleichen künstlerischen Werk feststellen.

In den entsprechenden *Texten* ist der Anteil der konkreten Substantive noch höher:

	Künstlerische Sprache	Autorensprache in Tammsaare	Sprache der Personen in Tammsaare
Konkrete	63,2	65,6	70,8
Abstrakte	36,8	34,4	29,2.

Die konkreten Substantive decken daher etwa 60...70% des estnischen künstlerischen Textes ab, wobei der größte Anteil konkreter Substantive (70,8%) in der Personensprache vorkommt. Das Übergewicht konkreter Substantive in diesen Texten erklärt sich durch die Spezifizität des künstlerischen Stils im Vergleich mit anderen Funktionalstilen. Dies wird auch durch Untersuchungen in anderen Sprachen bestätigt. In deutschen Texten beispielsweise ist der Anteil konkreter Substantive in der künstlerischen Prosa 73% und im wissenschaftlich-technischen Text nur 28% (Kul'gav, 1971: 18-19). Es ist zu bemerken, daß abstrakte Substantive üblicherweise dort überwiegen, wo man eine Darstellung allgemein oder knapp gestalten will.

Bei der weiteren Gliederung der LSG *konkreter* Substantive kann man kategorial-semantische Merkmale von niedrigerem Allgemeinheitsgrad benutzen (*Belebtheit - Unbelebtheit, Person - Nicht-Person* usw.; vgl. Abb. 3.3). Aufgrund dieses Schemas erkennt man innerhalb der konkreten Substantive vier grundlegende Untergruppen: LSG von Menschenbezeichnungen (N 1 1 1), LSG von Tierbezeichnungen (N 1 1 0), LSG von natürlichen Gegenständen (N 1 0 1) und LSG von Artefakten (N 1 0 0). Ihre allgemeine Häufigkeitsverteilung (in %) im Vokabular und im Text der estnischen künstlerischen Prosa findet man in Tabelle 3.8.

Tabelle 3.8
Verteilung der Subgruppen konkreter Substantive

LSG (Typ)	Wörterbuch KS	Wörterbuch AS	Wörterbuch SP	Text KS	Text AS	Text SP
N 1 1 1	23	21	24	23	29	35
N 1 1 0	4	6	9	3	6	7
N 1 0 1	25	26	25	33	33	28
N 1 0 0	48	47	42	41	32	30
Σ (%)	100	100	100	100	100	100

Wie man aus diesem Vergleich sehen kann, ist die Verteilung der Häufigkeiten der LSG konkreter Substantive auf der kategorial-semantischen Ebene, trotz einiger individueller Unterschiede, im Großen und Ganzen recht stetig und kann als Cha-

rakteristikum künstlerischer Prosa der gegebenen Sprache dienen. Am häufigsten ist die LSG der Bezeichnungen von Artefakten (N 1 0 0). Unterschiede zwischen Verteilungen in einzelnen Stilen treten im Text stärker in den Vordergrund als im Vokabular.

Die Untersuchung der Verteilung der LSG *abstrakter* Substantive in Teilgruppen führt man auch aufgrund binärer kategorial-semantischer Merkmale niedrigerer Allgemeinheitsstufe (*aktiv - nichtaktiv, Handlung - Nichthandlung* u.a.) durch. Auf dieser Grundlage kann man die abstrakten Substantive in vier Grundteilgruppen aufteilen: LSG der Substantive, die eine Tätigkeit ausdrücken (N 0 1 1), LSG der Substantive, die einen Zustand oder das Resultat einer Handlung ausdrücken (N 0 1 0), LSG der Substantive, die eine Eigenschaft ausrücken (N 0 0 1) und die LSG anderer Typen abstrakter Substantive (N 0 0 0). Ihre Verteilung im Vokabular und im Text estnischer künstlerischer Prosa (s. Tabelle 3.9.) zeigt, daß, wenn man die "Restgruppe" (N 0 0 0) nicht einbezieht, die erste Stelle nach Vorkommenshäufigkeit sowohl im Vokabular als auch im Text diejenige LSG einnimmt, die aktive Handlungen (N 0 1 1) ausdrückt.

Tabelle 3.9
Häufigkeitsverteilung der Untergruppen abstrakter Substantive

LSG (Typ)	Wörterbuch KS	AS	SP	Text KS	AS	SP
N 0 1 1	37	43	27	22	25	18
N 0 1 0	10	10	7	9	10	6
N 0 0 1	5	8	7	5	5	8
N 0 0 0	48	39	59	64	60	68
Σ (%)	100	100	100	100	100	100

Aufgrund der Häufigkeitsverteilung der Untergruppen bilden die abstrakten in etwas höherem Maße als konkrete Substantive ein stildifferenzierendes Charakteristikum in der künstlerischen Prosa. Auch die Unterschiede zwischen Stilen zeigen sich an der Verteilung der abstrakten Substantive hinreichend klar nicht nur im Text (wie bei konkreten Substantiven) sondern auch im Vokabular, d.h. bei der Gliederung des jeweiligen Inventars der Wörter in lexikalisch-semantische Gruppen.

Polysemie

In der modernen Linguistik wird die Polysemie als ein wichtiges Universale betrachtet, die in verschiedenen Sprachen verschiedene Formen annehmen kann, aber grundsätzlich allgemeinen Gesetzen folgt. In der letzten Zeit erschien eine Reihe von Untersuchungen, in denen die Frage nach der Systemizität der Polysemie unter quantitativem Aspekt angesprochen wurde (s. z.B. Polikarpov, 1976, 1987; Višnjakova, 1976; Krylov, Jakubovskaja, 1977; Andrukovič, Korolev, 1977; Tuldava, 1979; Köhler, 1986). Das große Interesse für quantitative Gesetzmäßigkeiten der Polysemie in der heutigen Zeit erklärt sich vor allen Dingen mit praktischen Bedürfnissen der Lexikographie und der automatischen Textverarbeitung (Informationssuche, automatische Übersetzung). Die Entdeckung quantitativ-systemischer Charakteristika der Polysemie kann gleichzeitig ein neues Licht auf die allgemeinen Gesetzmäßigkeiten der Funktion des lexikalisch-semantischen Systems der Sprache insgesamt werfen. Eine vergleichende Untersuchung der quantitativen Seite der Polysemie kann ergänzende Kriterien für die semantische Typologie liefern.

Untersucht man die Polysemie als System, dann ist es nicht zweckmäßig, sie von der Monosemie getrennt zu betrachten. Man muß sie als Ausprägung einer einzigen Eigenschaft betrachten, nämlich, eine oder mehrere Bedeutungen zu haben. Monosemie tritt in diesem Fall als die "Nullstufe" der Polysemie auf. Die Realisierung der Eigenschaft "Zahl der Bedeutungen" ist der *semantische Umfang* der Wortes, den man quantitativ messen kann. Dabei gehen wir davon aus, daß ein polysemes Wort (Lexem) als Einheit des Sprachsystems ein einziges Bedeutungsganzes darstellt, das im Bedeutungsraum eine Reihe von virtuellen semantischen Varianten oder unterschiedliche "Bedeutungen" des Wortes vereinigt. Es gibt verschiedene Arten der Bestimmung des Sinninhaltes des Wortes und der Abgrenzung seiner einzelnen Bedeutungen. Hierbei gibt es beträchtliche Diskrepanzen in der praktischen Lexikographie. Der Versuch der Erforschung der Polysemie zeigt jedoch, daß die Abgrenzung der Wortbedeutungen innerhalb eines erklärenden Wörterbuches recht konsequent verfolgt wird. Dies bezeugen erfolgreiche Versuche der Erforschung allgemeiner quantitativer Gesetzmäßigkeiten der Polysemie auf der Grundlage derartiger Wörterbücher (s. Papp, 1969; Krylov, Jakubovskaja, 1977).

Die empirischen Daten bezeugen, daß es zwischen den Kenngrößen des semantischen Umfangs verschiedener Wortarten signifikante Unterschiede gibt. Beispielsweise ergibt sich aus den Daten von Papp (1967), daß im vollständigen erklärenden Wörterbuch des Ungarischen die mittlere Zahl der Bedeutungen bei den Verben 2,3 ist; bei den Adjektiven 1,9; bei den Substantiven 1,6. In der gleichen Reihenfolge findet man auch die Wortarten in Rogets Thesaurus des Englischen (s. Višnjakova, 1976): Verben 3,5, Adjektive 2,5, Substantive 2,1. Unter den Autosemantika hat das Adverb den kleinsten semantischen Umfang (im Englischen durschschnittlich 1.4 Bedeutungen). Im erklärenden Wörterbuch des Estnischen steht an der ersten Stelle ebenfalls das Verb (Tuldava, 1979). Dies unterstützt die Hypothese, daß "der Bedeutungsinhalt des Verbs größer ist als der des Substantivs" (Ufimceva, 1968: 89).

Man kann weiter feststellen, daß der semantische Umfang des Wortes und die Zahl der Wörter mit dem gegebenen semantischen Umfang im Vokabular in einer statistischen Relation stehen: Den größten Anteil im Vokabular haben Monosemantika, danach folgen die anderen Wörter in der Reihenfolge zunehmender Polysemie. Eine derartige Verteilung stellt offensichtlich eine universelle quantitativ-systemische Eigenschaft der Polysemie natürlicher Sprachen dar. Wie in vielen anderen Fällen, zeigt sich auch hier das bekannte Prinzip der Konzentration und Dispersion linguistischer Einheiten.

Um die Frage des analytischen Ausdrucks des Zusammenhangs zwischen der Zahl der Bedeutungen und dem Anteil der Wörter mit der gegebenen Bedeutungszahl zu klären, untersuchten wir Publikationen über Polysemie im Englischen, Ungarischen und Russischen.[6] Eine empirische Formel kann man auf verschiedene Weisen aufstellen (s. Tuldava 1979). Eine gute Übereinstimmung zwischen den empirischen und theoretischen Daten liefert die Potenzfunktion mit einem Korrekturparameter (vom Typ des Zipf-Mandelbrotschen Gesetzes), aber zwecks besserer Interpretation der Verteilung, besonders für die Klassifikation der Wörter mit unterschiedlichem semantischen Umfang, ist es zweckmäßiger, eine modifizierte Exponentialfunktion des Typs

$$(3.10) \qquad p(m) = ae^{-b\sqrt{m}}$$

zu benutzen, wobei $p(m)$ der Anteil der Wörter mit der gegebenen Zahl der Bedeutungen, m die Anzahl der Bedeutungen (semantischer Umfang), a und b die Parameter sind (e ist die Basis des natürlichen Logarithmus). Den Vergleich der erwarteten und der beobachteten Anteile der Wörter findet man in Tabelle 3.10.

In dieser Interpretation stellt die Wurzel aus der Zahl der Bedeutungen quasi eine neue Maßeinheit des semantischen Umfangs dar: Die Folge von natürlichen Zahlen 1 (=√1), 2 (=√4), 3 (=√9) usw. markiert die Intervalle, die man mit der natürlichen Gruppierung der Wörter in Unterklassen aufgrund des *Polysemiegrades* verbinden kann, beispielsweise:

der nullte Polysemiegrad: Wörter mit einer Bedeutung
der erste Polysemiegrad: Wörter mit 2-4 Bedeutungen
der zweite Polysemiegrad: Wörter mit 5-9 Bedeutungen
der dritte Polysemiegrad: Wörter mit 10-16 Bedeutungen
usw.

[6] Die Daten des Englischen stammen von Višnjakova (1976), wo die Polysemie der Autosemantika in Rogets Thesaurus (etwa 30000 Wörter) untersucht wurde. Die ungarischen Daten sind von Papp (1967) und schließen alle Wörter des ungarischen erklärenden Wörterbuchs ein (etwa 58000 Wörter). Das russiche Beispiel stammt von Krylov und Jakubovskaja (1977), wo nur die Verben (etwa 9500) aus Ožegovs "Wörterbuch der russischen Sprache" untersucht wurden.

Zur theoretischen Begründung einer derartigen Abhängigkeit kann das "energetische Prinzip" dienen, gemäß dem die Wahrscheinlichkeit der Realisierung des Phänomens proportional zum Exponenten seiner Komplexität abnimmt (analog zur Energie in der Thermodynamik, vgl. Šrejder, 1967). Die Komplexität interpretiert man in diesem Fall als die Komplexität der semantischen Wortstruktur, die man mit der Zahl der Bedeutungen mißt (genauer mit der Wurzel aus der Bedeutungszahl). Die Verteilung nach (3.10) kann man als einen Spezialfall einer allgemeineren Verteilung des Typs

$$(3.11) \qquad p(m) = ae^{-bm^c}$$

betrachten, wobei a, b und c Parameter sind. Unsere empirischen Daten sagen, daß der Parameter $c \approx 0{,}5$, wodurch Formel (3.11) in Formel (3.10) übergeht, da $m^{0,5} = \sqrt{m}$.

Tabelle 3.10
Beziehung zwischen der Anzahl der Bedeutungen m und den Anteilen von Wörtern mit m Bedeutungen $p(m)$ im englischen, ungarischen und russischen Wörterbuch. Theoretische Anteile nach Formel (3.10)

	Englisch		Ungarisch		Russisch (Verben)	
m	$p(m)$ beob.	$p(m)$ erw.	$p(m)$ beob.	$p(m)$ erw.	$p(m)$ beob.	$p(m)$ erw.
1	0,427	0,426	0,504	0,558	0,615	0,627
2	0,203	0,205	0,265	0,200	0,254	0,189
3	0,117	0,117	0,118	0,090	0,071	0,075
4	0,072	0,073	0,052	0,046	0,030	0,035
5	0,048	0,048	0,024	0,025	0,013	0,017
6	0,035	0,033	0,013	0,015	0,007	0,009
7	0,023	0,023	0,008	0,009	0,003	0,005
8	0,016	0,017	0,005	0,006	0,002	0,003
9	0,013	0,012	0,003	0,004	0,002	0,002
10	0,009	0,009	0,002	0,003	0,002	0,001
11	0,0073	0,0071	0,0014	0,0017		
12	0,0060	0,0055	0,0012	0,0012		
13	0,0053	0,0042	0,0009	0,0008		
14	0,0034	0,0033	0,0007	0,0006	0,001	(0,03)
15	0,0032	0,0026	0,0007	0,0004		
>15	0,014	(0,015)	0,002	(0,04)		
Gesamt	1,0	(1,0)	1,0	(1,0)	1,0	(1,0)
Parameter	$a = 2{,}5$	$b = 1{,}77$	$a = 6{,}8$	$b = 2{,}5$	$a = 11{,}4$	$b = 2{,}9$

Wir müssen ergänzen, daß es noch andere Behandlungen des Zusammenhangs zwischen dem semantischen Umfang und der Zahl der Wörter gibt. Krylov und

Jakubovskaja (1977) stellen beispielsweise die These von der optimalen Verteilung nach dem Prinzip des "maximalen semantischen Inhalts der Lexik" auf, was zu der Exponentialverteilung des Typs $p(m) = e^{-bm}$ ("Krylov-Gesetz") führt. Auch dies kann man als Speziallfall von (3.11) betrachten, wenn $a = 1$, $c = 1$ und im "idealen" Fall $e^{-b} = 0{,}5$, d.h. die Wahrscheinlichkeit, daß ein Wort m Bedeutungen hat, nimmt geometrisch mit dem Parameter $q = 0{,}5$ ab. Dabei muß die Anzahl von Monosemantika die Hälfte des Vokabulars ausmachen, die Bisemantika die Hälfte der Monosemantika usw.

Von besonderem Interesse sind Untersuchungen, in denen die Polysemie mit anderen strukturellen Eigenschaften der Wörter in Verbindung gebracht wird, z.B. mit der Wortlänge (Fickermann et al., 1984; Sambor, 1984; Köhler, 1986), in denen festgestellt wurde, daß die Abhängigkeit zwischen Polysemie und Wortlänge dem Potenzgesetz folgt (das die erwähnten Autoren als Spezialfall des Menzerathschen Gesetzes betrachten). Eine starke Korrelation wurde auch zwischen der Polysemie und der morphologischen Aktivität des Wortes entdeckt (vgl. z.B. Tichonov, 1983).

Der Zusammenhang mit der Worthäufigkeit

Es ist bekannt, daß jedes Wort (auch ein mehrdeutiges) im Text in der Regel nur in einer bestimmten, "aktualisierten" Bedeutung verwendet wird. Ungeachtet der großen Anzahl möglicher Nuancen der Wortbedeutung in konkreten Kontexten kann man eine bestimmte Anzahl gleichartiger Verwendungen identifizieren, die man im erklärenden Wörterbuch als separate, usuale Bedeutungen fixiert. Dadurch wird die Untersuchung des direkten Zusammenhangs zwischen dem *semantischen Umfang* des Wortes im gegebenen Lexikon und seiner *Verwendung* im Text gerechtfertigt. Was den Charakter dieses Zusammenhangs betrifft, so kann man rein spekulativ schließen, daß diese zwei quantitativen Merkmale des Wortes in direkt proportionalem Zusammenhang stehen: Je mehr Bedeutungen das Wort hat, desto häufiger wird es im Text (in der Rede) verwendet. Wenn alle einzelnen Bedeutungen des Wortes das gleiche "Gewicht" hätten und gleich häufig wären, dann könnte man den Zusammenhang zwischen Bedeutungszahl und Worthäufigkeit eindeutig bestimmen: Ein Wort mit zwei Bedeutungen würde man im Text doppelt so häufig benutzen wie ein monosemantisches Wort, ein Wort mit drei Bedeutungen dreimal so oft wie ein monosemantisches Wort usw. Bekanntlich bildet die Wort-interne semantische Abgrenzung selbst ein kompliziertes Netz, insbesondere in Form der Hierarchie der Bedeutungen (unter den Bedeutungen eines polysemen Wortes gibt es üblicherweise ein Wort mit einer "Grundbedeutung"). Auch die aktualisierten Bedeutungen unterschiedlicher Wörter können sich in ihrer Verwendungshäufigkeit unterscheiden, und zwar in Abhängigkeit von kommunikativen Bedürfnissen und

von Bedürfnissen der Sprachstruktur. Außerdem muß man der stetigen Veränderung und der Entwicklung des lexikalisch-semantischen Systems der Sprache in Verbindung mit der gesellschaftlichen Entwicklung Rechnung tragen. Die Beziehung zwischen dem semantischen Umfang des Wortes und seiner Verwendung im Text hat daher einen komplexen Charakter, und beim gegenwärtigen Stand der Forschung können wir lediglich versuchen, den Charakter dieses Zusammenhang in einem integrativen Modell zu beleuchten. Wir gehen dabei von der Voraussetzung aus, daß das System, ungeachtet allerlei lokaler Abweichungen, als Ganzes während seiner gesamten Existenz eine relative Stabilität aufrechterhält.

Es interessiert uns vor allen Dingen die Etablierung der Form des Zusammenhangs zwischen dem semantischen Umfang, den wir als die Anzahl der Bedeutungen des Wortes im Lexikon messen, und der Verwendungshäufigkeit des Wortes im Text. Es ist daher zweckmäßig, sich die Zahl der Bedeutungen des Wortes als eine Funktion der mittleren Verwendungshäufigkeit vorzustellen (obwohl die Abhängigkeit tatsächlich gegenseitig ist). Als Beispiel nehmen wir russische (aus dem Material von Polikarpov, 1976) und estnische (Tuldava, 1979) Daten. Gruppiert man die Wörter nach Häufigkeitszonen um den Zonenmittelwert (F) und die mittlere Zahl von Wortbedeutungen in der gegebenen Zone (m), dann kann man die Beziehung zwischen diesen Variablen mit Hilfe der Potenzfunktion

$$m = \alpha F^{\gamma} \tag{3.12}$$

erfassen, wobei α und γ Parameter sind. Für das Häufigkeitswörterbuch des Russischen (Zasorina 1977) werden die Häufigkeitszonen durch Intervalle der Ränge bestimmt, wobei die mittlere Häufigkeit in Intervallen von jeweils 100 Wörtern berechnet wird. Die estnischen Daten werden etwas anders dargestellt: Als Grundlage dient die Intervallmitte der Häufigkeiten im Häufigkeitswörterbuch der Lexeme estnischer künstlerischer Prosa, z.B. 15 für das Intervall 10...20, und in diesem Intervall wird die mittlere Zahl der Wortbedeutungen festgelegt. Die unterschiedliche Zählung ändert aber nichts an den Tatsachen, und in beiden Fällen findet man eine gute Übereinstimmung zwischen empirischen (beobachteten) und theoretischen (erwarteten) Daten innerhalb der Grenzen der Erfahrungswerte; außerdem liefert die Funktion akzeptable Resultate auch bei Extrapolation im Bereich kleinerer Häufigkeiten. Es ist zu bemerken, daß es bei den empirischen Daten im Bereich der großen Häufigkeiten eine Verlangsamung des Anwachsens der Bedeutungszahl gibt (dieses Phänomen wurde bereits in der Untersuchung von Andrukovič und Korolev (1977) beobachtet). Dies bezieht sich auf die kleine Zahl der häufigsten Wörter, und bei der Verwendung mittlerer Werte glättet die gegebene Funktion einfach die Irregularität in der Zone der großen Häufigkeiten.

Wie bekannt, wurde die Abhängigkeit des semantischen Umfangs (Zahl der Bedeutungen) von der Verwendungshäufigkeit des Wortes zum ersten Mal von

Zipf (1949) formuliert und zwar in der Form $m = \sqrt{F} = F^{0,5}$ (m - Zahl der Bedeutungen, F - Verwendungshäufigkeit). Im Lichte der vorliegenden Untersuchung kann man sagen, daß Zipfs Formel in dem Sinne korrekt ist, daß sie für den Zusammenhang zwischen m und F die Potenzfunktion $m = \alpha F^{\gamma}$ ansetzt, wobei die Konstante α gleich 1 ist (inhaltlich interpretiert ist die Anzahl der Bedeutungen bei hapax legomena gleich 1). Oben haben wir gesehen, daß die Konstante α bei empirischen Daten tatsächlich nahe bei 1 liegt und man sie in einigen Fällen vernachlässigen kann. Im allgemeinen stellt Zipfs Formel aber einen Spezialfall der Funktion $m = F^{\gamma}$ dar, wenn $\gamma = 0,5$ ist. In Wirklichkeit variiert der Wert von γ in Abhängigkeit vom Vokabularumfang und vom Sprachtyp.

Tabelle 3.11
Abhängigkeit zwischen der mittleren Zahl der Wortbedeutungen im Lexikon (m) und der mittleren Häufigkeit der Wortverwendung im Text (F). Berechnung nach Formel (3.12)

Russisch				Estnisch			
Intervall der Ränge	F	m beob.	m erw.	Intervall der Häuf.	F	m beob.	m erw.
1-100	4369	8,2	8,2	10-20	15	2,9	2,9
101-200	758	5,4	5,5	21-30	25	3,2	3,4
201-300	468	4,9	4,9	31-40	35	4,4	3,9
301-400	374	4,4	4,7	41-50	45	3,7	4,2
401-500	250	4,6	4,3	51-60	55	4,8	4,6
501-600	234	4,3	4,2	61-70	65	5,1	4,9
601-700	200	3,9	4,0	71-80	75	5,3	5,2
701-800	176	4,4	3,9	81-90	85	5,0	5,4
801-900	155	3,3	3,8	91-100	95	6,5	5,6
901-1000	140	3,7	3,7	101-150	125	6,3	6,3
Prognose	100	-	3,5	Prognose	10	-	2,5
	50	-	2,9		5	-	1,9
	10	-	2,0		2	-	1,4
	1	-	1,2		1	-	1,05
Parameter	$\alpha = 1,2$; $\gamma = 0,23$			Parameter	$\alpha = 1,05$; $\gamma = 0,38$		

Die Berechnung mit Hilfe von $m = \alpha F^{\gamma}$ liefert die Regressionslinie des Zusammenhangs zwischen Anzahl der Bedeutungen (m) und Worthäufigkeit (F). Diese Formel bestimmt aber die Abhängigkeit zwischen m und F nur im Durchschnitt und

gibt keine exakte Voraussage für jedes einzelne Wort separat. Dies spiegelt eine der wichtigsten Besonderheiten probabilistischer Systeme wieder, daß nämlich die Organisation (die Struktur) den Zustand einzelner Elemente bei vielen Ereignissen nur allgemein, ganzheitlich wiedergibt. Außerdem muß man berücksichtigen, daß die Regressionsgerade, die man aufgrund von Stichprobendaten berechnet hat, Konfidenzgrenzen (auf dem gegebenen Signifikanzniveau) hat. Es ist am günstigsten, die Konfidenzintervalle aufgrund der Mittelwerte von m und F zu berechnen und sie dann in einem bilogarithmischen Koordinatensystem (bei linearem Zusammenhang zwischen den Logarithmen von m und F) darzustellen. Wenn man die Konfidenzintervalle bei dreifacher Standardabweichung bestimmt, dann kann man ungefähr annehmen, daß die Regressiongerade mit 99% Wahrscheinlichkeit in diesem Intervall liegt. Wichtig ist hier, daß wir aufgrund der Angaben über Konfidenzintervalle der Regressionsgeraden die Möglichkeit erhalten, einige praktische Aufgaben der Lexikographie mit hinreichend großer statistischer Zuverlässigkeit zu lösen.

Für Autosemantika des Estnischen (bei $m \approx F^{0,38}$) wurde der *Konfidenzbereich* für die Regressionsgerade im bilogarithmischen Maßstab errichtet (vgl. Abb. 3.4 und Tabelle 3.11). (Zur Methode der Berechnung der Konfidenzintervalle s. z.B. Förster, Rönz 1979).

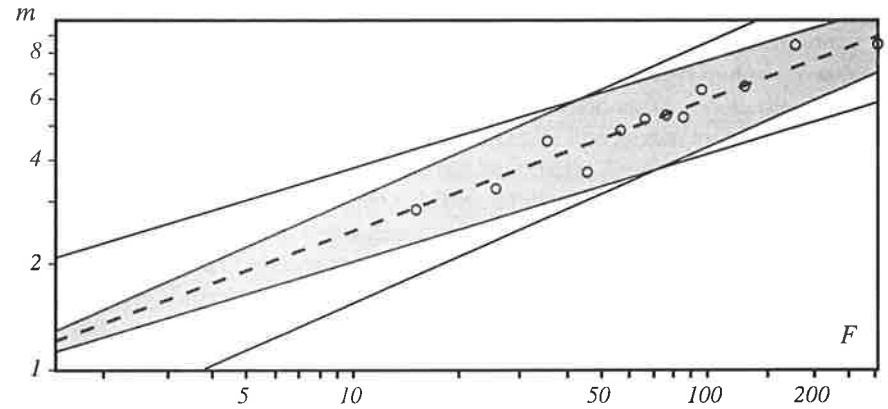


Abbildung 3.4. Konfidenzbereich der Regressionsgeraden für den Zusammenhang zwischen der Anzahl der Bedeutungen (m) und der Worthäufigkeit im Text (F) bei estnischen Autosemantika in bilogarithmischer Darstellung.

Man kann sich darauf einigen, daß der Konfidenzbereich die Zone der "Norm" angibt, in die die polysemen Wörter bei zulässiger Schwankung der Bedeutungszahl in Abhängigkeit von ihrer Verwendungshäufigkeit fallen. In Fällen, in denen

einzelne Wörter außerhalb der Normgrenzen liegen (weil sie zu viele oder zu wenige Bedeutungen haben), kann man eine "Über-" oder "Unter-Polysemie" konstatieren. Im Estnischen beispielsweise sind "überpolysemische" Wörter einige abstrakte Substantive (*elu* 'Leben', *meel* 'Gefühl'), Bezeichnungen aus der Tierwelt (*pesa* 'Nest', *sarv* 'Horn'), Bezeichnungen von Menschen *laps* 'Kind' u.a. Ein Teil dieser Wörter wird im übertragenen Sinn gebraucht, wodurch ihre Polysemie anwächst. "Unterpolysemische" Wörter sind einige hochfrequente Substantive, die einen relativ schmalen semantischen Umfang haben, z.B. *inimene* 'Mensch', *nägu* 'Gesicht', *päev* 'Tag', *aasta* 'Jahr', *kevad* 'Frühling'. Insektion und qualitative Analyse der "über-" und "unterpolysemischen" Wörter erlauben es, nicht nur einige Eigenschaften und Besonderheiten des semantischen Aspekts der Lexik aufzudecken, sondern bieten auch die Möglichkeit, die Qualität einer lexikographischen Arbeit zu beurteilen.

4. Soziale und stilistische Aspekte der Untersuchung

In diesem Kapitel werden Fragen der sozialen und funktionalen Differenzierung der Lexik und einige Probleme des Wachstums und der Entwicklung der Lexik von der quantitativen Seite erörtert. Im Abschnitt über Stilanalyse wird der Messung lexikalischer Differenzierung ("Vokabularreichtum") und der lexikalischen Nähe von Texten besondere Aufmerksamkeit gewidmet.

4.1. Soziale Differenzierung der Lexik

Verwendungsbereiche der Lexik

Die Lexik der Sprache ist nach verschiedenen Merkmalen differenziert, darunter nach Verwendungsbereichen, die eine außersprachliche, soziale Begründung haben. Die Lexik heutiger Nationalsprachen zerfällt in der Regel in drei Grundgruppen oder Varietäten: Die allgemein gebräuchliche (nationale) Lexik, die dialektale Lexik und die spezielle oder terminologische Lexik. Die letzten zwei Gruppen kann man unter den Sammelbegriff "Lexik mit begrenzter Anwendung" fassen. Jargon und Argot, die zu den Soziolekten gehören, kann man (im weitesten Sinne) als besondere Untergruppen der dialektalen Lexik betrachten. Gleichzeitig muß man auch die sogenannten Professionalismen (halboffizielle, hauptsächlich umgangssprachliche Varietäten von Termini) als Untergruppe der speziellen Lexik unterscheiden.

Die grundlegenden Varietäten der Lexik sind reale Gegebenheiten. Da sie aber Subsysteme des lexikalischen Supersystems der gegebenen Sprache sind, haben sie alle Charakteristika des Subsystems, d.h. sie sind systemisch verknüpft, haben Überschneidungen usw. Zur Illustration kann man das gesamte lexikalische System in Form eines *Venn*-Diagramms darstellen (vgl. Abb. 4.1).

Zu der allgemein gebräuchlichen Lexik gehören Wörter, deren Verwendung und Verständnis "durch Lokalität, Beruf oder Tätigkeit nicht begrenzt ist" (Kalinin 1978: 119). Diese Lexik bildet die beständige Grundlage der Sprache, auch wenn sie nicht homogen ist. Man kann in ihr Schichten der mündlichen Umgangssprache (darunter auch einfache Volkssprache) und der Schriftsprache (darunter "literarische" Lexik) unterscheiden. Von der historischen Perspektive aus kann man in ihr verschiedene genetische Schichten, Neologismen und Archaismen finden. Mit Hilfe

statistischer Kriterien kann man die Kern- oder Basislexik und die periphere Lexik unterscheiden. In allen Varietäten bleibt das Grundmerkmal der allgemein gebräuchlichen Lexik - ihre Verwendung - unabhängig von Lokalität und Tätigkeitsart.

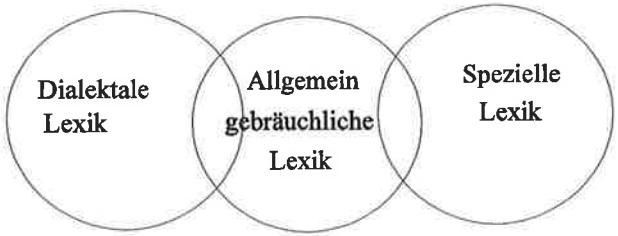


Abbildung 4.1. Die Grundkomponenten der Lexik einer modernen Nationalsprache (aufgeteilt nach Verwendungsbereichen)

Es gibt keine scharfe Demarkationslinie zwischen der allgemein gebräuchlichen Lexik und den anderen Varietäten. Die allgemein gebräuchliche Lexik reichert sich ständig durch Übernahme von Dialektwörtern an, die durch Verschmelzung ländlicher und städtischer Gebiete, durch künstlerische Literatur und durch bewußte Übernahme von Dialektismen durch Linguisten zum Gemeingut werden. Gleichzeitig findet auch der umgekehrte Prozeß statt, nämlich der Einfluß der allgemein gebräuchlichen Lexik auf die Dialekte, wodurch sie sich der allgemeinen Nationalsprache annähern. Allgemein verständlich und gebräuchlich werden mit der Zeit auch viele Wörter der sozialen Dialekte (Jargon und Argot), die man in mündlicher oder schriftlicher Sprache wegen ihrer Expressivität benutzt.

Es gibt auch keine scharfe Grenze zwischen der allgemein gebräuchlichen und der speziellen Lexik. In unserer Zeit kommen Termini immer öfter in allgemeinen Gebrauch. Infolgedessen vergößert sich in vielen Sprachen die Schicht allgemein verständlicher Termini. Dies ist eine der Konsequenzen der weit verbreiteten "Intellektualisierung" heutiger Sprachen. Auf der anderen Seite bekommen viele allgemein gebräuchlichen Wörter zusätzlich zu ihren allgemeinen Bedeutungen noch spezielle Bedeutungen (z.B. 'Base' in der Chemie), ganz zu schweigen davon, daß viele Termini mit Hilfe der allgemein gebräuchlichen Lexik gebildet werden (z.B. 'Wurzel' in der Linguistik).

Daher muß man feststellen, daß Wörter aus unterschiedlichen Verwendungsbereichen keine festen, unbeweglichen Grenzen haben; man beobachtet graduelle, "stetige" Übergänge zwischen den lexikalischen Schichten. Den Umfang und die Grenzen lexikalischer Gruppen verschiedener Verwendungsbereiche kann man objektiv mit Hilfe von probabilistisch-statistischen Kriterien (Häufigkeit, Verwendbarkeit, Besonderheiten der Häufigkeitsverteilung usw.) bestimmen. Bekannt sind beispielsweise die Abgrenzungskriterien für allgemein gebräuchliche und spezielle Lexik aufgrund differenzierter Angaben über den Grad der Verwendbarkeit der Wörter in einzelnen Subsprachen (Andreev, 1967) und die Kriterien der Bestimmung der terminologischen Lexik aufgrund der Worthäufigkeit (Bektaev, 1978). Von Interesse sind auch empirische Beobachtungen in Arbeiten, die sich der Analyse spezieller Texte widmen, z.B. die Feststellung, daß in einem Vokabular, das aus einem speziellen Text von 100000 laufenden Wörtern zusammengestellt wurde, sich die sehr speziellen Wörter, die sich auf die Hauptthematik des Bereichs beziehen, in der Zone der mittleren Häufigkeit $50 > F > 15$ konzentrieren, wo F die Häufigkeit im Text bedeutet (Neguljaev et al., 1973).

Auf dem Hintergrund der lexikalischen Basisgruppen, die sich infolge der Gliederung des allgemeinen Systems der Lexik nach Verwendungssphären bilden, zeichnet sich besonders die Lexik der Literatursprache, oder die "literarische Lexik", als eine Zwischenschicht aus und gleichzeitig als die wichtigste Komponente der lexikalischen Systeme moderner Sprachen. Im allgemeinen nimmt man an, daß die literarische Lexik sich mit der allgemein gebräuchlichen Lexik im oben erwähnten Sinne nicht vollständig deckt. Sie unterschiedet sich von der allgemein gebräuchlichen Lexik durch recht strenge Normierung (die Normen werden durch entsprechende Regeln und Wörterbücher der Literatursprache festgesetzt). Zum Bestand der literarischen Lexik gehören keine salopp-umgangssprachlichen Wörter, auch wenn man sie in der künstlerischen Literatur z.B. aus stilistischen Gründen benutzten kann. Weiter, zum Bestand der literarischen Lexik gehören keine dialektalen oder Jargonwörter, auch wenn man auch diese in literarischen Texten als "nichtliterarische" Einsprengsel finden kann. Schließlich, die literarische Lexik schließt einige Schichten der speziellen Lexik aus, besonders Wörter des professionellen Jargons. Die Frage der Einbeziehung der engspezialisierten Terminologie in den Bestand der literarischen Lexik löst man auf unterschiedliche Weisen. Es ist offensichtlich möglich, von einer literarischen Lexik im weiten und im engen Sinne des Wortes zu sprechen. Unter den letzteren Fall fällt die "allgemein gebräuchliche" literarische Lexik, die nur einen Teil der speziellen Lexik einschließt, die in orthographischen (orthologischen) oder erklärenden Wörterbüchern fixiert ist.

Allgemeine und spezifische Lexik

Es ist bekannt, daß die literarische Sprache kein homogenes Ganzes ist, sondern in viele Genre und Stile aufgeteilt ist. So unterscheidet man in der Literatursprache sogenannte Funktionalstile aufgrund der gesellschaftlichen Funktionen der Sprache (Mitteilung, Verständigung, Einflußnahme) unter Einbeziehung der Verständigungssphäre (ästhetische, wissenschaftliche, amtliche usw.). Wenn man den Ge-

genstandsbereich hervorhebt, dann spricht man von "Fachsprachen". Manchmal spricht man etwas unbestimmt von "Genre-Gruppen" der Texte oder von Texten unterschiedlicher Sphären der funktionalen Rede usw. In allen diesen Fällen leitet man die Prinzipien der Bildung von einzelnen Textgruppen aus außersprachlichen, extralinguistischen Grundlagen ab. Obwohl sie aus außersprachlichen Grundlagen hervorgegangen und mit dem Gegenstand der Äußerung eng verbunden sind, unterscheiden sich solche Gruppen (und Untergruppen) von Texten untereinander auch durch innersprachliche Merkmale, vor allem durch die Besonderheiten der Auswahl und Verwendung sprachlicher Mittel. Was die Lexik dieser Texte betrifft, so sind zwei Momente offensichtlich. Erstens, den grundlegenden Teil des Vokabulars eines jeden Textes bildet die in stilistischer Hinsicht neutrale allgemeingebräuchliche Lexik. Zweitens, für jede Textgruppe gibt es eine spezifische Lexik, die nur für die betreffende Gruppe charakteristisch ist. Bei vergleichender Untersuchung der Texte unterschiedlicher Gruppen zeigt sich die spezielle Kernschicht der allgemein gebräuchlichen Lexik, die allen Texten sogar bei kleinen Umfängen buchstäblich gemeinsam ist. Außerdem zeigen sich auch für eine bestimmte Menge der Texte (Gruppen von Texten) gemeinsame Schichten der Lexik usw. bis zu der spezifischen Lexik einer engen Gruppe von Texten, sogar bis zu einem individuellen Text (Kožina, 1977). Bei diesem Ansatz kann man von einer Opposition zwischen der "allgemeinen" und der "spezifischen" Lexik sprechen, jedoch auf unterschiedlichen Ebenen der Allgemeinheit/Spezifizität, deren Wahl von den Zielen und Aufgaben der Forschung abhängt.

Beim quantitativen Ansatz muß man in allen Fällen der Opposition zwischen allgemeiner und spezifischer Lexik die Bereiche Vokabular und Text unterscheiden, d.h. zwischen dem Inventar und der Verwendung differenzieren. Illustrieren wir dies an einem Beispiel. Das Häufigkeitswörterbuch des Russischen (Zasorina 1977) wurde aufgrund von Texten aus vier Funktionalstilen der Rede (künstlerische Prosa, Dramen, Zeitungen und Zeitschriften, wissenschaftliche Literatur) von gleichem Umfang zusammengestellt (je 250 Tausend laufender Wörter). Den gemeinsamen Teil der Lexik, d.h. Wörter, die in allen vier Textsorten vorkommen, bilden 6440 Einheiten (Lexeme) oder 16,4% des Gesamtvokabulars von 39268 Einheiten. Aber diese 6440 Wörter decken 82,2% aller Texte ab, auf denen das Häufigkeistwörterbuch beruht (868577 laufende Wörter von der Gesamtzahl von 1056382). Daher besteht ein großer Unterschied zwischen den quantitativen Charakteristika des Vokabulars und des Texts. Dieser Unterschied begründet sich dadurch, daß ein kleiner Teil allgemein gebräuchlicher Wörter - im strengen Sinne die "gemeinsame" oder "Kern-"Lexik - häufiger ist und in quantitativer Hinsicht in der aktuellen Rede (in Texten) aller Kommunikationssphären dominiert.

Es darf aber nicht vergessen werden, daß die "allgemeine" Lexik sich in verschiedenen Kommunikationssphären unterschiedlich verhält; d.h. in unterschiedlichen Texten gibt es Unterschiede zwischen den Häufigkeitscharakteristika allgemein gebräuchlicher Wörter. Dieser Umstand erlaubt es (unter Berücksichtigung der Quantität und Struktur einer spezifischen Lexik), Stile und Substile aufgrund objektiver probabilistisch-statistischer Kriterien zu differenzieren. So sind nach den Angaben des erwähnten Häufigkeitswörterbuches einige Auto- und Synsemantika in unterschiedlichen Texten folgendermaßen verteilt:

	Künstlerische Prosa	Dramaturgie	Zeitschriften Zeitungen	Wissenschaft Technik
a (Konj.)	2909	5203	1355	1252
kotoryj (Pron.)	673	314	1381	1600
god 'Jahr'	246	286	1080	555
bol'šoj 'groß'	297	708	571	490
skazat' 'sagen'	1278	978	359	294

Große Bedeutung hat zweifellos die Tatsache, daß die Einheiten der allgemein gebräuchlichen Lexik in unterschiedlichen Funktionalstilen weder nach ihren qualitativen (Bedeutung, semantischer Umfang, Expressivität, usw.) noch nach ihren quantitativen Eigenschaften völlig deckungsgleich sein können.

Der Unterschied der Worthäufigkeiten führt auch zu Unterschieden in Verteilungen, in der Abdeckung des Textes und in anderen quantitativ-typologischen Textcharakteristika, die zu unterschiedlichen Kommunikationssphären und dadurch zu unterschiedlichen Funktionalstilen (oder Subsprachen) gehören. Wenn wir beispielsweise die Listen der häufigsten Substantive in unterschiedlichen Funktionalstilen und Sprachen vergleichen (vgl. Tabelle 4.1), so entdecken wir folgendes:

Unterschiedliche Stile weisen große Unterschiede auf, so findet sich unter den zehn häufigsten Substantiven in künstlerischer Prosa und in Zeitungstexten im Estnischen nur ein gemeinsames Wort (aeg 'Zeit'); im Russischen ist es auch nur ein Wort (den' 'Tag'). Gleichzeitig kann man innerhalb eines Stils eine Ähnlichkeit des Vorkommens hochfrequenter Substantive feststellen, besonders in künstlerischer Prosa unterschiedlicher Sprachen. Beispielsweise haben Estnisch und Finnisch von zehn Wörtern sieben gemeinsam (im gegebenen Fall auch genetisch verwandte), nämlich (in Übersetzung) *Mann, Mensch, Hand, Kopf, Zeit, Tag, Werk (Ding, Sache)*. Estnisch und Russisch haben 8 gemeinsame Wörter, nämlich *Mensch, Hand, Auge, Kopf, Gesicht, Zeit, Tag, Werk (Ding, Sache)*. Auffällig ist, daß unter den gemeinsam hochfrequenten Wörtern der künstlerischen Prosa unterschiedlicher Sprachen Wörter, die den Menschen (*Mensch, Mann, Frau)* und Körperteile (*Hand, Auge, Kopf, Gesicht)* bezeichnen, prominent vorkommen. In Zeitungstexten verschiedener Sprachen gibt es weniger Übereinstimmungen. Beispielsweise gibt es im Estnischen und Finnischen unter den zehn häufigsten substantivischen Begriffen nur drei gemeinsame (*Jahr, Zeit, Teil*), und im Estnischen

Tabelle 4.1
Die zehn häufigsten Substantive in unterschiedlichen Funktionalstilen des Estnischen, Finnischen und Russischen

Estnisch				Finnisch				Russisch			
Künstl. Prosa		Zeitungen		Künstl. Prosa		Zeitungen		Künstl. Prosa		Zeitungen	
mees	Mann	aasta	Jahr	aika	Zeit	vuosi	Jahr	čelovek	Mensch	pravitel'stvo	Regierung
aeg	Zeit	valitsus	Regierung	mies	Mann	aika	Zeit	ruka	Hand	strana	Land
silm	Auge	riik	Staat	päivä	Tag	osa	Teil	glaz	Auge	god	Jahr
inimene	Mensch	rahvas	Volk	käsi	Hand	maa	Erde; Land	delo	Werk	partija	Partei
käsi	Hand	partei	Partei	poika	Sohn; Junge	asia	Werk;Ding	žizn'	Leben	bor'ba	Kampf
naine	Frau	politika	Politik	ihminen	Mensch	ihminen	Mensch	golova	Kopf	den'	Tag
päev	Tag	suhe	Beziehung	lapsi	Kind	työ	Arbeit	lico	Gesicht	gazeta	Zeitung
asi	Sache;Ding	aeg	Zeit	asia	Werk;Ding	kysymys	Frage	den'	Tag	gosudarstvo	Staat
pea	Kopf	osa	Teil	maa	Erde; Land	päivä	Tag	mat'	Mutter	sila	Kraft
nägu	Gesicht	õigus	Recht	pää	Kopf	määrä	Grad	vremja	Zeit	vopros	Frage

und Russischen vier (*Jahr, Regierung, Partei, Staat)*. Hier zeigt sich der Unterschied in der Thematik, teilweise bedingt durch historische und sozial-politische Umstände und auch dadurch, daß die Texte zu unterschiedlichen Untertypen des publizistischen Stils gehören (im Estnischen wurden außenpolitische, im Finnischen und Russischen gemischte Zeitungstexte untersucht).

"Markierte" Lexik

Eine Vorstellung von der quantitativen Verteilung der stilistischen Schichten der Wörter in der Lexik moderner Sprachen kann man sich durch die Analyse markierter Wörter in großen erklärenden oder normativen Wörterbüchern verschaffen. Im siebten Band des "Wörterbuches der zeitgenössischenen russischen Literatursprache" (1948-1965) sind aus der Gesamtzahl von 15530 Eintragungen 3925, d.h. 25,3%, mit Markierungen versehen (Filin, 1973). In Ožegovs "Wörterbuch der Russischen Sprache" (1963) von 51533 Wörtern wurden 17003 Mal Markierungen benutzt (Denisov, Kostomarov, 1970). In diesem Wörterbuch gibt es etwa 83000 Eintragungen, d.h. die markierten Einheiten machen etwa 20% aus. Im orthologischen Wörterbuch der estnischen Sprache (Õigekeelsussõnaraamat 1976), das 115000 Lexeme enthält, gibt es 42083 Markierungen, d.h. 33% des gesamten Wörterbuchs. Im Probeheft des erklärenden Wörterbuches des Estnischen (Eesti kirjakeele sõnaraamat 1969) gibt es für 1610 Lexeme 2211 Eintragungen, von denen 686 markiert sind, d.h. 29,2%. Der Anteil markierter Einheiten in den untersuchten normativen Wörterbüchern der Literatursprache bewegt sich daher zwischen 20 und 33%. Die Analyse der Verteilung von markierten Einheiten nach Typen der stilistischen Charakteristika zeigte Unterschiede in der Struktur der Wörterbücher (vgl. Tabelle 4.2). Beispielsweise dominiert in estnischen Wörterbüchern unter den markierten Wörtern spezielle Lexik, während in russischen Wörterbüchern vor allem umgangsprachliche Lexik gekennzeichet wird. Im Estnischen gibt es keinen großen Unterschied zwischen der umgangssprachlichen und der allgemein gebräuchlichen Literatursprache, und daher ist der Anteil markierter umgangssprachlicher Wörter vernachlässigbar klein.

Die Verteilung der Markierungen nach Häufigkeit im orthologischen Wörterbuch des Estnischen findet man in Tabelle 4.3. Der Unterschied in der Häufigkeit der Markierungen zeigt den Unterschied in der Relevanz einzelner Wortgruppen, die die markierte Schicht im System der literarischen Lexik ausmachen. Auffällig ist die kontinuierliche, monotone Abnahme der Häufigkeiten der Markierungen, von denen die meisten zur speziellen Lexik gehören. Diese Regularität der Ranghäufigkeitsverteilung erinnert an die latente Wirkung des Gesetzes der Konzentration und Streuung, das für komplexe selbstregulierende Systeme charakteristisch ist (vgl. Abschnitt 2.2). Wenn man - unter Beachtung des Homogenitäts-

Tabelle 4.2
Verteilung der stilistisch markierten lexikalischen Einheiten im Russischen und Estnischen (in %)

Wörterbuch Lexik	W. der zeitg. russ. Literaturspr.	Russ. W. von Ožegov	Ortholog. W. des Estnischen	Erklärendes W. des Estnischen
Speziell	7	17,0	92,7	72,8
Dialektal	3,7	1,8	1,5	2,9
Umgangsprachlich	38,4	33,9	1,9	6,8
Salopp	24,6	9,3		
Archaismen	?	13,5	1,9	6,2
Andere	?	24,5	2,0	11,3
Gesamt (%)		100,0	100,0	100,0
Anzahl von Markierungen	3925 (Stichpr.)	17003	42083	646 (Stichpr.)

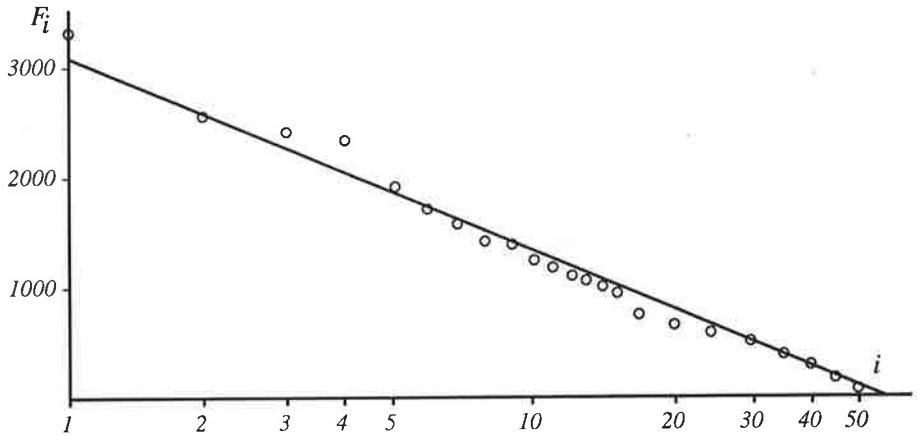


Abbildung 4.2. Ranghäufigkeitsverteilung der speziellen Lexik im Wörterbuch. Abszisse logarithmisch

prinzips - nur die Markierungen der speziellen Lexik betrachtet (52 Markierungen von 61), dann stellt sich heraus, daß die empirische Rangverteilung der Häufigkeiten dem logarithmischen Gesetz folgt (vgl. Abb. 4.2), auch wenn man im "Kern", d.h. unter den häufigsten Markierungen, einige Abweichungen von der allgemeinen

Tabelle 4.3
Verteilung stilistischer Markierungen im Orthologischen Wörterbuch des Estnischen. Wörterbuchumfang: 115000 Wörter; Zahl der Markierungen: 61; Zahl der Vorkommen von Markierungen: 42083

3433 tehn.	(Technik)	484 farm.	(Pharmazie)
2585 bot.	(Botanik)	435 kunst	(bildende Kunst)
2475 zool.	(Zoologie)	423 mets.	(Forstwissenschaft)
2437 med.	(Medizin)	413 vet.	(Veterinärmedizin)
1991 pôll.	(Landwirtschaft)	406 (etn.)	(Ethnographie)
1701 maj.	(Ökonomie)	383 kirj.	(Literatur)
1589 sport	(Sport)	372 trük.	(Typographie)
1465 keem.	(Chemie)	353 füsiol.	(Physiologie)
1442 aj.	(Geschichte)	343 min.	(Mineralogie)
1269 el.	(Elektrizitätswesen)	308 kirikl.	(Kirchlich)
1191 ehit.	(Bauwesen)	286 astr.	(Astronomie)
1117 anat.	(Anatomie)	272 folkl.	(Folklore)
1093 sôj.	(Militärwesen)	262 fot.	(Fotographie)
1924 lgv.	(Linguistik)	227 mäend.	(Bergbau)
990 füüs.	(Physik)	194 teatr.	(Theater)
805 kônek.	(Umgangsprache)	194 meteor.	(Meteorologie)
799 van.	(veraltet)	182 kal.	(Fischerei)
744 mat.	(Mathematik)	173 filos.	(Philosophie)
672 piltl.	(bildlich)	134 ped.	(Pädagogik)
659 biol.	(Biologie)	132 arheol.	(Archeologie)
634 geol.	(Geologie)	129 psühh.	(Psychologie)
625 tekst.	(Textilwesen)	122 bibl.	(Bibliographie)
623 muus.	(Musik)	92 loog.	(Logik)
611 murd.	(dialektal)	74 paleont.	(Paleontologie)
607 aiand.	(Gartenbau)	63 müt.	(Mythologie)
605 geogr.	(Geographie)	57 vulg.	(vulgär)
595 pol.	(Politik)	49 antr.	(Anthropologie)
575 jur.	(Jurisprudenz)	49 lastek.	(Kindersprache)
546 kok.	(kulinarisch)	36 nalj.	(scherzhaft)
507 mer.	(Ozeanologie)	21 luulek.	(poetisch)
		8 halv.	(verächtlich)

Tendenz beobachten kann[1]. Diesen Verteilungstyp findet man auch in einigen

[1] Lineare Abhängigkeit zwischen Häufigkeit (F) und dem Logarithmus der Ranges ($ln\ i$) nach der Formel $F_i = a - b\ ln\ i$, wobei a und b Konstanten sind. Im gegebenen Fall ist $a \approx 3200$ (theoretische maximale Häufigkeit), $b \approx 800$ (zeigt das Tempo der Abnahme).

anderen Bereichen der Linguistik, beispielsweise bei der Rangverteilung von Buchstaben im Text.

Die Gesamtheit der speziellen Lexik (Termini) im Orthologischen Wörterbuch kann man nach drei grundlegenden Bereichen der Wissenschaft gruppieren (vgl. Tuldava 1983):

	Zahl der Markierungen	Häufigkeit	%
Ingenieurwissenschaften	18	17651	45,2
Naturwissenschaften	14	12497	32,0
Humanwissenschaften	20	8877	22,8
Gesamt	52	39025	100,0

Die Verteilung einzelner Sparten der speziellen Lexik im Wörterbuch spiegelt zweifellos das gesellschaftliche Gewicht und die Aktualität der gegebenen Wissenschaftszweige in unserer Zeit wider[2]. Wir müssen jedoch anmerken, daß die spezielle Lexik der Humanwissenschaften eine größere Verbreitung hat, als es aufgrund der Tabelle ersichtlich ist. Viele Termini trifft man täglich in Presse, Radiosendungen, Fernsehen usw. an, wodurch eine große Menge dieser Termini in den Bereich der allgemein gebräuchlichen Wörter übergeht und im Wörterbuch keine spezielle Kennzeichnung bekommt. Hier zeigt sich der Zusammenhang zwischen Häufigkeit und Bekanntheit der Wörter.

4.2. Das Wachstum und Entwicklung der Lexik

Lexikalische Wachstumsmodelle

Die These, daß "infolge der stetigen Erweiterung der Sphäre der menschlichen Tätigkeit die Lexik jeder Sprache, besonders ihr terminologisches Vakobular, ungeachtet des Verschwindens einiger Wörter ständig wächst" (Piotrovskij u.a., 1977: 56), ist allgemein bekannt. Vom mathematischen Gesichtspunkt entspricht diesem ständigen Anwachsen des Vokabularumfangs vor allen Dingen das exponentielle Wachstum nach der Formel (Piotrovskij u.a., 1977: 57):

$$(4.1) \qquad L_T = L_0 e^{kT},$$

[2] Über den Status terminologischer Wörterbücher s. Gerd (1986).

wobei T die Zeitspanne (z.B. Jahrhundert, Jahrtausend), L_T den Vokabularumfang am Ende der Zeitspanne T, L_0 den Anfangsumfang des Vokabulars, k den Koeffizienten des Zuwachses und e die Basis des natürlichen Logarithmus darstellen. Laut diesem Gesetz hat die Geschwindigkeit des Vokabularwachstums einen "lawinenartigen" Charakter (die Geschwindigkeit des Anwachsens des Vokabulars ist proportional zum erreichten Zustand), den man mit Hilfe der Differentialgleichung

$$(4.2) \qquad \frac{dL_T}{dT} = kL_T \quad (k > 0)$$

darstellen kann, wobei k eine Konstante ist. Aus dieser Gleichung folgt, daß die Zuwachsrate linear von dem erreichten Zustand L_T abhängig ist und die relative Zuwachsrate $\frac{dL_T/dT}{L_T}$ (das Zuwachstempo) konstant bleibt. Löst man diese Gleichung, so erhält man (4.1).

Zur Überprüfung dieser Hypothese wurden Daten über Lexikwachstum aus der estnischen Literatursprache aufgrund "repräsentativer" Wörterbücher des 17. bis 20. Jahrhunderts herangezogen (Tuldava 1984a). Die estnische Literatursprache trat im 16. Jahrhundert in Erscheinung und durchlief die Stadien des "Entstehens, Formierens und Stabilisierens". Diese Stadien spiegeln sich in der Zusammensetzung und im Wachstum der repräsentativen (für ihre Zeit vollständigsten und normativen) Wörterbüchern wieder (s. Tabelle 4.4).

Tabelle 4.4
Anwachsen der Lexik nach Wörterbüchern des 17.-20. Jhdts

Nr.	Jahr	Zahl der Wörter
1.	1660	10000
2.	1780	14000
3.	1818	21000
4.	1869	50000
5.	1893	60000
6.	1930	120000
7.	1960	105000
8.	1976	115000

Die Analyse zeigt, daß das Wachstum des Vokabulars in diesen Wörterbüchern im Zeitraum von 1780 bis 1930 exakt dem Exponentialgesetz folgt (d.h. von Wörterbuch Nr. 2 bis Nr. 6). Die Parameter von (4.1) sind $L_0 = 1000$ und $k = 1{,}45$. Die Übereinstimmung der empirischen und der theoretischen Werte ist gut (Kurve

I in Abb. 4.3; s. auch Tabelle 4.5; die Zahlen sind auf Tausende gerundet)[3]. Nach dem Exponentialgesetz des Wachstums verdoppelt sich im gegebenen Fall der Umfang alle 48 Jahre,[4] d.h. der Vokabularumfang verdoppelt sich ungefähr jeweils in einem halben Jahrhundert. Bei diesem Wachstumstempo kann man voraussagen, daß der Umfang des "repräsentativen" (orthologischen oder erklärenden) Wörterbuches im Jahre 2000 330000 Wörter betragen wird, und im Jahr 2100 etwa anderthalb Millionen.

Dieses exponentielle Gesetz spiegelt die Realität offensichtlich in dem Falle wider, wenn man bei der Ermittlung des Vokabularumfangs der Literatursprache nicht nur allgemein gebräuchliche Wörter, sondern auch eng spezialisierte Termini einbezieht, wobei man den Wegfall eines Wortes aus der lebendigen Sprache (Absterben oder Austausch des Wortes) außer Acht läßt, mit anderen Worten, wenn man das Vokabularwachstum als einen kumulativen Prozeß betrachtet. Im Leben und in der lexikographischen Praxis ist es jedoch normalerweise nicht so. Man muß annehmen, daß ein unaufhaltsames Wachstum des Vokabulars nicht für immer währen kann und früher oder später bremsende Faktoren erscheinen, die sowohl durch interne (z.B. Sättigung des Vokabulars mit allgemein gebräuchlichen Wörtern in einer entwickelten Literatursprache) als auch externe Ursachen bedingt sind, worunter auch die Bedürfnisse der Gemeinschaft und die regulierende Tätigkeit der Lexikographen fallen.

Es ergibt sich also, daß das Wachstum der allgemein gebräuchlichen Lexik und das Wachstum des Umfangs der Literatursprache in verschiedenen Zeitspannen sich mit dem exponentiellen Gesetz nur in bestimmten Perioden der Sprachentwicklung charakterisieren lassen. In der Realität beginnt der Prozeß der Entwicklung der Lexik langsam (Periode der Entstehung der Literatursprache), dann beschleunigt er sich und nimmt eine "lawinenartige" Form an (Periode der Formierung der Literatursprache), aber von einem bestimmten Moment an verlangsamt er sich notwendigerweise (Periode der Stabilisierung). Diesem Entwicklungsschema entspricht ein anderes mathematisches Modell, die sogenannte logistische Funktion:

$$(4.3) \qquad L_T = \frac{L_n}{1 + ae^{-kT}},$$

wobei L_T der Vokabularumfang am Ende der Zeitspanne T, L_N die theoretische Wachstumsgrenze (Asymptote) und a und k Parameter der Funktion sind. Graphisch kann man dieses Modell als eine S-Kurve darstellen, die zuerst ein beschleunigtes Anwachsen, dann aber Verlangsamung und fast Stillstand bei asymptotischer Annäherung an eine Grenze darstellt (Kurve II in Abb. 4.3). Das Gesetz des logistischen Wachstums gilt für die Lexik der estnischen Literatursprache dann, wenn man die Wörterbücher Nr. 1 (1660) und Nr. 6 (1930; in diesem Wörterbuch werden auch veraltete Wörter kumulativ aufgeführt) unberücksichtigt läßt. Die Parameter der Funktion sind L_T = 150000, a = 280, k = 1,8[5]. Die Übereinstimmung ist gut (s. Tabelle 4.5).

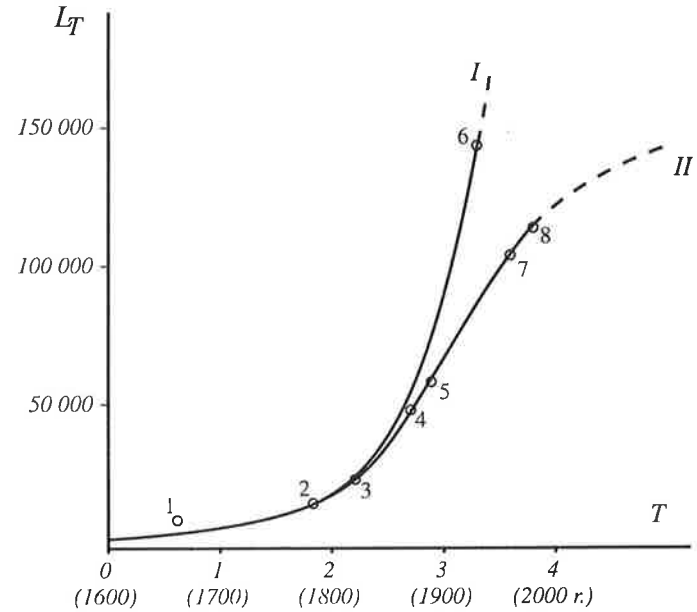


Abbildung 4.3. Lexikonwachstum der estnischen Literatursprache nach Wörterbüchern des 16.-20. Jhds. Ausgleich und Progonose nach der exponentiellen (I) und der logistischen (II) Funktion. Die Zahlen in der Abbildung stehen für die einzelnen Wörterbücher (vgl. Tabelle 4.4)

Die Formeln (4.1) und (4.3) stimmen insofern überein, als die beiden Kurven im Anfangsbereich ($L_n >> L_T$) praktisch zusammenfallen - wie auch gut in der Abbildung zu sehen ist. Auf diese Weise subsumiert (4.3) den exponentiellen Trend im ersten Stadium des Wachstums. Die logistische Kurve hat einen Wendepunkt, nach

[3] Den Anfang bildet das Jahr 1600 (T = 0). Die weiteren Zahlen bedeuten den Abstand vom Anfang, z.B. Jahr 1780 ergibt T = 1,8.

[4] Den Verdoppelungzeitraum berechnet man als $T_y = (ln\ 2)/k$. Im gegebenen Fall ist k = 1,45, daher ist $T_y = (ln\ 2)/1{,}45 = 0{,}48$, also 48 Jahre (*mit T_y in Jahrhunderten*).

[5] Verschiedene Methoden für die Schätzung der Parameter von (4.3) findet man in Altmann (1983).

Tabelle 4.5
Empirische und theoretische Daten des Vokabularwachstums nach Wörterbüchern aus dem 16.-20. Jhds.
(L'_T - exponentielles Gesetz, L''_T - logistisches Gesetz)

Nr.	Jahr	T	L_T (beob.)	L'_T (theor.)	L''_T (theor.)
1.	1660	0.6	10000	2000	2000
2.	1780	1.8	14000	14000	13000
3.	1818	2.2	21000	24000	24000
4.	1869	2.7	50000	50000	50000
5.	1893	2.9	60000	67000	60000
6.	1930	3.3	120000	120000	86000
7.	1960	3.6	105000	185000	105000
8.	1976	3.8	115000	247000	115000
-	2000	4.0	Prognose:	330000	124000

dem eine ständige Verlangsamung der Wachstumsgeschwindigkeit beginnt. Nach diesem Punkt erinnert die Kurve an die logarithmische Funktion, die das "Gesetz der adaptiven Verlangsamung" darstellt (Nalimov, Mul'čenko, 1969). Die Analyse zeigt, daß die Kurve der logarithmischen Funktion

(4.4) $$L_T = \alpha + \beta \ln T$$

(im konkreten Fall mit Parametern α = -160000 und $\beta \approx 206000$) mit der logistischen Funktion (4.3) in der Stabilisierungsphase der Entwicklung der Literatursprache (anfangend ungefähr um1900) praktisch zusammenfällt.

Das Gesetz des logistischen Wachstums in seiner allgemeinen Form (Beschleunigung - Wendepunkt - Verlangsamung) hat mit großer Wahrscheinlichkeit eine allgemeine sozial-linguistische Relevanz und charakterisiert das Wachstum der Lexik der meisten Literatursprachen, auch wenn es eine konkrete Form in Abhängigkeit von den Bedingungen der historischen Entwicklung des gegebenen Volkes - der Sprachträger - annimmt. Man kann hinzufügen, daß das Gesetz des logistischen Wachstums in seinen verschiedenen konkreten Realisierungen (es gibt eine Reihe von Varianten des logistischen Wachstums) zu den grundlegenden Gesetzen der Entwicklung von selbstorganisierenden Systemen gehört, wenn man ihre Entwicklung in hinreichend großen Zeiträumen betrachtet. Die S-förmige Kurve charakterisiert auch einige andere diachrone linguistische Prozesse ("das Piotrowski-Gesetz"; vgl. Altmann, 1983), und in unserer Zeit verwendet man sie häufig in anderen wissenschaftlichen Disziplinen, z.B. sogar bei der Modellierung der Entwicklung der Wissenschaft selbst (z.B. Price, 1966).

Neben der Veränderung des Lexikonumfangs interessiert uns auch die Entwicklung des lexikalischen Bestandes im Laufe der Zeit. Der Vokabularbestand entwickelt sich ständig, indem er sich durch neue Wörter ergänzt und alte Wörter eliminiert, wobei ein bestimmter Teil für lange Zeit unverändert bleibt. Ein Vergleich von Wörterbüchern aus verschiedenen Epochen kann uns Material zur Klärung der Gesetzmäßigkeiten der Entwicklung des Vokabulars sowohl in qualitativer als auch in quantitativer Hinsicht liefern. In diesem Abschnitt werden wir nur einige Möglichkeiten des quantitativen Vergleichs der Vokabularbestände erörtern (für eine entsprechende qualitativ-quantitative Analyse s. Tuldava, 1984a).

Um eine allgemeine Vorstellung vom Ausmaß der Veränderung des Lexikons in der Periode der Entwicklung der estnischen literarischen Nationalsprache vom Ende des 19. Jhdts. bis heute zu bekommen, haben wir die Lexeme der Wörterbücher Nr. 5 und Nr. 8 verglichen (vgl. Tabelle 4.4). Eine Stichprobenanalyse zeigte, daß die Gesamtzahl der Wörter mit anlautendem L im Wörterbuch Nr. 5 (Estnisch-deutsches Wörterbuch von Wiedemann, 2. Auflage, 1893) sich auf 4700 beläuft, während es im Wörterbuch Nr. 8 (Orthologisches Wörterbuch, 1976) 7732 gibt. Die Zahl gemeinsamer Wörter in den beiden Wörterbüchern beträgt 1550 (als gemeinsam galten Wörter, die formal identisch waren und wenigstens eine gemeinsame Bedeutung hatten). Die Resultate des Vergleichs der Fragmente aus den beiden Wörterbüchern kann man folgendermaßen darstellen:

Wörterbuch 1893:	Zahl der Wörter im Fragment:	4700
	Davon identisch mit Wörterbuch 1976:	1550 (d.h. 33%)
	Nur im Wörterbuch 1893:	3150 (d.h. 67%)
Wörterbuch 1976:	Zahl der Wörter im Fragment:	7732
	Davon identisch mit Wörterbuch 1976:	1550 (d.h. 20%)
	Nur im Wörterbuch 1976:	6182 (d.h. 80%).

Zur Illustration kann man diese Daten in Form eines Schemas darstellen, indem man die beiden Fragmente vereinigt und den gemeinsamen Teil hervorhebt (Abb. 4.4).

Die Daten des quantitativen Vergleichs sagen aus, daß 33%, d.h. nur ein Drittel des Fragments des Wörterbuchs von 1893 bis heute erhalten blieb und zwei Drittel keinen Platz im modernen orthologischen Wörterbuch gefunden haben. Andererseits bilden die erhaltenen Wörter aus dem Wörterbuch 1893 etwa 20% des Wörterbuches von 1976, d.h. ein Fünftel, und alle anderen scheinen "neue" Wörter zu sein. Dies alles zeugt von tiefgreifenden qualitativen Veränderungen in der Entwicklung der Literatursprache des 20 Jhdts.

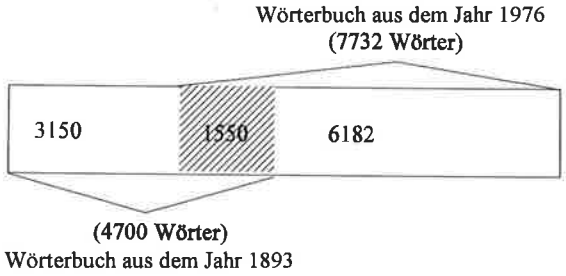


Abbildung 4.4. Verteilung der gemeinsamen und nicht gemeinsamen Lexik beim Vergleich (Vereinigung) beider Wörterbücher (Fragmente). Der gemeinsame Teil ist schraffiert.

Die lexikalische Nähe zweier Vokabularfragmente kann man mit Hilfe verschiedener Indizes messen. Bei der Vereinigung der Fragmente legen wir sie sozusagen übereinander und trennen ihren gemeinsamen und die beiden spezifischen Teile voneinander (Abb. 4.4). Das Verhältnis des gemeinsamen Teiles (C) zu dem Umfang des vereinigten Vokabulars ($A+B$) ergibt sich als

$$(4.5) \qquad R_I = \frac{C}{A + B - C}.$$

Der Assoziationsindex R_I nimmt Werte zwischen 0 und 1 an. Wenn $C = 0$, dann ist offensichtlich $R_I = 0$, d.h. die Wörterbücher haben keine gemeinsamen Wörter. $R_I = 1$ bei theoretisch maximaler Assoziation, wenn nämlich die beiden Wörterbücher identisch sind ($A = B = C$). In unserem Fall ergab die Auswertung der Assoziation zwischen den Wörterbüchern

$$R_I = \frac{1550}{4700 + 7732 - 1550} = 0{,}142.$$

Dies bedeutet, daß der gemeinsame Teil 14,2% des vereinigten Vokabulars ausmacht.

Es gibt auch andere Assoziationsindizes. Man kann beispielsweise das Verhältnis des gemeinsamen Teils (C) zum durchschnittlichen Umfang beider Wörterbücher als

$$(4.6) \qquad R_{II} = \frac{C}{(A + B)/2} = \frac{2C}{A + B}$$

ausdrücken, was in unserem konkreten Fall

$$R_{II} = \frac{2(1550)}{4700 + 7732} = 0{,}249$$

ergibt. Auch R_{II} liegt zwischen 0 und 1, läßt sich aber mit R_I nicht direkt vergleichen, da sie Unterschiedliches ausdrücken. Der konkrete Wert $R_{II} = 0{,}249$ bedeutet inhaltlich interpretiert, daß die gemeinsamen Wörter durchschnittlich 24,9% des Umfangs jedes einzelnen Vokabulars (Fragments) bilden.

Bei stark abweichenden Umfängen der verglichenen Wörterbücher empfiehlt es sich, nicht den arithmetischen Mittelwert wie in Formel (4.6) zu benutzen, sondern den geometrischen Mittelwert nach der Formel

$$(4.7) \qquad R_{III} = \frac{C}{\sqrt{A \cdot B}}.$$

In unserem Beispiel ergibt sich nach (4.7) das Resultat $R_{III} = 0{,}257$. Der Unterschied zu R_{II} ist nicht allzu groß, da sich die Vokabularumfänge (Fragmentumfänge) nicht sehr unterscheiden. R_{III} ist jedoch theoretisch besser begründet, da er eine Verallgemeinerung darstellt und den Index R_{II} als Spezialfall einschließt: wenn $A = B$, dann $R_{III} = R_{II}$.

Alter und Häufigkeit der Wörter

In der letzten Zeit hat man in vielen Untersuchungen beobachtet, daß es einen Zusammenhang zwischen dem "Alter" des Wortes, d.h. der Zeit seines Entstehens in der Sprache, und seiner Verwendungshäufigkeit in der heutigen Sprache gibt (s. z.B. Arapov, Cherc 1974, 1983; Embleton 1986).

Beim quantitativ-systemischen Ansatz zur Erforschung des Zusammenhangs zwischen Alter und Häufigkeit des Wortes ist es zweckmäßig, auf das Modellieren mit Hilfe von Verteilungen zurückzugreifen. Man muß solche systemischen Eigenschaften der Objekte finden, die der Aufstellung von Modellen und ihrer inhaltlichen Interpretation dienen können. Im gegebenen Fall ist es am zweckmäßigsten, die Methode von Arapov und Cherc (1974, 1983) zu benutzen, die in einer speziellen Untersuchung den Zusammenhang zwischen Alter und Häufigkeit des Wortes für eine adäquate mathematische Modellierung der Evolution des Vokabulars erforschen.

Nach dieser Methode werden die empirischen Daten - im gegebenen Fall Wörter im heutigen Häufigkeitswörterbuch - in Häufigkeitszonen (Gruppen) von jeweils 100 Wörtern je Zone zusammengefaßt. Den geordneten Zonen werden Zahlen oder Ränge (i) zugeschrieben. In jeder Zone sucht man die Zahl der "älteren" Wörter (d.h. Wörter, deren Entstehung zu einem früheren Zeitpunkt datiert ist) und die entsprechende relative Häufigkeit bzw. den Anteil.

Zur Illustration bringen wir Angaben über Verteilungen älterer Wörter nach Häufigkeitszonen in Häufigkeitswörterbüchern verschiedener Sprachen (s.Tabelle 4.6). Die Daten wurden aus dem Werk von Arapov und Cherc (1974) übernommen

und beruhen auf folgenden Häufigkeitswörterbüchern: Französisch (Gougenheim), Deutsch (Kaeding), Englisch (Dewey), Russisch (Štejnfeldt), Tschechisch (Jelinek, Bečka, Těšitelová).

Tabelle 4.6
Verteilung der Wörter älteren Ursprungs in Häufigkeitszonen (*i*) in sechs Sprachen

Rang (*i*) (Zone)	Sprachen					
	Estn.	Franz.	Dt.	Engl.	Russ.	Tschech.
	Datierung (Jahr)					
	1200	1200	1100	1100	600	600
1 (1-100)	91	91	94	92	84	75
2 (101-200)	77	84	88	70	57	63
3 (201-300)	70	84	83	53	52	50
4 (301-400)	66	71	73	40	51	43
5 (401-500)	60	73	63	47	42	37
6 (501-600)	54	52	55	32	32	36
7 (601-700)	46	57	55	29	42	45
8 (701-800)	44	55	59	36	35	42
9 (801-900)	43	52	52	31	33	32
10 (901-1000)	37	61	53	31	35	32
Gesamt: Wörter älteren Ursprungs	588	680	675	461	463	455

Unter den ersten tausend häufigsten Wörtern findet man die größte Zahl älterer Wörter im französischen und im deutschen Wörterbuch (680 bzw. 675); im Estnischen gibt es etwas weniger (588) und beträchtlich weniger im Englischen (461). Die Daten dieser vier Sprachen beruhen auf derselben Datierung älterer Wörter (1100-1200 n.Chr.), wobei die Regeln für die Identifizierung älterer Wörter ungefähr gleich sind. Im Russischen und Tschechischen bezieht sich die Datierung auf eine frühere Periode, nämlich auf die Zeit des Zerfalls der urslawischen Einheit (etwa 600 n.Chr.). Die Anzahl der Wörter in beiden Wörterbüchern ist ungefähr gleich (463 bzw. 455).

In allen analysierten Sprachen kann man den Zusammenhang zwischen Alter und Häufigkeit der Wörter insofern beobachten, als der Anteil älterer Wörter mit der Zunahme des Ranges der Häufigkeitszone, d.h. mit Abnahme der durchschnittlichen Häufigkeit abnimmt. Es stellt sich die Frage nach der Form der mathematischen Abhängigkeit zwischen der Zahl älterer Wörter $F(i)$ oder ihren Anteilen $p(i)$ und dem Rang der Häufigkeitszone i. Diese Abhängigkeit muß man probabilistisch behandeln, formal aber als eine Funktion.

Tabelle 4.7
Empirische und theoretische Verteilung von Häufigkeiten alterer Wörter in drei Sprachen; der Parameter a und der Koeffizient e^{-a} des Exponentialgesetzes $p(i) = e^{-ai}$; in der Tabelle sind die absoluten Häufigkeiten $F(i) = 100\ p(i)$ aufgeführt

Rang (*i*) (Zone)	Estnisch		Französisch		Deutsch	
	Beob.	Theor.	Beob.	Theor.	Beob.	Theor.
1 (1-100)	91	91	91	94	94	93
2 (101-200)	77	82	84	89	88	87
3 (201-300)	70	74	84	84	83	81
4 (301-400)	66	67	71	79	73	76
5 (401-500)	60	61	73	74	63	70
6 (501-600)	54	55	52	70	55	66
7 (601-700)	46	50	57	66	55	61
8 (701-800)	44	45	55	62	59	57
9 (801-900)	43	41	52	58	52	53
10 (901-1000)	37	37	61	55	53	50
Parameter: a	0,1		0,06		0,07	
e^{-a}	0,905		0,942		0,932	

In erster Approximation kann man annehmen, daß die Abnahme der Zahl (des Anteils) älterer Wörter mit wachsendem Rang dem Exponentialgesetz folgt, wobei das mittlere Tempo der Abnahme konstant bleibt. In solch einem Fall kann man die Verteilung der älteren Wörter in Häufigkeitszonen in Form einer fallenden Exponentialfunktion darstellen, die sich der Abszisse asymptotisch nähert (vgl. Abb. 4.5). Der analytische Ausdruck dieser Kurve hat die Form

$$p(i) = e^{-ai}, \tag{4.8}$$

wobei $p(i)$ die relative Häufigkeit (Proportion) älterer Wörter in der i-ten Zone, e die Basis des natürlichen Logarithmus und a der Parameter (Konstante) sind. Wie bekannt, hat M. Swadesh (1960) diese einfache Abhängigkeit in seiner Theorie der Glottochronologie im Bezug auf die Wahrscheinlichkeit der Bewahrung älterer

Wörter in bestimmten Zeitintervallen in der Sprachgeschichte postuliert.[6]

Funktion (4.8) beschreibt hinreichend genau die Dynamik der Abnahme älterer Wörter im Rahmen von 10-12 Häufigkeitszonen im Estnischen, Französischen und Deutschen (vgl. Tabelle 4.7). Besonders gut zeigt sich dieser Zusammenhang im Estnischen, wo man als Ausgangspunkt das Basisvokabular einer Fachsprache gewählt hat (Tuldava, 1982). Der Graph (vgl. Abb. 4.6) zeigt, daß die lineare Abhängigkeit zwischen *ln p(i)* und *i*, die dem Exponentialgesetz entspricht, zufriedenstellend verläuft.

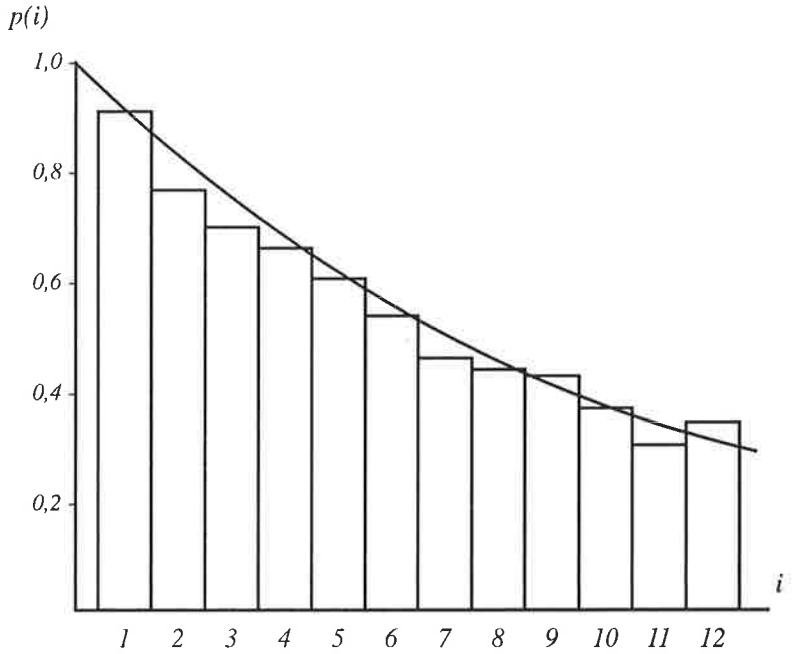


Abbildung 4.5. Verteilung der älteren Lexik im Häufigkeitswörterbuch des Estnischen. Zusammenhang zwischen dem Anteil älterer Wörter $p(i)$ und dem Rang der Häufigkeitszone (i). Graph der Funktion $p(i) = e^{-0,1\,i}$

[6] Formel (4.8) entspricht der Abhängigkeit, die man in der Archäologie Radiokarbondatierung nennt. Dies war der unmittelbare Stimulus für die Glottochronologie. Wir können noch anmerken, daß die Komponente e^{-a} in Formel (4.8) als das mittlere Abnahmetempo interpretiert werden kann, beispielsweise mit $e^{-a} = 0,905$ beträgt der Anteil älterer Wörter 0,905 oder 90,5% des vorherigen Zustandes.

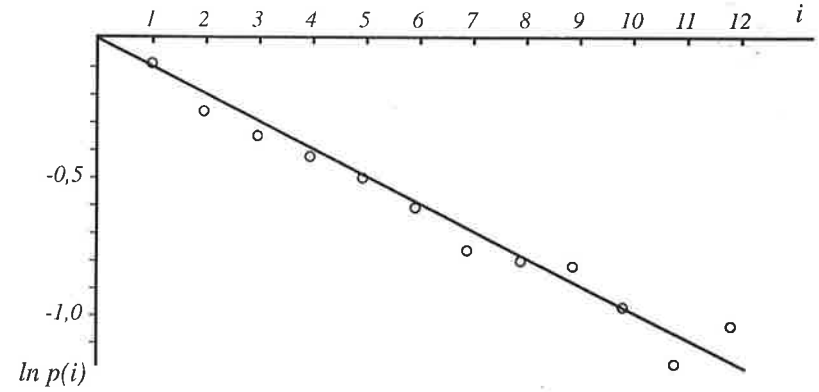


Abbildung 4.6. Linearer Zusammenhang zwischen dem Logarithmus des Anteils *ln p(i)* und dem Rang *i*.

Funktion (4.8) gilt jedoch nicht für alle Sprachen und Vokabulare, z.B. Russisch, Tschechisch und Englisch (gemäß den Daten der analysierten Wörterbücher). M.V. Arapov und M.M. Cherc (1974) schlugen die folgende verallgemeinernde Formel vor:

$$p(i) = e^{-a\sqrt{i}}. \tag{4.9}$$

Auch wenn diese Formel z.B. für das Russische geeignet ist, liefert sie für die anderen analysierten Wörterbücher zu approximative Schätzungen. Man muß daher nach einer anderen verallgemeinernden Formel suchen. Es ist leicht zu zeigen, daß Formeln (4.8) und (4.9) Spezialfälle einer allgemeineren Funktion mit drei Parametern *c*, *a* und *b* sind:

$$p(i) = ce^{-ai^b} \tag{4.10}$$

In den Formeln (4.8) und (4.9) ist der Parameter $c = 1$ von vornherein festgelegt, d.h. er gibt das Maximum der Wahrscheinlichkeit älterer Wörter an. Dies sieht man auch in der Modifikation der Formel (4.10) für absolute Häufigkeiten:

$$F(i) = ne^{-ai^b}, \tag{4.11}$$

wobei $F(i)$ die absolute Häufigkeit älterer Wörter und n der Umfang der Gruppe (Häufigkeitszone) sind, d.h. das Maximum oder der Ausgangspunkt, auf Grund dessen man die Wahrscheinlichkeit $p(i) = F(i)/n$ bestimmt. Im gegebenen Fall ist $n = 100$.

Der Unterschied zwischen den Formeln (4.8) und (4.9) liegt darin, daß im ersten Fall der Parameter b auf 1 festgelegt ist, während im zweiten Fall $b = 0{,}5$ (d.h. $\sqrt{i} = i^{0,5}$). Es scheint, daß die freie Variation der Parameters b zweckmäßiger ist, da er als quantitativ-linguistische Kenngröße für Sprachen und Wörterbücher charakteristisch ist. Inhaltlich drückt der Parameter b das Abnahmetempo der Wahrscheinlichkeit des Auftretens älterer Wörter mit Abnahme der Worthäufigkeit aus. Auch der Parameter a läßt sich inhaltlich interpretieren: Er stellt eine Charakteristik der Konzentration älterer Wörter im Anfangsteil des Häufigkeitswörterbuches dar (kleineres a bedeutet größere Konzentration). Für die estnische Sprache bekommen wir aus (4.10) $a = 0{,}11$ und $b = 0{,}96$ (da $b \approx 1$, ist in diesem Fall die Anwendung der Exponentialfunktion (4.8) gerechtfertigt). Für das Russische (im Intervall i = 1...25) sind die Parameter $a = 0{,}32$ und $b = 0{,}51$ (wegen dieses Wertes von b ist die Anwendung von (4.9) gerechtfertigt).

Die Frage der analytischen Beschreibung der Verteilung älterer Wörter im Vokabular kann man auch auf andere Weise erörtern, nämlich aufgrund der *kumulativen* Verteilung der Wörter. In diesem Falle muß man die Konzentration der Häufigkeiten älterer Wörter nach kumulativen Häufigkeitszonen (erste Zone, erste und zweite Zone zusammen usw.) betrachten. Bei der Wahl entsprechender Funktionen muß man berücksichtigen, daß das Anwachsen der Zahl älterer Wörter eine Grenze hat und daß das Anwachsen sich verlangsamt, je mehr man sich der Grenze nähert. Unter Berücksichtigung des Zusammenhangs der Variablen nach (4.10) kann man die Funktion der kumulativen Verteilung als

$$p^*(i) = 1 - e^{-ai^b} \qquad (4.12)$$

ableiten. Hier ist $p^*(i)$ die Wahrscheinlichkeit, die dem Verhältnis $F^*(i)/F_n$ entspricht, wobei $F^*(i)$ die kumulative Häufigkeit älterer Wörter und F_n die Grenze der Anzahl älterer Wörter in der gegebenen Stichprobe ist.[7] Formel (4.12) ist identisch mit dem Weibull-Gesetz (Weibull 1939; vgl. auch Bektaev, Piotrovskij, 1973: 136-138).

In Tabelle 4.8 findet man die Resultate der Anwendung von (4.12) für sechs Sprachen, die untergleichen experimentellen Bedingungen untersucht wurden (in Intervallen zu Hundert für alle Sprachen). Man kann eine gute Übereinstimmung zwischen empirischen und theoretischen Werten beobachten. Im Estnischen ergibt sich die Prognose für die Grenze der älteren Wörter auf $F_n \approx 1000$ für das gegebe-

[7] Die Parameter a und b kann man mit Hilfe der Methode der kleinsten Quadrate nach der Linearisierung $ln\ ln(1/(1-p^*(i))) = ln\ a + b\ ln\ i$ schätzen. Hier ist $ln\ a$ die Anfangsordinate, b der Winkelkoeffizient. F_n schätzt man iterativ; man behält den Wert von F_n, bei dem die Übereinstimmung zwischen den empirischen und theoretischen Werten am besten ist. Den ersten Schritt kann man nach der Linearisierung mit Hilfe des Graphen durchführen.

Tabelle 4.8
Kumulative Häufigkeitsverteilung älterer Wörter in sechs Sprachen: empirische und theoretische Werte; Parameter der Weibullverteilung

Rang (Zone)	Estnisch		Französisch		Deutsch		Englisch		Russisch		Tschechisch	
	Beob.	Theor.	Beob.	Theor.	Beob.	Theor.	Beob.	Theor.	Beob.	Theor.	Beob.	Theor.
1 (1-100)	91	91	91	93	94	96	92	94	84	84	75	76
2 (101-200)	168	170	175	177	182	180	162	158	141	142	138	133
3 (201-300)	238	241	259	255	265	257	215	212	193	193	188	184
4 (301-400)	304	305	330	327	338	329	255	258	244	242	231	230
5 (401-500)	364	364	403	395	401	395	302	300	286	282	268	273
6 (501-600)	418	417	455	459	456	458	334	337	318	321	304	313
7 (601-700)	464	466	512	519	511	516	363	371	360	359	349	351
8 (701-800)	508	510	567	576	570	570	399	408	395	394	391	387
9 (801-900)	551	551	619	630	622	622	430	431	428	428	423	421
10 (901-1000)	588	588	680	680	670	670	461	458	463	460	455	453
Parameter: F	1000		1600		1500		900		1900		1500	
a	0,095		0,060		0,066		0,110		0,045		0,052	
b	0,971		0,965		0,953		0,807		0,792		0,836	

ne Wörterbuch (Autorensprache in künstlerischer Prosa), d.h. etwa 7% des ganzen Vokabulars von etwa 14,7 Tausend Wörtern. Man muß berücksichtigen, daß F_n die Zahl älterer Wörter im ganzen Wörterbuch voraussagt, und zwar unter der Bedingung, daß das Wachstum auch jenseits der empirischen Daten konstant bleibt. Es ist selbstverständlich, daß man zur Erhöhung der Zuverlässigkeit der Prognose das empirische Material erweitern muß. Aber für eine vergleichende typologische Analyse reicht es offensichtlich, wenn man die ersten tausend häufigsten Wörter betrachtet (vgl. auch Arapov, Cherc, 1974: 59). Die Kenngröße F_n kann man in diesem Fall als die relative Schätzung der "Archaizität" des gegebenen Wörterbuches betrachten. Auch den anderen Parametern der Verteilung kann man inhaltlich interpretieren. Der Parameter a drückt (in bezug auf F_n) den Grad der Konzentration älterer Wörter am Anfang des Vokabulars aus (größeres a bedeutet größere Konzentration). Der Parameter b drückt das Wachstumstempo aus.

Die Anwendung der Weibullfunktion als Modell der Verteilung älterer Wörter im Vokabular zeigt, daß die Parameter dieser theoretischen Verteilung direkt oder indirekt als quantitativ-linguistische Charakteristika und stildifferenzierende Faktoren bei der Analyse der Lexik dienen können. Mit Hilfe dieser Verteilungsfunktion kann man im Rahmen des Wörterbuches sowohl extra- als auch interpolieren. Im großen und ganzen spiegeln die Parameter der Weibullverteilung sowohl globale Eigenschaften als auch systemische Wechselbeziehungen zwischen den Elementen des lexikalischen Systems wider und verweisen gleichzeitig auf die systemische Beziehung zwischen Alter und Häufigkeit der Wörter.

4.3. Lexikalisch-stilistische Analyse von Texten

Unter dem Gesichtspunkt der quantitativ-linguistischen Analyse von Texten gibt es eine Reihe aktueller Probleme, die mit dem stilistischen Aspekt der Lexik eines gegebenen Textes verbunden sind, insbesondere Fragen zum "Vokabularreichtum" des Textes, zur lexikalischen Nähe von Texten und zur Klassifizierung von Texten aufgrund quantitativer lexikalisch-stilistischer Charakteristika.

Vokabularreichtum des Textes

Die Suche nach objektiven Methoden der Bewertung des Vokabularreichtums von Texten ist schon seit langem Gegenstand des Interesses zahlreicher Forscher, die sich mit den Problemen der quantitativen Untersuchung von Individual- und Funktionalstilen beschäftigen (Mistrík, 1967; Muller, 1968; Woronczak, 1972; Těšitelová, 1972; Ratkowsky et al., 1980 usw.). Der Vokabularreichtum wird im quantitativen Sinne allgemein als die Anzahl unterschiedlicher Wörter im Text definiert,

d.h. als Vokabularumfang des Textes oder als das Verhältnis der Zahl unterschiedlicher Wortformen (V) oder Lexeme (L) zum Textumfang (N), d.h. als V/N oder L/N. Dieses Verhältnis bezeichnet man als "Diversitätsindex" oder "TTR-Index" (type-token ratio). Je größer der numerische Wert des Index, desto mehr unterschiedliche Wörter (Wortformen oder Lexeme) benutzt der Schreiber oder der Sprecher im gegebenen Text.

Beim Vergleich des Vokabularreichtums verschiedener Texte muß man aber berücksichtigen, daß ein direkter Vergleich zweier Indizes nur dann möglich ist, wenn die Texte den gleichen Umfang haben, da das Verhältnis zwischen Textumfang und Vokabularumfang in Abhängigkeit von N variiert (vgl. Abschnitt 2.1). In Fällen, wo man Vokabulare von Texten verschiedener Umfänge vergleicht, muß man auf spezielle Methoden zurückgreifen (s. u.).

Ein hoher Stellenwert bei der Bewertung des Vokabularreichtums kommt auch den selten vorkommenden Wörtern im Vokabular des gegebenen Textes zu. Üblicherweise berechnet man das Verhältnis der sogenannten *hapax legomena* zum Vokabular- oder zum Textumfang. Die Anzahl der hapax legomena bezeichnet man gewöhnlich mit m_1 oder differenzierter mit V_1 für Wortformen und L_1 für Lexeme. So erhält man den "Einmaligkeitsindex" als V_1/V oder V_1/N (bei Wortformen) und L_1/L oder L_1/N für Lexeme. Wenn in einem Text sehr viele Wörter mit der Häufigkeit 1 vorkommen, dann könnte dies den Wunsch des Autors widerspiegeln, bildhafte Ausdrücke zu finden, seltene oder originelle Wörter zu wählen, die Wiederholung von Wörtern zu vermeiden usw. In diesem Fall bezeugt ein großer Anteil von Wörtern mit der Häufigkeit 1 den Reichtum und die Heterogenität der Lexik des Textes. Jedoch sind weder dieser Einmaligkeitsindex noch der Diversitätsindex ein ästhetisches Kriterium. Nur eine qualitative Analyse kann nachweisen, was sich hinter der großen Anzahl einmal vorkommender Wörter im Text verbirgt: guter Stil oder Weitschweifigkeit. Manchmal kann der Reichtum an seltenen Wörtern das Textverständnis erschweren oder Zeichen eines schlechten Stils sein; ein geringer Anteil an hapax legomena kann funktional begründet sein, z.B. bei der Wiedergabe einer spontanen Rede. Bei der Anwendung des Einmaligkeitsindexes sollte man auch bedenken, daß der Anteil von hapax legomena direkt vom Textumfang abhängt.

Zur Illustration bringen wir einen Vergleich der lexikalischen Komposition von russischen Texten verschiedener Autoren, wobei gleichgroße Stichproben von jeweils $N = 1000$ laufenden Wörtern aus der Autorensprache dreier Schriftsteller gewählt wurden: Für "Die achte Verletzung" von K. Simonov, "Das Schicksal des Menschen" von M. Šolochov und "Wir, sowjetische Menschen" von B. Polevoj, wurden die obigen quantitativen Charakteristika berechnet (vgl. Tabelle 4.9)[8]

[8] Die drei Erzählungen stammen aus dem Buch "Der russische Charakter", Moskau: Molodaja gvardija 1970.

Tabelle 4.9
Quantitative Charakteristika der Lexik für Stichproben aus der Autorensprache dreier russischer Schriftsteller der Sowjetzeit

Autor	N	L	L_1	L/N	L_1/L	L_1/N
M. Šolochov	1000	564	437	0,564	0,775	0,437
B. Polevoj	1000	524	392	0,524	0,749	0,392
K. Simonov	1000	443	297	0,443	0,671	0,297

Aus der Tabelle ist ersichtlich, daß man das größte Diversitätsmaß, also den größten Vokabularreichtum, bei Šolochov findet (0,564). Dasselbe kann man auch aus dem Vergleich der Einmaligkeitsindizes schließen, wobei die individuellen Unterschiede besonders stark auf der Textebene erscheinen (die Variationsspanne ist 0,437...0,297). Der Wert des Indexes $L_1/N = 0{,}437$ deutet darauf hin, daß hapax legomena (Lexeme) bei Šolochov 43,7% des gesamten Textes (d.h. der Stichprobe von 1000 laufenden Wörtern) ausmachen.

Die obigen Indizes - Diversitäts- und Einmaligkeitsindex - kann man als fundamentale, Standard-Charakteristika des Vokabularreichtums betrachten. In der quantitativen Linguistik gibt es aber noch viele andere quantitative Kenngrößen, die direkt oder indirekt den "Vokabularreichtum" im obigen Sinne diagnostizieren können. Traditionell wird der Parameter γ in der bekannten Zipfschen Formel $F_i = Ci^{-\gamma}$ (vgl. Abschnitt 2.2) als Maß des Vokabularreichtums betrachtet. Der Parameter γ - der Tangens des Winkels der Geraden in doppeltlogarithmischer Darstellung - weist direkt auf die "Ausdehnung" des Vokabulars des gegebenen Textes hin. Je kleiner γ, desto größer muß der Vokabularumfang sein. Wir wissen aber, daß Zipfs Formel die Häufigkeitsstruktur des Textes normalerweise nicht hinreichend erfaßt: Es gibt Strukturbruchstellen, die die fundamentale Zone der Wörter mittlerer Häufigkeit von der Zone der häufigen und der der seltenen Wörter trennen. Wenn man die Werte γ_1, γ_2, γ_3 für die drei Häufigkeitszonen (vgl. Abschnitt 2.2) vergleicht, dann sieht man, daß γ_3, die Kenngröße der seltenen Wörter, am empfindlichsten auf den Vokabularreichtum reagiert. Ein entsprechendes Experiment zeigte, daß die Kenngröße γ_3 eng mit dem Diversitätsindex und dem Einmaligkeitsindex korreliert ist. Der Parameter γ_2 (für die Zone der mittleren Häufigkeiten) zeigt einen mittelstarken Zusammenhang mit den Reichtumsindizes an, und der Parameter γ_1, also die Charakteristik der Zone der häufigsten Wörtern, steht mit den Indizes in keinem direkten Zusammenhang (für Details s. Tuldava, 1977a).

Oben haben wir erwähnt, daß man beim Vergleich von Texten ungleichen Umfangs besondere Methoden für die Bestimmung des Vokabularreichtums verwenden muß. Bekannt ist z.B. Guirauds (1954) Formel $R = L/\sqrt{N}$, die jedoch nur für Texte mit kleinem Umfangsunterschied verwendet werden kann. Etwas stabiler

ist der Koeffizient von Somers (1966): $R = (ln\ ln\ L) / (ln\ ln\ N)$. Theoretisch und praktisch besser begründet sind die Methoden, die die Dynamik des Vokabularwachstums bei wachsendem Textumfang berücksichtigen (z.B. Orlov, 1982). Wir verweisen hier auf die Möglichkeit, den Vokabularreichtum von Texten mit unterschiedlichem Umfang mit Hilfe von Formeln zu bestimmen, die den Zusammenhang zwischen Vokabular- und Textumfang analytisch ausdrücken (vgl. Abschnitt 2.3). Man kann beispielsweise Formel (2.40) verwenden, die es ermöglicht, die Umfänge verschiedener Texte "auszugleichen" und deren Parameter in Hinblick auf Tendenz und Form der Kurve des Vokabularwachstums interpretierbar sind.

Die Wahl einer geeigneten Formel hängt von den konkreten Bedingungen, Zielen und Aufgaben ab. Beim Vergleich kleiner Texte (etwa bei $N \leq 10000$; vgl. Tuldava, 1995, Abschn. 8.7) von ungefähr gleichem Umfang kann man eine einfache Formel benutzen, bei der man die Parameter inhaltlich interpretieren kann. Dazu eignet sich Formel (2.36), die man in folgender Form darstellen kann (auch Formel des ersten Grades von Tornquist genannt):

$$L = \frac{aN}{N + b}, \tag{4.13}$$

wobei a und b Parameter sind, die man nach Linearisierung mit der Methode der kleinsten Quadrate schätzen kann. Die lineare Beziehung besteht zwischen $1/L$ und $1/N$ (oder zwischen N/L und N); anstelle von L (Lexemzahl) kann man V (Zahl der Wortformen) benutzen. Funktion (4.13) läßt sich nämlich folgendermaßen in eine lineare Form transformieren:

$$L = \frac{aN}{N + b} \Rightarrow \frac{1}{L} = \frac{N + b}{aN} = \frac{1}{a} + \frac{b}{a}\cdot\frac{1}{N}.$$

Substituiert man $1/L = Y$, $1/a = A$, $b/a = B$ und $1/N = X$, dann ergibt sich $Y = A + BX$, d.h. zwischen $1/L$ und $1/N$ besteht eine lineare Beziehung . Analog kann man auch die Linearität zwischen N/L und N zeigen:

$$L = \frac{aN}{N + b} \Rightarrow \frac{N}{L} = \frac{N + b}{a} \Rightarrow \frac{N}{L} = \frac{b}{a} + \frac{1}{a}N,$$

d.h. $Y = A + BX$.

Die Formel drückt einen Typ von Hyperbelfunktion aus, die Asymptoten besitzt (s. Abb. 4.7).

Für unsere Zwecke hat nur der erste Quadrant Bedeutung, da $N \geq 0$ und $L \geq 0$. Parameter a steht für die Asymptote, die die obere Grenze von L bei $N \to +\infty$ angibt[9]. In unserem Fall können wir diesen Parameter als eine quantitativ-stilistische

[9] Die Funktion hat die Eigenschaft, für $N > 0$ zu einem Grenzwert zu streben. Wie man leicht berechnet, ist es $\lim_{N \to +\infty} \frac{aN}{N + b} = a$,

Kenngröße interpretieren, die die Wachstumstendenz des Vokabulars ausdrückt und dadurch den "potentiellen Reichtum" der Lexik beim Vergleich mehrerer Texte schätzt. Der Parameter *b* bestimmt das Tempo des Anwachsens der Lexik im Vergleich mit dem Anwachsen des Textumfangs (in Abb. 4.7 ist zu sehen man, wie die Veränderung von *b* die Form der Kurve bei festem *a* beeinflußt). Es ist zweckmäßig, das Verhältnis $a/b = \varphi$ als Stilcharakteristikum zu benutzten, da es das relative Tempo des Vokabularwachstums beschreibt: Je größer φ, desto stärker das Anwachsen des Vokabulars in den Anfangsstadien des Textes.

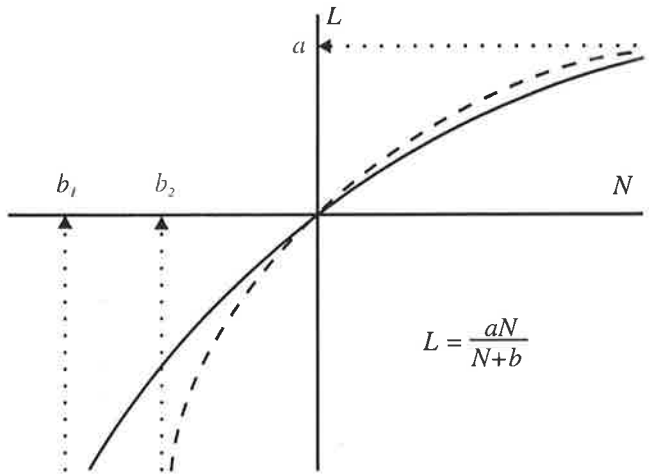


Abbildung 4.7. Die Kurve der Funktion $L = aN/(N+b)$ bei unterschiedlichen Werten von *b* und gleichem *a*

Die oben erwähnten quantitativ-stilistischen Kenngrößen, die die Dynamik des Vokabularwachstums charakterisieren, kann man auch kombinieren. Wenn man bei mehreren Untersuchungen mittlere Werte der Größen *a* und φ für ein Genre oder einen Stil erhält, dann kann man sie in ihren Konfidenzintervallen als "Normen" des gegebenen Genres oder Stils betrachten. Es ergeben sich 9 Kombinationen ("+" und "-" bedeuten "höhere" bzw. "niedrigere" Genrenormen, mit "=" bezeichnen wir die Werte innerhalb des Konfidenzintervalls):

a	φ	*a*	φ	*a*	φ
+	+	-	+	=	+
+	-	-	-	=	-
+	=	-	=	=	=

Die letzte Kombination (=/=) steht für den lexikalisch-stilistischen Texttyp, der für das gegebene Genre oder Stil am typischsten ist, da beide Charakteristika im Rahmen der Norm liegen. Besonders wichtig ist die Interpretation der folgenden vier Typen von Normabweichungen:

$+a/+\varphi$: ein großer Vokabularreichtum wird prognostiziert, wobei das Vokabular in den ersten Stadien des Textes mit schnellem Tempo anwächst; eine derartige Kombination von Charakteristika ist Autoren eigen, die in allen Phasen des Werkes beharrlich und zielstrebig an der Lexik arbeiten;

$+a/-\varphi$: potentiell großer Wortschatz, aber das Anwachsen des Vokabulars geschieht langsam, graduell;

$-a/+\varphi$: es wird ein kleiner Vokabularumfang prognostiziert; das Anfangstempo ist recht schnell, flaut aber bald ab (z.B. in Verbindung mit einer homogenen Thematik);

$-a/-\varphi$: kleiner Vokabularreichtum und langsames Vokabularwachstum.

Die Abweichungstypen -/+ (I) und +/- (II) sind in Abbildung 4.8. illustriert.

Zum Schluß möchten wir noch festhalten, daß das Problem der quantitativen Schätzung des Vokabularreichtums, das für die Stilistik und die Texttypologie so wichtig ist, mit den obigen Methoden nicht erschöpft ist. Das Problem hat noch darüber hinaus reichende Aspekte, wie z.B. die Untersuchung des Vokabularreichtums beschränkt auf einzelne Wortarten, Autosemantika, expressive Wörter usw. In allen Fällen ist es aber nötig, die quantitativen Resultate der Untersuchung qualitativ zu analysieren.

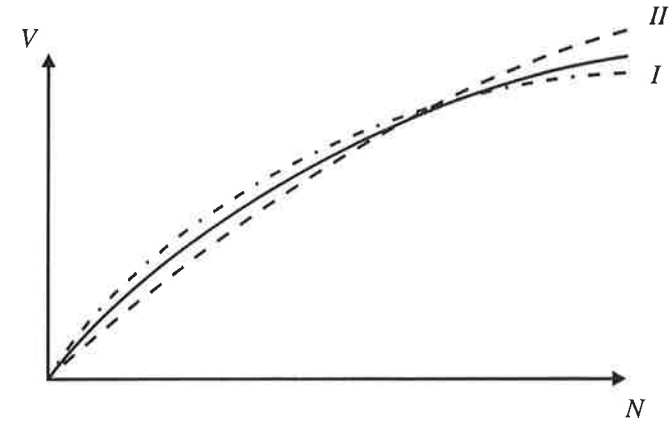


Abbildung 4.8. "Norm" (volle Linie) und die Varianten I und II: Die Abhängigkeit des Vokabularumfangs (*L*) vom Textumfang (*N*).

In der quantitativen Linguistik gibt es zur Zeit zwei grundlegende Ansätze zur Messung der lexikalischen Übereinstimmung (Nähe) zweier Texte. Im ersten Fall mißt man die Nähe der Texte nach dem Grad der Übereinstimmung des Vokabulars, d.h. ohne Rücksicht auf die Häufigkeit der Wörter im Text. Dafür benutzt man verschiedene Indizes der lexikalischen Übereinstimmung. Zu diesem Zweck kann man beispielsweise die Formeln (4.5), (4.6) und (4.7) benutzen, die wir zum Vergleich von Wörterbüchern aus verschiedenen Epochen der Sprache verwendet haben (vgl. Abschnitt 4.2). Eine andere Methode beruht auf der Korrelation der Häufigkeiten in den verglichenen Texten. Dazu benutzt man verschiedene Varianten der Korrelationsanalyse (z.B. Kļaviņa, 1977; Marusenko, 1981). Eine besondere Variante des Textvergleiches unter Einbeziehung von Häufigkeiten ist die sogenannte distributionell-statistische Methode, wobei man das gleichzeitige Vorkommen von Wörtern in Fragmenten bestimmter Länge berücksichtigt (Šajkevič, 1968, 1982). In der vorliegenden Arbeit werden wir eine weniger bekannte Art der Messung der lexikalischen Nähe von Texten unter Berücksichtigung der Häufigkeit erörtern, nämlich die Methode der Vereinigung der Vokabulare.

Beim Vergleich der lexikalischen Zusammensetzung von Texten kann man, statt einfache Übereinstimmung des Vokabulars zu fordern, von der probabilistischen Verteilung der Wörter auf die Weise ausgehen, daß man eine bestimmte Grundgesamtheit zufällig in zwei Teile trennt. Daraus kann man schließen, daß man beim Vergleich zweier Texte das Experiment durch Zusammensetzung der verglichenen Texte durchführt (s. Muller, 1968; Tuldava, 1971; Darčuk, 1975). Man setzt voraus, daß im Falle der Homogenität der Lexik zweier Texte (A und B) die Verteilung der Teilhäufigkeiten im vereinigten Text nach der Formel $(p + q)^F$ verläuft, wobei F die Häufigkeit des Wortes im vereinigten Text ist, p und q die Wahrscheinlichkeiten dafür sind, daß ein zufällig gewähltes Wort aus der vereinigten Stichprobe zum Text A bzw. B gehört. Daraus kann man die theoretischen Wahrscheinlichkeiten der Teilhäufigkeiten der Wörter in den Texten A und B bei ihrer Vereinigung berechnen (s. Tabelle 4.10).

Im günstigsten Fall, wenn die Umfänge der verglichenen Texte gleich sind ($N_A = N_B$) und folglich $p = q = 1/2$, nehmen die Wahrscheinlichkeiten der Teilhäufigkeiten in beiden Texten konkrete Werte an (Tabelle 4.11).

In Tabelle 4.11 sieht man, daß sich Wörter mit der Häufigkeit 1 (in dem vereinigten Text) auf die Texte A und B im probabilistischen Modell gleichmäßig verteilen, d.h. 50% fehlen in einem Text ($f = 0$) und genau so viele Wörter erscheinen 1 mal in dem anderen Text ($f = 1$). Wörter, die im vereinigten Text die Häufigkeit 2 ($F = 2$) haben, werden folgendermaßen verteilt: 1/4 oder 25% erscheint zweimal in einem Text, z.B. im Text A, und genau so viel fehlt im Text A, d.h. erscheint zweimal nur im Text B; die restlichen 50% erscheinen jeweils einmal

im ersten und einmal im zweiten Text. Auf diese Weise deutet man auch die restlichen Zeilen der Tabelle. Zu bemerken ist, daß die äußeren Zahlen in jeder Zeile immer der Wahrscheinlichkeit entsprechen, daß Wörter mit der gegebenen Häufigkeit nur in einem der Texte vorkommen.

Tabelle 4.10
Verteilung der Häufigkeiten im vereinigten Text nach der Formel $(p + q)^F$

p \ f	0	1	2	3	4	...	$(p + q)^F$
1	p	q					$(p + q)^1$
2	p^2	$2pq$	q^2				$(p + q)^2$
3	p^3	$3p^2q$	$3pq^2$	q^3			$(p + q)^3$
4	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4		$(p + q)^4$
...	.....						.....

p und q sind die Wahrscheinlichkeiten, daß das Wort zum Text A bzw. B gehört; F ist die Häufigkeit des Wortes im vereinigten Text, f ist die Teilhäufigkeit.

Tabelle 4.11
Verteilung der Häufigkeiten im vereinigten Text im Falle, daß $p = q = 1/2$.

p \ f	0	1	2	3	4	.....
1	1/2	1/2				
2	1/4	2/4	1/4			
3	1/8	3/8	3/8	1/8		
4	1/16	4/16	6/16	4/16	1/16	
...	.....					

Um den Vergleich zu illustrieren, haben wir zwei kurze Werke aus der künstlerischen Literatur, nämlich "Die achte Verletzung" von K. Simonov und "Wir, sowjetische Menschen" von B. Polevoj verwendet. Aus der Autorensprache beider Erzählungen wurden Stichproben von jeweils 1000 laufenden Wörtern erhoben. Die Analyse zeigte, daß in der Stichprobe aus Simonov (Text A) unter 1000 Wortverwendungen 443 unterschiedliche Wörter gefunden wurden und in der Stichprobe aus Polevoj (Text B) 524. In dem vereinigten Text mit $N_A + N_B = N_{AB} = 2000$ ist der Vokabularumfang $L_{AB} = 841$ (vgl. die Häufigkeitsspektren der einzelnen Texte und des vereinigten Textes in Tabelle 4.12).

Der nächste Schritt des Experiments besteht in der Aufstellung des theoreti-

schen und des empirischen Modells der Verteilung der Teilhäufigkeiten und im Vergleich beider Modelle. Die theoretischen Häufigkeiten berechnet man mit Hilfe der Häufigkeitsverteilung der Wörter, d.h. des Häufigkeitsspektrums des vereinigten Textes, wobei man als Grundlage die theoretische Wahrscheinlichkeitsverteilung der Teilhäufigkeiten in Abhängigkeit von der Größe von p und q heranzieht

Tabelle 4.12
Häufigkeitsspektren individueller Texte und des vereinigten Textes

F	Text A $N_A = 1000$	Text B $N_B = 1000$	Text $A + B$ $N_{AB} = 2000$
	m_F	m_F	m_F
1	297	392	585
2	73	68	116
3	25	26	56
4	11	12	26
5	7	6	10
6	9	3	6
7	4	2	6
8	4	0	7
9	1	2	3
>10	12	13	26
	443	524	841

m_F = Anzahl der Wörter mit der Häufigkeit F

(im gegebenen Fall ist $p = q = 1/2$; s. Tabelle 4.11). Die Berechnung ist in Tabelle 4.13 dargestellt. Die Verteilung der Teilhäufigkeiten im empirischen Modell berechnet man direkt aus der zusammengesetzten Wortliste beider Texte, z.B. nach dem Schema:

	A	B	$A+B$
a	5	4	9
avtomat	-	1	1
artillerist	2	-	2
atakovat'	1	1	2
bol'šoj	1	3	4
usw.			

Das empirische Modell für den Text A ist in Tabelle 4.14 dargestellt (analog erhält man die Verteilung im Text B, wenn man die Zeilen in der Tabelle umdreht).

Der Vergleich theoretischer und empirischer Modelle für die Verteilung von Teilhäufigkeiten im vereinigten Text zeigt einige interessante Unterschiede, besonders bei der Verteilung seltener Wörter. Man sieht, z.B. in Tabelle 4.14, daß Wörter mit der Häufigkeit 1 folgendermaßen verteilt sind: Im Text A 248 und im Text B 337 (entsprechend die Teilhäufigkeiten 0 und 1). Der Vergleich mit den entsprechenden theoretischen Werten (Tabelle 4.13) zeigt, daß im Text A 44,5 Einheiten fehlen,

Tabelle 4.13
Verteilung der theoretischen Teilhäufigkeiten

F	m_F	Teilhäufigkeiten									
		0	1	2	3	4	5	6	7	8	...
1	585	292,5	292,5								
2	116	29	58	29							
3	56	7	21	21	7						
4	26	1,5	6,5	10	6,5	1,5					
5	10	0,5	1,5	3	3	1,5	0,5				
6	6	0	0,5	1,5	2	1,5	0,5	0			
7	6	0	0,5	1	1,5	1,5	1	0,5	0		
8	7	0	0	1	1,5	2	1,5	1	0	0	
...	...	...	...	...	...	...	...	...	...	...	

Tabelle 4.14
Verteilung der empirischen Teilhäufigkeiten

F	m_F	Teilhäufigkeiten in Text A									
		0	1	2	3	4	5	6	7	8	...
1	585	337	248								
2	116	42	30	44							
3	56	14	12	14	16						
4	26	3	4	8	6	5					
5	10	0	1	3	1	4	1				
6	6	1	1	1	0	1	0	2			
7	6	0	1	1	2	0	1	1	0		
8	7	0	0	1	0	1	2	2	0	1	
...	...	...	...	...	...	...	...	...	...	...	

während im Text *B* genau diese Zahl überschüssig ist:

	Text *A*	Text *B*
Empirisch (E)	248	337
Theoretisch (T)	292,5	292,5
E - T	-44,5	+44,5

Für Wörter mit der Häufigkeit 2 im vereinigten Text führen wir eine unvollständige Liste konkreter Wörter in Tabelle 4.15 auf. Aus dieser Liste sieht man, daß sich unter den gemeinsamen Wörtern, die jeweils einmal in jedem Text vorkommen, Wörter wie *sestra* (Schwester), *zemlja* (Erde), *noč* (Nacht), *solnce* (Sonne) und Adjektive wie *medicinskij* (medizinisch) und *ranennyj* (verletzt) finden. Was die Wörter betrifft, die nur in einem der Texte vorkommen (jeweils zweimal), so kann man aufgrund ihrer geringen Häufigkeit nicht feststellen, welche für den gegebenen Text spezifisch sind.

In Tabelle 4.16 bringen wir die volle Liste der Wörter mit der Häufigkeit 3 im vereinigten Text. Unter den gemeinsamen Wörtern findet man solche wie *vojna* (Krieg), *front* (Front), *vrač* (Arzt), *nosilki* (Bahre), *nemeckij* (deutsch), was von thematischer Nähe zeugt. In der Tat betreffen beide verglichenen Werke Ereignisse der Kriegszeit. Unter den Wörtern, die nur jeweils in einem Text vorkommen, kann man offensichtlich zwei Schichten ausmachen, die jeweils für den gegebenen Text spezifisch sind, und solche, die gemäß unserem theoretischen Modell zufällig in nur einem Text erschienen sind. Ihr zahlenmäßiges Verhältnis kann man aufgrund der empirischen und der theoretischen Größen approximativ berechnen, z.B. sind im Text *A* aus der Gesamtzahl der Wörter mit der Häufigkeit 3 (E = 16) theoretisch 7 Wörter "zulässig" (T = 7), während E - T = 9 nicht als zufällige Wörter zu betrachten sind (für weitere Einsichten ist eine qualitative Analyse nötig).

Ebenso kann man die konkreten Verteilungen von Wörtern mit den Häufigkeiten 4, 5 usw. im vereinigten Text analysieren (s. Tuldava 1971).

Zur Berechnung der statistischen Signifikanz des Unterschieds zwischen den empirischen und theoretischen Modellen und gleichzeitig zur Messung der *Nähe* der lexikalischen Struktur beider Texte kann man den Ähnlichkeitskoeffizienten (*K*) nach der Formel

$$K = 1 - \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad (4.14)$$

benutzen, wobei *n* die Zahl der Beobachtungen ist.

Die Berechnung des Chi-Quadrats ergibt 67,1, was den kritischen Wert auf der 0,1%-Ebene signifikant überschreitet, d.h. es gibt einen statistisch signifikanten

Unterschied zwischen den empirischen und theoretischen Modellen der Teilhäufigkeiten in Tabelle 4.13 und 4.14. Der Ähnlichkeitskoeffizient ist

$$K = 1 - \sqrt{\frac{67,1}{841 + 67,1}} = 0,728.$$

Beim Vergleich einer großen Anzahl von Texten untereinander kann man Paare oder Gruppen von Texten bestimmen, die sich voneinander in bezug auf die Worthäufigkeiten im Text durch größere oder kleinere Nähe unterscheiden.

Tabelle 4.15
Verteilung der Wörter mit der Häufigkeit *F* = 2 im vereinigten Text

	A 2	*A* 1 *B* 1	*B* 2
	artillerist gimnasterka kontuzija kostyl' majatnik naveščat' nagradit' poduška pensionnyj sneg samoljubie samoljubivyj tank šinel' ujti u.a.	blednyj zemlja medicinskij ranennyj sestra privyknut' pokazat'sja pogovorit' popytat'sja rasstat'sja slyšat' solnce noč' načalo molčat' u.a.	baryšnja izjaščnyj ljubov' nerv omerzenie palač podpol'nyj rodina sapožnik smert' tjur'ma utešat' fašistskij chatka znat' u.a.
E T	44 29	30 58	42 29
E-T	+15	-28	+13

Clusteranalyse

Die Clusteranalyse kann man als eine Gesamtheit von Methoden definieren, die man zur Zerlegung einer Menge von Objekten in Gruppen, oder *Cluster* benutzen kann, die in einem bestimmten Sinne die einander ähnlichsten Elemente enthalten. Die Methoden der Clusteranalyse gehören zu einer Gruppe von Prozeduren, die

Tabelle 4.16
Verteilung der Wörter mit der Häufigkeit $F = 3$ im vereinigten Text

	A 3	*A* *B* 2 1	*A* *B* 1 2	*B* 3
	batareja vyjti govorit' davno esli karman ni ničego ostal'noj pokazyvat' predstavit' polk ranit' sejčas časy Čujko	bez vojna vmeste god dolgo žizn' žit' kazat'sja posle rasskazat' ruka sovsem tovarišč jasnyj	vzgljad vrač žena idti nemeckij otec rabota slovo takoj tut front nosilki	volja general dat' echat' zadanie inoj kursy letčik naš roditeli sam sovetskij šef ėvakuacija
E T	16 7	14 21	12 21	14 7
E-T	+9	-7 -16	-9	+7

man allgemein als Methoden der Mustererkennung bezeichnet, und im engeren Sinne gehören sie zu den Methoden der Klassifikation multidimensionaler Beobachtungen. Die Besonderheit der Klassifikation multidimensionaler Beobachtungen besteht darin, daß man jedes Objekt anhand der Menge der an ihm beobachteten Merkmale beschreibt und für die Aufstellung der Klassifikation die gegebene Menge der Merkmale und ihrer Zusammenhänge benutzt.

Es gibt eine Reihe von Spielarten der Clusteranalyse. Sie alle haben drei grundlegende Voraussetzungen gemeinsam, die zur Durchführung der Analyse notwendig sind: die multidimensionalen Ausgangsdaten,. Angaben zu Ähnlichkeit oder Nähe und Angaben über Cluster (vgl. Ryzin, 1977). Dementsprechend kann man drei Phasen der Untersuchung unterscheiden: In der ersten, vorbereitenden Phase ordnet man die Ausgangsdaten, in der zweiten Phase mißt man die Nähe (Ähnlichkeit oder Differenz) zwischen den zu klassifizierenden Objekten und in der dritten Phase konstruiert man das Clustersystem, das Objekte auf verschiedenen Ähnlichkeitsebenen zusammenfaßt. Die letzten zwei Phasen führt man üblicherweise mit Hilfe automatischer Methoden auf dem Rechner durch. Das Ergebnis der Clusteranalyse besteht in einer Zerlegung, die die aufgestellten Kriterien erfüllt.

Es muß betont werden, daß die Clusteranalyse und sämtliche anderen Klassifikationsmethoden subjektiv und relativ in dem Sinne sind, daß die Resultate der Analyse völlig durch die zugrundegelegten Merkmale determiniert sind. Klassifikationen, die auf vielen und unterschiedlichen Merkmalen begründet sind, sind effektiver für die Aufdeckung einer "natürlichen" Ordnung der Objekte und Erscheinungen (wenn es möglich ist, die gesamte verfügbare Information über die Merkmale der zu klassifizierenden Objekte auszunutzen). In anderen Fällen, wenn der Forscher nur an einigen Eigenschaften der Objekte interessiert ist oder wenn die Clusteranalyse nur einem bestimmten praktischen Zweck dient, kann man sich mit einer kleinen Anzahl von ausgewählten Merkmalen begnügen. In der vorliegenden Arbeit stellen wir uns gerade eine derartige begrenzte Aufgabe, nämlich die Möglichkeiten der Textklassifikation mit Hilfe der Clusteranalyse aufgrund einiger in der Praxis der quantitativen Linguistik bekannter formaler Charakteristika der lexikalisch-statistischen Textstruktur. Dabei stellt sich die Frage nach der Übereinstimmung der Resultate unterschiedlicher Prozeduren, die zwar am gleichen Material, aber mit unterschiedlichen Mengen von Merkmalen durchgeführt werden.[10]

Die Aufgabe der Klassifikation von Texten, darunter künstlerischen, ergibt sich bei texttypologischen Untersuchungen (für stilistische, pädagogische u.a. Zwecke) und bei Problemen in den Bereichen Informatik, Textindexierung, Autorenbestimmung usw.

In der vorliegenden Arbeit werden 20 Texte aus der modernen estnischen künstlerischen Prosa, bestehend aus jeweils 5000 laufenden Wörtern der Autorensprache, klassifiziert (vgl. Tuldava 1981). Man nimmt an, daß eine Stichprobe von 5000 laufenden Wörtern, unterteilt in 5 Teile von jeweils 1000 laufenden Wörtern, hinreichend ist, um einige fundamentale formale Merkmale der uns interessierenden statistischen Organisation der Texte (in einer vergleichenden Analyse bei gleichen Textumfängen) zu offenbaren. Ausgehend von diesen 20 Texten wurden drei Experimente aufgrund von unterschiedlichen Messungen von quantitativ-linguistischen Textcharakteristika durchgeführt. Die Charakteristika waren wie folgt:

- Textabdeckung durch Wortformen (Fragestellung 1);
- Häufigkeitsspektrum (Fragestellung 2);
- Dynamik des Vokabularwachstums (Fragestellung 3).

Die konkreten Ausgangsdaten findet man in den Tabellen 4.17, 4.18 und 4.19.

Es ist festzuhalten, daß alle oben beschriebenen Charakteristika als eng miteinander zusammenhängende quantitative Kenngrößen der statistischen Textstruktur betrachtet werden (s. z.B. Orlov 1982). Die Frage ist nur, ob unser Experiment in drei verschiedenen Versuchen der Klassifikation realer Texte ähnliche Resultate

[10] In dem Experiment wurde das Programm für die Clusteranalyse von Äärema (1978) verwendet.

aufweisen wird, wenn wir den angenommenen Zusammenhang der Kenngrößen berücksichtigen.

Die mathematische Grundlage für die Klassifikation von Objekten mit Hilfe der Clusteranalyse bildet die Berechnung von Funktionen für Objektpaare aufgrund von numerischen Merkmalswerten. Als Resultat erhält man im allgemeinen eine *Ähnlichkeitsmatrix* (Proximitäts- oder Distanzmatrix) der Objekte. In solchen Matrizen werden $n(n\text{-}1)/2$ Ähnlichkeitswerte für alle Paare von n Objekten, die man klassifizieren will, dargestellt.

Die Aufgabe der Clusteranalyse kann man mit Hilfe von Proximitäts- oder Distanzmatrizen lösen. Proximitätsmatrizen werden gewöhnlich aufgrund von Ähnlichkeits- oder Assoziationskoeffizienten aufgestellt. Distanzmatrizen werden aufgrund von "Abstandskoeffizienten" konstruiert. Die Wahl der Metrik für die Messung des Abstands wird durch die Natur der Ausgangsmerkmale und die Ziele der Klassifikation bestimmt.

In der vorliegenden Untersuchung wurde als Ähnlichkeitsmaß die bekannte Euklidische Distanz aufgrund von folgenden Überlegungen gewählt: Bei gegebener Merkmalsauswahl und bei gleichem Textumfang kann man alle Merkmalswerte (d.h. die einzelnen Elemente des Vektors) als gleichwertig betrachten, und die numerischen Unterschiede zwischen den verschiedenen Werten der Merkmale der verglichenen Texte kann man für die Bestimmung des Abstands zwischen den Texten als wesentlich betrachten. Um jedoch das allzu große Gewicht der großen Werte von Merkmalen im Vergleich mit den kleinen zu relativieren, ist es empfehlenswert, die Mutungsintervalle der Merkmalswerte von Ausgangsdaten zu normieren (üblicherweise so, daß man den Mittelwert subtrahiert und durch die Standardabweichung dividiert, damit man eine Einheitsdispersion bekommt). Die Euklidische Distanz ist gegeben durch

$$d(X_s,X_t) = \left[\sum_{j=1}^{k} (x_{js} - x_{jt})^2\right]^{0.5}, \tag{4.15}$$

wobei X_{js} und X_{jt} die normierten Werte der Merkmale sind, k die Zahl der Messungen. Der Wert $d(X_s,X_t)$ für die Vektoren X_s und X_t ist dem Abstand zwischen den Objekten (Texten) T_s und T_t bezüglich der gewählten Menge von Merkmalen äquivalent. Es wird vorausgesetzt, daß die Ähnlichkeit zwischen den Texten von einer Ähnlichkeit zwischen den Stilen der Autoren hinsichtlich einiger latenter Eigentümlichkeiten zeugt, die sich in stabilen quantitativen (linguostatistischen) Charakteristika der Texte manifestieren.

Bei der Konstruktion des Clustersystems geht man von den Angaben über die Ähnlichkeiten zwischen den Objekten aus, d.h. man betrachtet, bildhaft gesprochen, in den Algorithmen für die Clusterbildung die Ähnlichkeitsmatrix als Eingabe

und die Zerlegung auf Cluster als Ausgabe. Die Clusterbildungsmethoden kann man in hierarchische und nichthierarchische aufteilen. Es gibt zwei Typen von hierarchischen Clusterungsmethoden: Die agglomerative und die divisive (trennende) Methode. Das Prinzip der agglomerativen Algorithmen besteht in der schrittweisen Vereinigung der Objekte von den ähnlichsten bis hin zu den unähnlichsten im Cluster. In divisiven hierarchischen Prozeduren hingegen wird die Menge der Objekte schrittweise in Gruppen zerlegt. In der vorliegenden Arbeit wurde eine Variante der agglomerativen hierarchischen Prozedur gewählt. Für die praktische Ausführung mittels EDV wurde die B_k-Methode benutzt, die eine verbesserteVariante des sogenannten Cambridge-Algorithmus darstellt (s. Jardine, Sibson 1968). Bei Verwendung der B_k-Methode kann man im allgemeinen von einer k-Clusterbildung sprechen, in der der Parameter k die zulässige Überdeckung der Cluster bestimmt.

Tabelle 4.17
Ausgangsdaten: Abdeckung des Textes mit Wortformen (%)

Text Nr.	Ränge der Wortformen (i)								
	1	10	50	100	500	1000	1500	2000	2500
1	2,7	11,0	21,4	28,3	50,6	62,6	72,6	82,6	92,6
2	2,6	12,0	23,3	30,4	52,4	642	74,2	84,1	94,1
3	2,8	13,8	26,7	33,7	57,5	70,7	80,7	90,7	100,0
4	3,5	12,9	24,4	31,2	53,6	65,2	75,2	85,2	95,2
5	5,7	13,8	25,0	31,3	52,2	62,7	72,7	82,7	92,8
6	2,4	14,3	27,0	34-7	57,6	69,7	79,7	89,8	99,8
7	3,4	14,4	25,5	32,7	54,4	66,1	76,1	86,1	96,1
8	3,4	15,3	27,9	35,7	58,9	71,1	81,1	91,2	100,0
9	3,1	10,7	20,3	27,2	50,6	62,1	72,2	82,3	92,4
10	1,6	10,1	20,2	27,1	47,2	58,5	68,5	78,6	88,6
11	3,3	15,4	27,0	33,9	55,5	66,8	76,7	86,6	96,5
12	2,6	14,3	27,2	35,2	57,6	69,0	79,0	89,0	99,0
13	3,7	14,1	27,2	34,8	56,5	68,1	78,1	88,2	98,3
14	3,4	13,8	25,4	32,3	55,2	66,6	76,6	86,6	96,6
15	2,5	13,1	25,4	32,1	54,5	66,1	76,2	86,4	96,6
16	2,9	13,1	25,9	33,4	57,1	70,1	80,1	90,2	100,0
17	3,6	16,7	30,2	38,0	61,1	72,5	82,5	92,5	100,0
18	3,7	13,7	27,4	35,1	58,2	70,2	80,2	90,3	100,0
19	3,5	11,5	21,4	27,6	48,9	59,9	69,7	79,4	89,2
20	3,9	11,9	22,6	28,8	50,0	61,4	71,4	81,4	91,4

Hat man n zu klassifizierende Objekte, dann kann der Parameter k ganzzahlige Werte im Intervall <1,n-2> annehmen. Bei $k = 1$, d.h. bei 1-Clusterbildung, die mit der "Methode des nächsten Nachbarn" übereinstimmt, erhält man sich nicht überschneidende Cluster, die man dann in Form eines Dendrogramms darstellen kann. Bei $k > 1$ ist dies nicht mehr möglich. Hier benutzen wir die 1-Clusterbildung. Ein wichtiger Faktor der Clusteranalyse ist das Niveau der Klassifikation, das man mit h bezeichnet.

Wie bereits festgestellt wurde, werden die Objekte bei der agglomerativen hierarchischen Methode schrittweise in Cluster zerlegt. Der Clusterungsprozeß beginnt so, daß man im ersten Schritt die zwei ähnlichsten Objekte (im ersten Experiment Texte Nr. 6 und 12) vereinigt und als ein Cluster betrachtet. Dadurch verringert sich die Zahl der Objekte auf n-1, ein Cluster enthält zwei Objekte und die restlichen n-2 jeweils eins. Den Prozeß kann man solange wiederholen, bis alle

Tabelle 4.18
Häufigkeitsspektrum - Anteil der Wortformen (%) mit der gegebenen Häufigkeit

Text Nr.	Häufigkeit der Wortformen (F)											
	1	2	3	4	5	6	7	8	9	10	11-20	>20
1	79,12	10,32	4,11	1,95	0,98	0,80	0,66	0,21	0,38	0,21	0,77	0,49
2	78,70	10,65	4,22	1,93	0,96	0,64	0,43	0,39	0,25	0,22	1,04	0,57
3	73,29	13,60	4,22	3,21	1,38	0,89	0,81	0,37	0,61	0,04	0,65	0,93
4	78,74	10,45	3,94	2,01	1,24	0,91	0,37	0,33	0,29	0,18	0,99	0,55
5	81,65	9,40	2,83	2,17	0,94	0,66	0,28	0,25	0,28	0,11	0,91	0,52
6	76,00	12,16	4,38	1,83	1,04	0,80	0,88	0,44	0,32	0,36	1,03	0,76
7	78,28	10,97	4,30	1,74	1,08	0,63	0,30	0,59	0,41	0,22	0,96	0,52
8	75,20	12,46	4,39	2,05	1,23	0,94	0,53	0,53	0,33	0,33	1,31	0,70
9	80,18	10,05	3,72	1,32	1,11	0,80	0,73	0,52	0,25	0,17	0,63	0,52
10	81,74	10,08	3,36	1,47	0,65	0,52	0,39	0,23	0,13	0,29	0,72	0,42
11	78,61	10,85	3,96	1,91	0,82	0,71	0,75	0,41	0,22	0,19	0,97	0,60
12	77,70	11,25	3,96	1,53	1,10	0,74	0,47	0,59	0,59	0,19	1,25	0,63
13	77,80	11,29	4,60	1,24	1,08	0,74	0,50	0,23	0,27	0,43	1,12	0,70
14	78,57	9,89	4,19	2,21	1,31	0,94	0,49	0,34	0,30	0,22	0,94	0,60
15	78,64	10,49	3,90	2,17	1,27	0,90	0,38	0,22	0,19	0,19	0,94	0,71
16	74,07	12,81	5,18	2,09	1,25	0,96	0,80	0,52	0,32	0,12	1,08	0,80
17	75,73	11,38	5,10	1,98	1,14	0,84	0,59	0,42	0,46	0,34	1,22	0,80
18	76,12	11,44	4,43	2,42	1,41	0,64	0,56	0,32	0,48	0,24	1,13	0,81
19	81,56	9,01	3,54	1,80	1,08	0,62	0,59	0,46	0,16	0,07	0,72	0,39
20	80,60	10,33	3,14	1,77	1,19	0,68	0,34	0,38	0,14	0,24	0,68	0,51

Objekte in einem großen Cluster vereinigt wurden. Die Resultate dieses Prozesses stellt man graphisch als Dendrogramm und mit Hilfe einzelner Tabellen mit den Resultaten der Clusterbildung nach jedem Schritt dar (die Dendrogramme und Tabellen werden von Rechner ausgegeben). Das Dendrogramm gibt die Möglichkeit einer anschaulichen Interpretation des gesamten Vorgangs der Clusterbildung (vgl. für unsere Daten die Abb. 4.9 - 4.11). Für das gegebene Experiment mit $n = 20$ Objekten endet die Prozedur im 19-ten Schritt, wenn alle Objekte (Texte) in einem Cluster vereinigt sind.

Wir halten fest, daß bei den drei Fragestellungen, die wir an denselben 20 Texten untersucht haben, Merkmale gewählt wurden (Abdeckung des Textes, Häufigkeitsspektrum, Dynamik des Vokabularwachstums), die man allgemein als miteinander verknüpfte und nahe verwandte Charakteristika der statistischen Textstruktur

Tabelle 4.19
Dynamik des Vokabularwachstums
(Zahl der unterschiedlichen Wortformen bei unterschiedlichen Textumfängen)

Text Nr.	Autor	Textumfang (N)				
		1000	2000	3000	4000	5000
1	E. Beekman	731	1383	1865	2404	2869
2	V. Gross	677	1315	1859	2358	2791
3	A. Hint	649	1116	1597	2034	2463
4	H. Kiik	710	1351	1828	2315	2738
5	J. Kross	723	1315	1914	2382	2861
6	P. Kuusberg	645	1166	1674	2075	2508
7	L. Promet	674	1212	1720	2207	2694
8	V. Saar	633	1128	1700	2045	2439
9	H. Sergo	734	1326	1885	2416	2876
10	R. Sirge	764	1397	2017	2572	3067
11	M. Traat	689	1235	1722	2226	2656
12	E. Vetemaa	680	1208	1734	2162	2552
13	A. Kaal	651	1204	1690	2119	2586
14	T. Kallas	663	1226	1733	2179	2668
15	J. Peegel	690	1224	1700	2223	2669
16	J. Tuulik	624	1135	1560	2005	2491
17	A. Valton	588	1036	1468	1955	2373
18	M. Unt	658	1176	1678	2133	2483
19	E.Niit/J.Kross	740	1357	1923	2473	2983
20	J. Smuul	732	1361	1917	2473	2929

betrachtet. Folglich konnte man auch ähnliche Clusterbildungsresultate in den drei Versuchen erwarten. Vergleicht man die entsprechenden Dendrogramme der schrittweisen Clusterbildung (vgl. Abb. 4.9 - 4.11), dann findet man auf den ersten Blick in verschiedenen Punkten eine Ähnlichkeit: Beispielsweise im ersten und dritten Experiment werden Texte 1 und 9 sowie Texte 7 und 14 in einem Cluster in einer frühen Phase der Clusterbildung verbunden (im 4. und 1. bzw. im 2. und 4. Schritt). Aber im allgemeinen kann man feststellen, daß die Dendrogramme sich wenig ähneln. Daher werden wir diejenigen Phasen der Clusterbildung vergleichen, die aufgrund bestimmter Kriterien als "optimal" betrachtet werden können. Im vorliegenden Fall können die anvisierten (optimalen) Zerlegungsphasen empirisch aufgrund von Datenvergleich und unter Berücksichtigung der Schätzungen der einzelnen und mittleren "Stabilität" der Cluster in verschiedenen Zerlegungsphasen bestimmt werden (solche Schätzungen werden von dem verwendeten Programm automatisch ausgegeben). Davon ausgehend kann man unter den drei Fragestellungen folgende Clusterbildungen bei aufstellen:

	Fragestellung 1	Fragestellung 2	Fragestellung 3
A:	(4.7.11.14.15.)	(1.2.4.7.11.12.14.15.)	(7.11.14.15.)
B:	(8.13.18.)	(6.8.17.18.)	(6.8.13.18.)
C:	(1.9.)	(5.9.19.20.)	(1.5.9.19.20.)
D:	(3.6.12.16.)	-	-

Isolierte Texte (1-Element-Cluster):

(2.)(5.)(10.) (17.)(19.)(20.)	(3.)(10.)(13.)(16.)	(2.)(3.)(4.) (10.)(12.)(16.)(17.)

Bei einer vergleichenden Analyse stellt man fest, daß es sowohl bei der Bildung als auch in der Zusammensetzung der Cluster einige gemeinsame Momente gibt.

Das Cluster, das wir als "A" bezeichnen, erscheint in ähnlichen Varianten bei allen drei Fragestellungen. Den festen Kern des Clusters A bilden die Texte 7, 11, 14, 15, zu denen sich Text 4 gesellt, den man im Cluster A bei den ersten zwei Fragestellungen findet.

Im Cluster B sind die Texte 8 und 18 für alle Fragestellungen gemeinsam, zu denen sich Texte 6 und 13 gesellen.

Im Cluster C findet man bei der zweiten und dritten Fragestellung die Texte 5, 9, 19, 20; bei der ersten bleiben die Texte 5, 19, 20 in der gewählten Klassifikationsphase isoliert, d.h. sie bilden 1-Element Cluster, man kann aber feststellen, daß die Texte 19 und 20 sich nach einigen Schritten in einem Cluster vereinigen (vgl. Abb. 4.9). Text 5 bleibt aber gleichzeitig isoliert bis zum letzten Schritt der Clusterbildung.

Cluster D kommt nur bei der ersten Fragestellung zustande, und zu ihm gehören die Texte 3, 6, 12, 16. Dieses Cluster bildet sich aber in einer frühen Phase, nämlich im fünften Schritt, und bleibt bis zum 11. Schritt unverändert, was von der großen Stabilität des Clusters zeugt.

Neben den Clustern mit vielen Elementen sind auch solche mit einem Element (in einer gewählten Phase der Klassifikation) interessant. In allen Experimenten bleibt Text 10 unverändert isoliert. Eine Tendenz zur Isolierung haben auch die Texte 2, 3 und 17, die in zwei von drei Fällen einelementige Cluster bilden.

Mit parallelen Experimenten zur Clusterbildung von 20 Texten aufgrund von unterschiedlichen formalen Charakteristika der statistischen Textstruktur gelang es also, einige hinreichend stabile, sich nicht überschneidende Cluster zu etablieren, die im gegebenen Grad die für die jeweilige Sprache (oder Fachsprache) charakteristische *Typen* von Texten bilden. Die verwendete Methode erlaubt es aber nicht, alle Texte zu typisieren: Im Durchschnitt fallen 30% der Texte nicht in die stabilen Vielelement- oder 1-Element Cluster. Dies kann teilweise durch den "Verkettungseffekt" der Cluster erklärt werden, den man durch Clusterbildung in einer frühen Phase überwinden kann. Die Hauptursache für die unvollständige Zerlegung der Texte in sich nicht überschneidende Cluster beruht aber darauf, daß "die Mehrheit der realen Klassen in dem Sinne vage ist, daß der Übergang von Zugehörigkeit zu Nichtzugehörigkeit zu diesen Clustern eher graduell als sprunghaft ist" (Zadeh, 1980). Daher wird es zweckmäßig sein, die Algorithmen der Clusteranalyse auf der Vorstellung von einem Cluster (Klasse, Typ) als einer vagen, unscharfen Menge aufzubauen (zum Versuch der Clusteranalyse mit sich teilweise überdeckenden Klassen s. Liiv, Tuldava, 1993).

Bei der vergleichenden Analyse der Resultate der drei parallelen Experimente waren beträchtliche Unterschiede in den hierarchischen Strukturen der Clustersysteme festzustellen (vgl. die entsprechenden Dendrogramme). Dieser Unterschied rührt vor allem daher, daß die gewählten Merkmale, die als miteinander verknüpft betrachtet worden waren, in der Tat keine genaue Übereinstimmung aufweisen, wie sie für exaktere Berechnungen notwendig ist. In realen Texten gibt es keine festen Verbindungen zwischen den verschiedenen Charakteristika der statistischen Textstruktur. Daraus folgt, daß die Clusteranalyse aufgrund einer beliebigen Menge von Merkmalen, die die statistische Organisation des Textes charakterisieren, Resultate aufgrund anderer analoger (verwandter, ähnlicher) Merkmale nicht prädestiniert, auch wenn es zwischen den Resultaten der Analysen einige (nicht voraussagbare) Übereinstimmungen gibt. In solchen Fällen kann man nur übereinstimmende oder ähnliche Resultate als hinreichend zuverlässig betrachten.

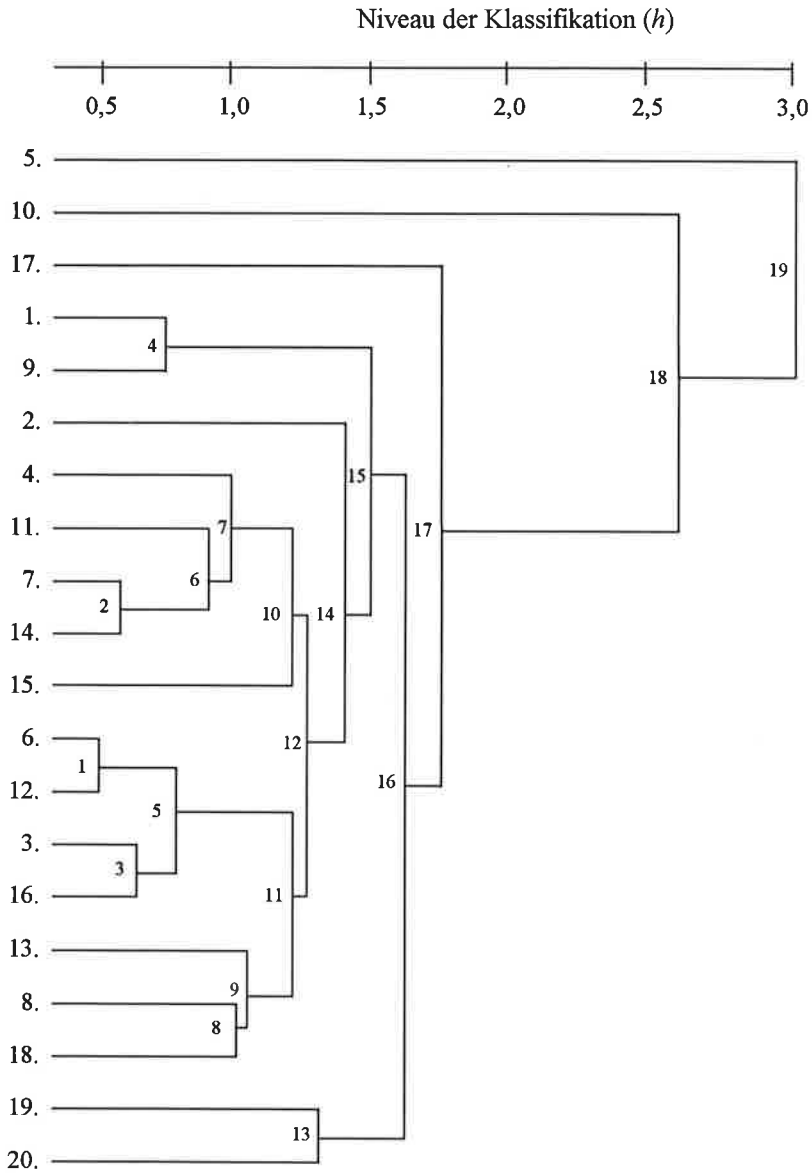


Abbildung 4.9. Dendrogramm der schrittweisen Clusterbildung von 20 Texten aufgrund von Kenngrößen der Textabdeckung mit Wortformen (Fragestellung 1). Die Zahlen links sind die Textnummern. Zahlen im Schema sind die Anzahlen von Schritten, die für die Vereinigung von Texten in Gruppen nötig sind.

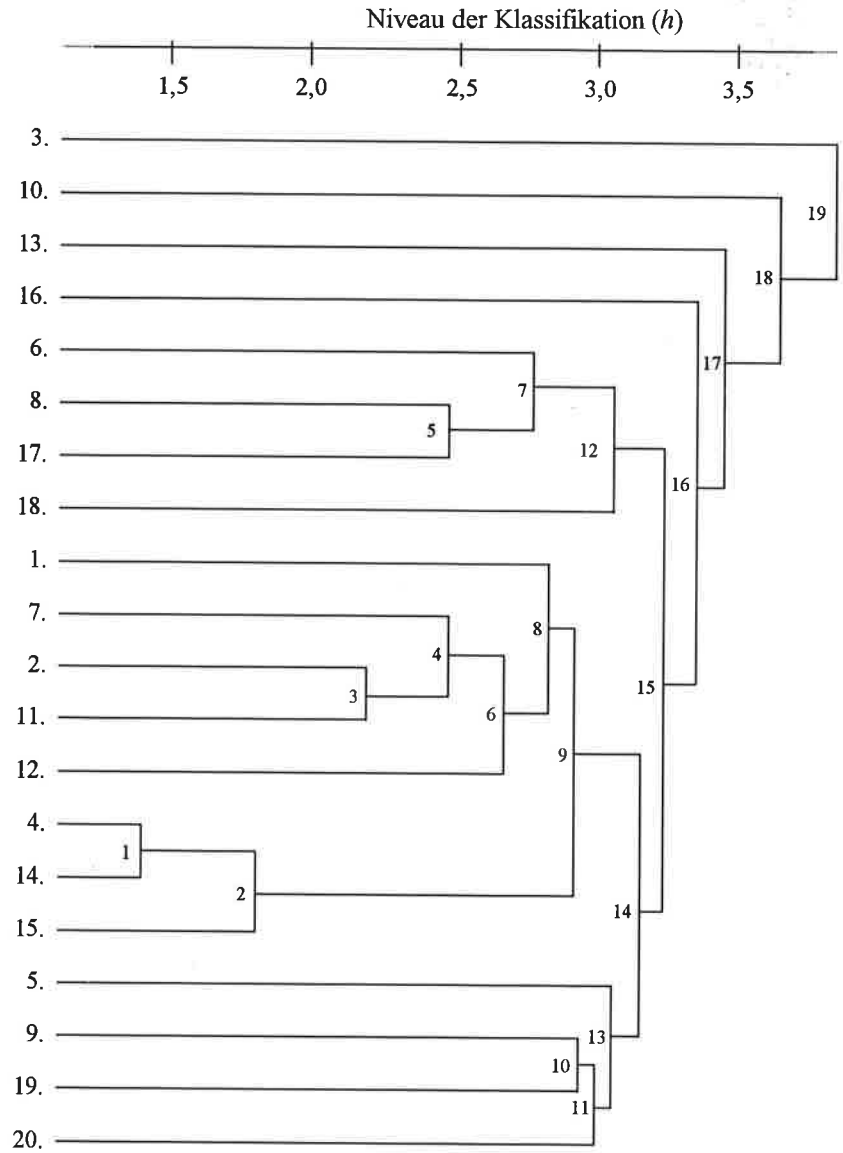


Abbildung 4.10. Dendrogramm der schrittweisen Clusterbildung von 20 Texten aufgrund von Häufigkeitsspektren auf der Vokabularebene (Fragestellung 2)

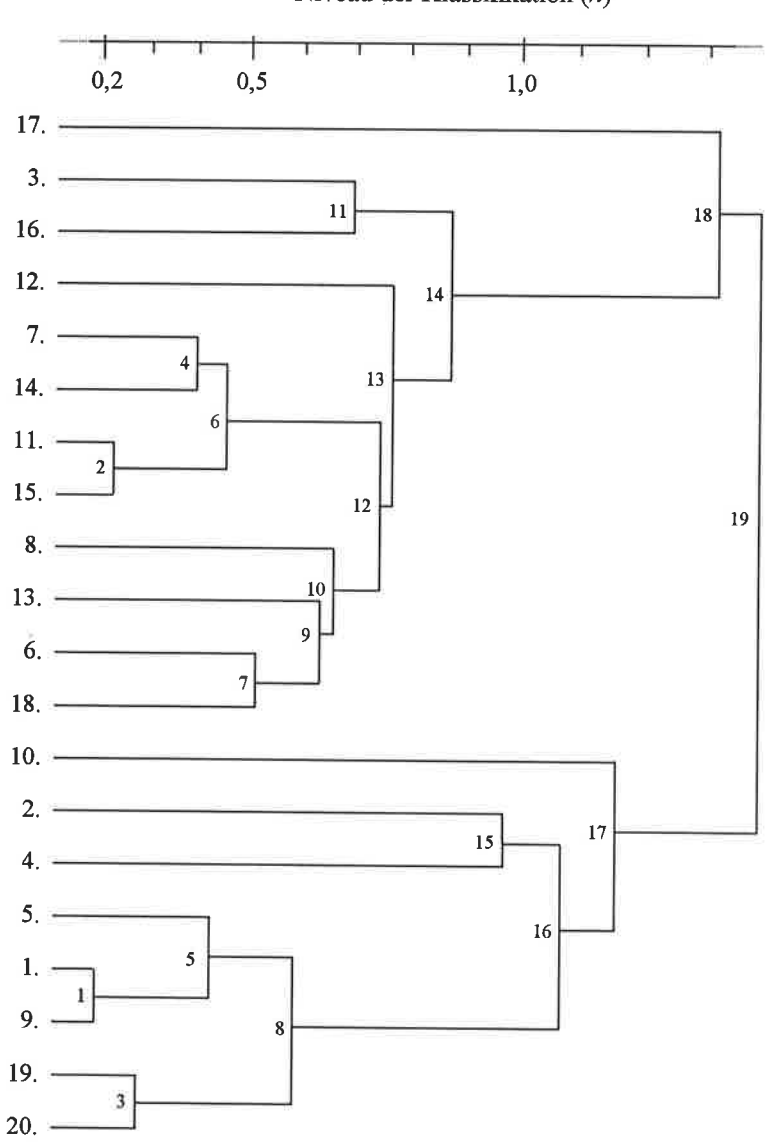


Abbildung 4.11. Dendrogramm der schrittweisen Clusterbildung von 20 Texten aufgrund der Dynamik des Vokabularwachstums mit zunehmendem Textumfang (Fragestellung 3)

Literatur

Ääremaa, R. (1978). Obščaja teorija konstruirovanija klaster-sistem i algoritmy dlja nachoždenija ich čislennych predstavlenij. *Trudy Vyčislitel'nogo centra (Tartu: TGU)* 42, 53-77.

Alekseev, P.M. (1968). Častotnye slovari i priemy ich sostavlenija. In: *Statistika reči:* 61-63. Moskva: Nauka.

Alekseev, P.M. (1977). *Kvantitativnaja tipologija teksta.* Leningrad: Diss.

Alekseev, P.M. (1978). O nelinejnych formulirovkach zakona Cipfa. *Voprosy kibernetiki* 41, *53-65.*

Alekseev, P.M. (1980). K osnovam statističeskoj leksikografii. In: *Problema slova i slovosočetanija:* 93-105. Leningrad: Leningradskij gosudarstvennyj pedagogičeskij institut.

Alekseev, P.M. (1981). O kvantitativnoj tipologii teksta. *Učenye zapiski TGU* 581, 3-13.

Alekseev, P.M. (1984). *Statistische Lexikographie.* Bochum: Brockmeyer.

Alekseev, P.M. (1986). Raspredelenie leksičeskich edinic po dline v tekste i slovare. *Učenye zapiski TGU* 745, 3-28.

Alimov, Ju.I. (1980). *Al'ternativa metodu matematičeskoj statistiki.* Moskva; Znanie.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika* 2, 1-10.

Altmann, G. (1980a). *Statistik für Linguisten.* Bochum: Brockmeyer.

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung:* 54-90. Göttingen: Herodot.

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics* 3, 1-7.

Andreev, N.D. (1967). *Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykoznanii.* Leningrad: Nauka.

Andrjuščenko, V.M. (1978). K voprosu ob ispol'zovanii koèffiicenta stabil'nosti v kačestve mery upotrebitel'nosti. In: *Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki:* 3-40. Moskva: MGU.

Andrukovič, I.F., Korolev, E.I. (1977). O statističeskich i leksiko-grammatičeskich svojstvach slov. *Naučno-techničeskaja informacija* 2, Nr. 4, 1-9.

Anochin, P.K. (1962). Teorija funkcional'noj sistemy kak predposylki k postroeniju fiziologičeskoj kibernetiki. In: *Biologičeskie aspekty kibernetiki:* 74-91. Moskva.

Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In:

Doležel, L., Bailey, R.W. (eds.), *Statistics and Style:* 57-67. New York-London: Elsevier.
Arapov, M.V. (1981). Sistemnyj analiz leksičeskoj struktury tekstov. *Sistemnye issledovania. Ežegodnik* 1980, 372-403.
Arapov, M.V., Cherc, M.M. (1974). *Matematičeskie metody v istoričeskoj lingvistike.* Moskva: Nauka.
Arapov, V.M., Cherc, M.M. (1983). *Mathematische Methoden in der historischen Linguistik.* Bochum: Brockmeyer.
Arapov, M.V., Efimova, E.N. (1975). Ponjatie leksičeskoj struktury teksta. *Naučno-techničeskaja informacija* 2, Nr. 6, 3-7.
Arapov, M.V., Efimova, E.N., Šrejder, Ju.A. (1975). O smysle rangovych raspredelenij. *Naučno-techničeskaja informacija* 2, Nr. 1, 9-20; Nr. 2, 9-20.
Arapov, M.V., Šrejder, Ju.A. (1978). Zakon Cipfa i princip dissimetrii sistemy. *Semiotika i grammatika* 10, 74-95.
Arapov, M.V., Ter-Gasparian, L.I., Cherc, M.M. (1978). Sravnenie častotnych slovarej. *Naučno-techničeskaja informacija* 2, Nr. 4, 20-29.
Bartkov, B.I. (1982). Količestvennaja morfemografija (derivatografija) anglijskogo, nemeckogo, francuzskogo i russkogo jazyka. In: *Osnovosloženie i poluaffiksacija v naučnom stile i literaturnoj norme:* 27-55. Vladivostok.
Bartkov, B.I. (1983). Količestvennye metody v derivatologii. In: *Issledovanija derivacionnoj podsistemy količestvennym metodom:* 3-40. Vladivostok.
Baudouin de Courtenay, I.A. (1963). *Izbrannye trudy po obščemu jazykoznaniju I-II.* Moskva.
Bektaev, K.B. (1978). *Statistiko-informacionnaja tipologija tjurkskogo teksta.* Alma-Ata: Izdatel'stvo Nauka Kazachskoj SSR.
Bektaev, K.B., Luk'janenkov, K.F. (1971). O zakonach raspredelenija edinic pis'mennoj reči. In: *Statistika reči i avtomatičeskij analiz teksta:* 47-112. Leningrad: Nauka.
Bektaev, K.B., Piotrovskij, R.G. (1973). *Matematičeskie metody v jazykoznanii I-II.* Alma-Ata.
Beljaeva, T.M., Vasil'eva, N.M. (1984). Slovoobrazovatel'noe gnezdo v slovare i ego funkcional'naja nagruzka v reči. In: *Derivatologija i derivatografija literaturnoj normy i naučnogo stilja:* 28-35. Vladivostok.
Belonogov, G.G. (1962). O nekotorych statističeskich zakonomernostjach v russkoj pis'mennoj reči. *Voprosy jazykoznanija* 1962; Nr. 1, 100-101.
Belonogov, G.G., Frolov, G.D. (1963). Empiričeskie dannye o raspredelenii bukv v russkoj pis'mennoj reči. *Problemy kibernetiki* 9, 287-305.
Belonogov, G.G., Novoselov, A.P. (1971). Nekotorye količestvennye zakonomernosti v avtomatizirovannych informacionnych sistemach. In: *Avtomatičeskaja pererabotka teksta metodami prikladnoj lingvistiki. Materialy vsesojuznoj konferencii:* 219-220. Kišinev.

Belonogov, G.G., Samodelkina, S.A. et al. (1985). Slovoobrazovatel'nye klassy russkich slov. *Naučno-techničeskja informacija* 2, Nr. 12, 22-24.
Belonogov, G.G., Zagika, E.A. et al. (1983). Avtomatizacija lingvističeskoj obrabotki slovarej. *Naučno-techničeskaja informacija* 2, Nr. 11, 20-24.
Benveniste, E. (1974). *Obščaja lingvistika.* Moskva.
Bernštejn, N.A. (1966). *Očerki po fiziologii dviženij i fiziologii aktivnosti.* Moskva: Medicina.
Best, K.-H., Altmann, G. (1996). Project report. *Journal of Quantitative Linguistics* 3, 85-88.
Bielfeldt, H.H. (1965[2]). *Rückläufiges Wörterbuch der russischen Sprache der Gegenwart.* Berlin: Akademie-Verlag.
Billmeier, G. (1968). Über die Signifikanz von Auswahltexten. *Forschungsberichte des Instituts für deutsche Sprache* 2, 126-171.
Blauberg, I.V. (1977). Celostnost' i sistemnost'. In: *Sistemnye issledovanija. Ežegodnik* 1977, 5-28. Moskva: Nauka.
Bogdanov, V.V. (1973). Statističeskaja koncepcija jazyka i reči. In: *Statistika reči i avtomatičeskij analiz teksta:* 9-19. Moskva: Nauka.
Boroda, M.G., Polikarpov, A.A. (1984). Zakon Cipfa-Mandel'brota i edinicy različnych urovnej organizacii teksta. *Učenye zapiski TGU* 689, 35-60.
Brookes, B.C. (1982). Quantitative analysis in the humanities. In: Guiter, H., Arapov, M.V. (1982), *Studies on Zipf's law:* 65-115. Bochum: Brockmeyer.
Brušlinskij, A.V. (1979). *Myšlenie i prognozirovanie (logiko-psichologičeskij analiz).* Moskva: Mysl'.
Budagov, R.A. (1978). Sistema i antisistema v nauke o jazyke. *Voprosy jazykoznanija* Nr. 4, 3-17.
Busemann, A. (1948). *Stil und Charakter. Untersuchungen zur Psychologie der individuellen Redeform.* Meisenheim/Glan.
Byčkov, V.N. (1984). K probleme obobščenija i interpretacija rangovych raspredelenij v statističeskoj lingvistike. *Učenye zapiski TGU 689,* 61-70.
Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distribution. In: Kučera, H., Francis, W.N. (eds.), *Computational analysis of present-day American English:* 406-424. Providence, R.I.: Brown UP.
Častotnyj slovar' sučasnoi ukrains'koi chudožnoi prozi. Bd. 1-2. Kiiv: Naukova dumka.
Čebanov, S.G. (1947). O podčinenii ukladov "indo-evropejskoj" gruppy zakonu Puassona. *Doklady AN SSSR, novaja serija* 55, Nr. 2.
Cherc, M.M. (1969). O predstavitel'nosti teksta zadannoj dliny. *Naučno-techničeskaja informacija* 2, Nr. 6, 26-29.
Coseriu, E. (1963). Sinchronija, diachronija i istorija. In: *Novoe v lingvistike, Vol.* 3: 143-343. Moskva.
Darčuk, N.P. (1975). *Individual'noe i obščee v leksičeskoj sisteme avtorskogo*

stilja (na materiale sovremennoj ukrainskoj chudožestvennoj prozy). Kiev: Diss.
Denisov, P.N., Kostomarov, V.G. (1970). Stilističeskaja differenciacija leksiki i problema razgovornoj reči. In. *Russkaja razgovornaja reč':* 69-75. Saratov.
Denisov, P.N., Morkovkin, V.V. (1978). *Kompleksnyj častotnyj slovar' russkoj naučnoj i techničeskoj leksiki.* Moskva: Russkij jazyk.
Dobrov, G.M. (1969). *Prognozirovanie nauki i techniki.* Moskva: Nauka.
Drujanov, L.A. (1980). *Zakony nauki, ich rol' i poznanie.* Moskva: Znanie.
Dugast, D. (1980). *La statistique lexicale.* Genève: Slatkine.
Eesti Kirjakeele seletussõnaraamat. Makett (1969). Tallin: Valgus.
Efremova, T.F. (1968). Iz nabljudenij nad strukturoj sovremennogo russkogo jazyka na urovne morfov. In: *Semantičeskie i fonologičeskie problemy prikladnoj lingvistiki:* 45-55. Moskva: MGU.
Embleton, S.M. (1986). *Statistics in historical linguistics.* Bochum: Brockmeyer.
Engwall, G. (1974). *Fréquence et distribution du vocabulaire dans un choix de romans français.* Stockholm: Skriptor.
Fickermann, I., Markner, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. *Glottometrika 6,* 115-126.
Filin, F.P. (1973). O strukture sovremennogo russkogo literaturnogo jazyka. *Voprosy jazykoznanija* 2, 5-12.
Filin, F.P. (1979). Nekotorye voprosy sovremennogo jazykoznanija. *Voprosy jazykoznanija Nr.* 4, 19-28.
Filosofskaja énciklopedija (1960). Moskva: Sovetskaja énciklopedija.
Förster, E., Rönz, B. (1979). *Methoden der Korrelations- und Regressionsanalyse.* Berlin: Die Wirtschaft.
Frumkina, R.M. (1960). Primenenie statističeskich metodov v jazykoznanii. *Voprosy jazykoznanija* 4.
Frumkina, R.M. (1961). K vorposu o tak nazyvaemom zakone Cipfa. *Voprosy jazykoznania* 2.
Frumkina, R.M. (1964). *Statističeskie metody izučenija leksiki.* Moskva: Nauka.
Frumkina, R.M. (1969). O verojatnostnom prognozirovanii v rečevom povedenii. In: *Problemy prikladnoj lingvistiki. Tezisy mežvuzovskoj konferencii Bd.* 2: 313-316. Moskva.
Fucks, W. (1956). Mathematical theory of word-formation. In: Cherry, C. (ed.), *Information theory:* 154-170. London: Butterworths Scientific Publications.
Fucks, W. (1957). Matematičeskaja teorija slovoobrazovanija. In: *Teorija peredači soobščenij:* 221-247. Moskva.
Gačečiladze, T.G., Cilosani, T.P. (1971). Ob odnom metode izučenija statističeskoj struktury teksta. In: *Statistika reči i avtomatičeskij analiz teksta:* 113-133. Leningrad: Nauka.
Garcia Hoz, V. (1953). *Vocabulario usual, común y fundamental.* Madrid.

Gerd, A.S. (1986). *Osnovy naučno-techničeskoj leksikografii.* Leningrad: LGU.
Ginzburg, R.S., Khidekel, S.S., Knyazeva, G.Y., Sankin, A.A. (1966). *A course in modern English lexicology.* Moskva: Vyssaja škola.
Golovin, B.N. (1968). O roli statistiki v opisanii jazykovych i rečevych stilej. In: *Častotnye slovari i avtomatičeskaja pererabotka tekstov. Tezisy dokladov:* 36-41. Minsk.
Golovin, B.N. (1971). *Jazyk i statistika.* Moskva: Prosveščenie.
Gor'kova, B.I. (1969). Rangovoe raspredelenie na množestvach naučno-techničeskoj informacii. *Naučno-techničeskaja informacija* 2, Nr. 7, 5-11.
Greenberg, J.H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26, 178-194.
Grigor'eva, A.S. (1981). *Statističeskaja struktura russkogo épistorjal'nogo teksta (leksika častnych pisem).* Leningrad: Diss.
Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie.* Paris. PUF
Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique.* Dordrecht: Reidel.
Haitun, S.D. (1983). *Naukometrija - sostojanie i perspektivy.* Moskva: Nauka.
Hakulinen, L. (1979). *Suomen kielen rakenne ja kehijtys.* Helsinki: IL.
Harlass, G., Vater, H. (1974). *Zum aktuellen deutschen Wortschatz.* Tübingen: Narr.
Herdan, G. (1964). *Quantitative linguistics.* London: Butterworths.
Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin: Springer.
Hint, M. (1988). *Eesti ilukirjanduskeele statistiline fonotaktika.* Tallinn: TPeDI.
Hoffmann, L., Piotrowski, R.G. (1979). *Beiträge zur Sprachstatistik.* Leipzig: VEB Verlag Enzyklopädie.
Jablonskaja, N.N. (1976). Častotnyj slovar' nemeckogo pod'jazyka chirurgii. In: *Voprosy prikladnoj lingvistiki Bd. 6,* 83-89. Dnepropetrovsk.
Jablonskij, A.I. (1977). Struktura i dinamika sovremennoj nauki. In: Gvišiani, D.M. (Hrsg.), *Sistemnye issledovanija. Ežegodnik* 1976: 66-90. Moskva: Nauka
Jakubajtis, T.A. (1963). Verojatnostnaja charakteristika slov s raznym količestvom slogov v latyšskom jazyke. *Izvestija AN Latyšskoj SSR 7,* 43-48.
Jakubajtis, T.A. (1981). *Časti reči i tipy tekstov.* Riga: Zinatne.
Jakubajtis, T.A., Skljarevič, A.N. (1978). *Verojatnostnye charakteristiki svjaznych tekstov.* Riga: Akademija Nauk LSSR.
Jarceva, V.N. (1970). Količestvennye i kačestvennye izmenenija v jazyke. In: *Leninizm i teoretičeskie problemy jazykoznanija.* Moskva.
Jardine, N., Sibson, R. (1968). A model for taxonomy. *Mathematical Biosciences* 2, 465-485.
Jiřáková, I. (1976). Zavisimost' količestvennogo sostava grammatičeskich kategorij

porjadka častej reči ot ob'ema častotnych slovarej russkogo jazyka. *Prague Studies in Mathematical Linguistics* 5, 37-52.
Juilland, A., Brodin, D., Davidovitch, C. (1970). *Frequency dictionary of French words.* The Hague-Paris: Mouton.
Kaasik, Ü., Tuldava, J. (1980). Sõnalõpu ja sõnapikkuse vahekorrast eestikeelses tekstis. *Tartu Ülikooli Toimetised* 477, 154-167.
Kaasik, Ü., Tuldava, J., Villup, A., Ääremaa, K. (1977). Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnaraamat. *Tartu Ülikooli Toimetised* 413, 5-140.
Kalinin, A.V. (1978). *Leksika russkogo jazyka.* Moskva: MGU.
Kalinin, V.M. (1964). Nekotorye statističeskie zakony matematičeskoj lingvistiki. In: *Problemy kibernetiki* 2. Moskva.
Kalinin, V.M. (1965). Funkcionaly, svjazannye s raspredeleniem Puassona i statističeskaja struktura teksta. *Trudy Matematičeskogo Instituta im. Steklova* 79. Moskva-Leningrad.
Kalinina, E.A. (1968). Izučenie leksiko-statističeskich zakonomernostej na osnove verojatnostnoj modeli. In: *Statistika reči:* 64-107. Leningrad: Nauka.
Kaširina, M.E. (1974). O tipach raspredelenija edinic v tekste. In: *Statistika reči i avtomatičeskij analiz teksta:* 335-360. Leningrad: Nauka.
Kļaviņa, S.P. (1977). *Sopostavlenie funkcional'nych stilej latyšskogo jazyka (lingvostatističeskoe issledovanie).* Vilnius: Diss.
Klimenko, N.F. (1974). Složnye glagoly v novogrečeskom i ukrainskom jazykach. In: *Strukturnaja i matematičeskaja lingvistika, Vol.* 2: 54-62. Kiev.
Kločkova, E.A. (1968). O raspredelenii klassov slov v nekotorych funkcional'nych stiljach russkogo jazyka. In: *Voprosy slavjanskogo jazykoznanija:* 109-118. Saratov.
Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Köhler, R. (1995). Maßeinheiten, Dimensionen und fraktale Strukturen in der Linguistik. *Zeitschrift für Empirische Textforschung* 2, 5-6.
Kohonen, T. (1980). *Associativnaja pamjat'.* Moskva: Mir.
Kolmogorov, A.N. (1956). Teorija verojatnostej. In: *Matematika, ee soderžanie i značenie, Bd.* 2. Moskva.
Kondakov, N.I. (1971). *Logičeskij slovar'.* Moskva: Nauka.
Korolev, E.I., Korsakova, I.I., Safronova, M.V. (1984). Častota upotreblenia slov v tekste i ich leksičeskie charakteristiki. *Naučno-techničeskaja informacija* 2, Nr. 2, 8-14.
Kozačkov, L.S. (1978). Informacionnye sistemy s ierarchičeskoj ("rangovoj") strukturoj. *Naučno-techničeskaja informacija* 2, Nr. 8, 15-24.
Kožina, M.N. (1977). *Stilistika russkogo jazyka.* Moskva: Prosveščenie.
Krallmann, D. (1966). *Statistische Methoden in der stilistischen Analyse.* Bonn: Diss.
Krámský, J. (1966). A quantitative analysis of Italian mono-, di- and trisyllabic words. *Travaux linguistiques de Prague 129-143.* Prague: Academia.
Kravec, A.S. (1976). *Priroda verojatnosti (filosofskie aspekty).* Moskva: Mysl'.
Krylov, Ju.K. (1982). Ob odnoj paradigme ligvostatističeskich raspredelenij. *Učenye zapiski TGU* 628, 80-102.
Krylov, Ju.K. (1985). K vorposu o dinamike narastanija ob'ema slovarja slučajnoj vyborki i svjaznogo teksta. *Učennye zapiski TGU* 711, 55-66.
Krylov, Ju. K. (1987). Stacionarnaja model' porožđenija svjaznogo teksta. *Učennye zapiski TGU* 774, 81-102.
Krylov, Ju.K., Jakubovskaja, M.D. (1977). Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova. *Naučno-techničeskaja informacija 2, Nr. 3,* 1-6.
Kubrjakova, E.S. (1970). Morfologičeskaja struktura slova v sovremennych slavjanskich jazykach. In: *Morfologičeskaja struktura slova v indoevropejskich jazykach:* 104-181. Moskva: Nauka.
Kučera, H., Francis, W.N. (1967). *Computational analysis of present-day American English.* Providence, R.I.: Brown UP.
Kul'gav, M.P. (1971). Imja suščestvitel'noe kak osnovnoe sredstvo voploščenija "nominal'nogo/substantivnogo stilja". In: *Nekotorye voprosy nemeckoj filologii:* 3-20. Pjatigorsk.
Kuraszkiewicz, W. (1958). Statystyczne badanie słownictwa polskich tekstów XVI wieku. In: *Z polskich studiów sławistycznych:* 240-257. Warszawa: Panstwowy Instytut Wydawniczny.
Land, K.Ch. (1977). Sravnitel'naja statika v sociologii. In: *Matematika i sociologija:* 371-401. Moskva: Mir.
Latviešu valodas biežuma vārdnīca (1972). Riga: Zinātne.
Lebedev, A.N. (1983). Zakonomernosti povtorenija slov v reči. *Psichologičeskij žurnal* 5, 11-22.
Lebedev, A.N. (1986). Nejrofiziologičeskie predely pamjati čeloveka i bogatstva ego leksiki. *Učenye zapiski TGU 745,* 95-108.
Leont'ev, A.A. (1969). *Jazyk, reč', rečevaja dejatel'nost'.* Moskva: Prosveščenie.
Levkovskaja, K.A. (1968). *Leksikologija sovremennogo nemeckogo jazyka.* Moskva: Vysšaja škola.
Leont'ev, A.A. (1974). Rečevaja dejatel'nost'. Problemy matematičeskogo modelirovanija rečevoj dejate'lnosti. In: *Osnovy teorii rečevoj dejatel'nosti:* 21-28, 73-80. Moskva: Nauka.
Liiv, H., Tuldava, J. (1987). O klassifikacii tekstov s pomoščju klaster-analiza. *Učennye zapiski TGU* 777, 55-68.
Liiv, H., Tuldava, J. (1993). On classifying texts with the help of cluster analysis. In: Hřebíček, G. Altmann (eds.), *Quantitative text analysis:* 253-262.Trier:

WVT.
Lounsbury, F. (1965²). Transitional probability. Linguistic structure and systems of habit-family hierarchies. In: *Psycholinguistics. A survey of theory and research problems:* 93-101. Bloomington.
Lurija, A.R. (1979). *Jazyk i soznanie*. Moskva; MGU.
Malachovskij, L.V. (1980). Principy častotnoj stratifikacii slovarnogo sostava jazyka. In: *Statistika reči i avtomatičeskij analiz teksta 1980:* 99-105. Leningrad. Nauka.
Manasjan, N.S. (1987). Ob ocenke parametrov lingvističeskich raspredelenij opredelennogo klassa. *Strukturnaja i prikladnaja lingvistika* 3, 94-97.
Mandelbrot, B. (1954). Structure formelle des textes et communication: deux études. *Word* 10, 1-27.
Maršakova, I.V. (1974). Issledovanie častotnogo slovarja ključevych slov. *Naučno-techničeskaja informacija* 2, Nr. 11, 7-13.
Martinet, A. (1963). Osnovy obščej lingvistiki. In: *Novoe v lingvistike, Vol.* 3: 347-566. Moskva: Izdatel'stvo inostrannoj literatury.
Martynenko, G.J. (1965). Nekotorye statističeskie nabljudenija na materiale bolgarskogo jazyka. In: *Statistiko-kombinatornoe modelirovanie jazykov:* 327-339. Moskva-Leningrad: Nauka.
Martynenko, G.J. (1978). Nekotorye zakonomernosti koncentracii i rassejanija ėlementov v lingvističeskich i drugich složnych sistemach. In: *Strukturnaja i prikladnaja lingvistika, Vol. 1:* 63-79. Leningrad: LGU.
Martynenko, G.J- (1982). Tipologija lingvostatističeskich raspredelenij. *Učenye zapiski TGU 628,* 103-120.
Marusenko, M.A. (1981). Ob izmerenii svjazi otraslevych terminosistem s primeneniem EVM. *Učenye zapiski TGU 591,* 74-81.
Maslov, Ju.S. (1975). *Vvedenie v jazykoznanie.* Moskva: Vyssšaja škola.
McKinnon, A. (1980). Aberrant frequency words: their identification and uses. *Glottometrika 2,* 108-124.
Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.
Mistrík, J. (1967). Matematiko-statističeskie metody v stilistike. *Voprosy jazykoznanija 3,* 42-52.
Mistrík, J. (1969). *Frekvencia slov v slovenčine.* Bratislava: Vydavatel'stvo SAV.
Mitropol'skij, A.K. (1971²). *Technika statističeskich vyčislenij.* Moskva; Nauka.
Monod J. (1970). *Le hazard et la nécéssité.* Paris: Seuil
Muller, Ch. (1968). *Initiation à la statistique linguistique.* Paris: Larousse.
Muller, Ch. (1976). Some recent contributions to statistical linguistics. *Statistical Methods in Linguistics 1976,* 136-147.
Müller, W. (1971). Wortschatzumfang und Textlänge. *Muttersprache 81,* 266-276.
Muravickaja, M.P., Slipčenko, L.D. (1982). Simmetrija v lingvističeskich sistemach. In: *Sistema i struktura jazyka v svete marksistsko-leninskoj metologii:* 70-84. Kiev: Naukova dumka.
Nalimov, V.V. (1979²). *Verojatnostnaja model' jazyka. O sootnošenii estestvennych i iskustvennych jazykov.* Moskva: Nauka.
Nalimov, V.V. (1979a). Funkcia raspredelenija verojatnostej kak sposob zadanija razmytych množestv. Nabroski metateorii (diskussia po rabotam L. Zadeh). *Avtomatika 6,* 80-87.
Nalimov, V.V., Mul'čenko, Z.M. (1969). *Naukometrija. Izučenie razvitija nauki kak informacionnogo processa.* Moskva: Nauka.
Neguljaev, G.S., Pokras, Ju.L., Kolesnikov, L.I. (1973). Avtomatizirovannyj otbor leksiki dlja informacionno-poiskovych tezaurusov. *Naučno-techničeskaja informacija 2, Nr. 2,* 16-24.
Neljubin, L.L. (1983). *Perevod i prikladnaja lingvistika.* Moskva: Vyssšaja škola.
Nemčenko, V.N. (1984). *Sovremennyj russkij jazyk. Slovoobrazovanie.* Moskva: Vyssšaja škola.
Nešitoj, V.V. (1975). Dlina teksta i ob'em slovarja. Pokazateli leksičeskogo bogatstva teksta. In: *Metody izučenija leksiki:* 110-118. Minsk: BGU
Nešitoj, V.V. (1984). Sistema nepreryvnych raspredelenij v informatike i lingvistike. *Naučno-techničeskaja informacija 2, Nr. 3,* 1-6
Nikonov, V.A. (1978). Dlina slova. *Voprosy jazykoznanija 6,* 104-111.*Ôigekeel-sussônaraamat.* Tallinn: Valgus.
Orlov, Ju.K. (1970). O statističeskoj strukture soobščenij, optimal'nych dlja čelovečeskogo rasprijatija (k postanovke voprosa). *Naučno-techničeskaja informacija* 2, Nr. 8, 11-16.
Orlov, Ju.K. (1976). Obobščennyj zakon Cipfa-Mandelbrota i častotnye struktury informacionnych edinic različnych urovnej. In: Guseva, E.K. (Hrsg:), *Vyčislitel'naja lingvistika:* 179-202. Moskva: Nauka.
Orlov Ju.K. (1978). Statističeskoe modelirovanie rečevych potokov. *Voprosy kibernetiki* 41, 66-99.
Orlov, Ju.K. (1980). Informacionnye potoki: statističeskij analiz i prognozirovanie. *Naučno-techničeskaja informacija* 2, Nr. 2, 23-30.
Orlov, Ju.K. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Guiter, H., Arapov, M.V. (Eds.), *Studies in Zipf's law:* 154-233. Bochum: Brockmeyer.
Ovčinnikov, N.F. (1966). *Principy sochranenija.* Moskva.
Ožegov, S.I. (1963⁵). *Slovar' russkogo jazyka.* Moskva.
Pankrac, G. Ja. (1981). Statističeskoe issledovanie fonologičeskoj struktury slova. *Učenye zapiski TGU* 591, 82-90.
Panov, E.N. (1980). *Znaki, simboly, jazyki.* Moskva: Znanie.
Pao, M.L. (1978). Automatic text analysis based on transition phenomena of word occurrences. *Journal of American Social Information Science* 29, 121-124.

Papp, F. (1967). O nekotorych količestvennych charakteristikach slovarnogo sostava jazyka. *Slavica (Debrecen)* 7, 51-58.
Papp, F. (1969). O mašinnoj obrabotke odnojazyčnych slovarej (na materiale vengerskogo jazyka). *Naučno-techničeskaja informacija* 2, Nr. 3, 20-29.
Papp, F. (1980). Lingvostatistika i vengerskij jazyk. *Učenye zapiski TGU* 518, 15-37.
Perebejnos, V.I. (1984). Opredelenie nadežnosti dannych častotnogo slovarja. *Učenye zapiski TGU* 689, 103-110.
Perebijnis, V.S. (1970). *Kil'kisni ta jakisni charakteristiki sistemi fonem sučasnoj ukrainskoi literaturnoi movi.* Kiiv: Naukova dumka.
Petrenko, B.V. (1974). Issledovanie dokumental'nogo informacionnogo potoka na osnove analiza zaprosov. *Naučno-techničeskaja informacija* 2, Nr. 10, 3-8.
Pikver, A. (1973). *O primemenenii distributivno-statističeskogo metoda v morfemike.* Moskva: Diss.
Piotrovskij, R.G. (1975). *Tekst, mašina, čelovek.* Leningrad: Nauka.
Piotrovskij, R.G. (1979). *Inženernaja lingvistika i teorija jazyka.* Leningrad: Nauka.
Piotrovskij, R.G., Bektaev, K.B., Piotrovskaja, A.A. (1977). *Matematičeskaja lingvistika.* Moskva: Vysšaja škola.
Piotrovskij, R.G., Turygina, L.A. (1971). Antonimija "jazyk - reč'" i statističeskaja interpretacija normy jazyka. In: *Statistika reči i avtomatičeskij analiz teksta:* 5-46. Leningrad: Nauka.
Piotrowski, R.G. (1984). *Text , Computer, Mensch.* Bochum: Brockmeyer.
Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985). *Mathematische Linguistik.* Bochum: Brockmeyer.
Polikarpov, A.A. (1976). *Faktory i zakonomernosti analitizacii jazykovogo stroja.* Moskva: Diss.
Polikarpov, A.A. (1987). Polisemija: sistemno-kvantitativnye aspekty. *Učenye zapiski TGU* 774, 135-154.
Price, D. (1966). *Malaja nauka, bol'šaja nauka.* In: *Nauka o nauke:* 281-384. Moskva: Progress.
Rakitov, A.I. (1977). *Filosofskie problemy nauki. Sistemnyj podchod.* Moskva: Mysl'.
Ratkowsky, D.A., Halstead, M.H., Hantrais, L. (1980). Measuring vocabulary richness in literary works: a new proposal and a re-assessment of some earlier measures. *Glottometrika* 2, 125-145.
Roberts, A.H. (1965). *Statistical analysis of American English.* The Hague: Mouton.
Rubaškin, V.Š. (1976). Priznak i značenie. *Naučno-techničeskaja informacija 2, Nr.* 3, 3-10.
Ruzavin, G.I. (1978). *Naučnaja teorija. Logiko-metodologičeskij analiz.* Moskva: Mysl'.

Rybnikov, K.A. (1979). *Vvedenie v metodologiju matematiki.* Moskva: MGU.
Ryzin, J.v. (ed.) (1977). *Classification and clustering.* New York: Academic Press.
Sačkov, Ju.V. (1971). *Vvedenie v verojatnostnyj mir. Voprosy metodologii.* Moskva: Nauka.
Sadovskij, V.N. (1974). *Osnovanija obščej teorii sistem.* Moskva: Nauka.
Sadovskij, V.N. (1979). Razvitie metodologii sistemnych issledovanij. *Obščestvennye nauki* 3, 78-93.
Šajkevič, A.J. (1968). Opyt statističeskogo vydelenija funkcional'nych stilej. *Voprosy jazykoznanija* 1, 64-76
Šajkevič, A.J. (1982). *Distributivno-statističeskij analiz tekstov.* Leningrad: Diss.
Sambor, J. (1984). Menzerath's law and the polysemy of words. *Glottometrika* 6, 94-114.
Saukkonen, P., Haipus, M., Niemikorpi, A., Sulkala, H. (1979). *Suomen kielen taajuussanasto. A frequency dictionary of Finnish.* Porvoo-Helsinki: Juva.
Saussure, F. de (1977). *Trudy po jazykoznaniju.* Moskva: Progress.
Serebrennikov, B.A. (Hrsg.) (1972). *Obščee jazykoznanie. Vnutrennaja struktura jazyka.* Moskva: Nauka.
Serebrennikov, B.A. (Hrsg.) (1973). *Obščee jazykoznanie. Metody lingvističeskich issledovanij.* Moskva: Nauka.
Setälä, V. (1972). *Suomen kielen dynamiikka.* Helsinki: SKS.
Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423, 623-656.
Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika* 42, 425-440
Slipčenko, L.D. (1973). Zakonomernosti fonemnoj struktury slova v anglijskom jazyke. In: *Matematičeskaja lingvistika* 1, 104-109. Kiev: KGU.
Smirnickij, A.I. (1956). *Leksikologija anglijskogo jazyka.* Moskva: Nauka.
Solncev, V.M. (1977[2]). *Jazyk kak sistemno-strukturnoe obrazovanie.* Moskva: Nauka.
Somers, H.H. (1959). *Analyse mathématique du langage: lois générales et mesures statistiques.* Louvain: Nauwelaerts.
Somers, H.H. (1966). Statistical methods in literary analysis. In: Leed, J. et al. (eds.), *The computer and literary style.*: 128-140. Kent, Ohio: Kent State UP.
Šrejder, Ju.A. (1967). O vozmožnosti teoretičeskogo vyvoda statističeskich zakonomernostej teksta (k obosnovaniju zakona Cipfa). In: *Problemy peredači informacii Bd.* 3, 57-63. Moskva.
Stevens, S.S. (1960). *Éksperimental'naja psichologija. Bd. I.* Moskva.
Stoff, V.A. (1972). *Vvedenie v metodologiju naučnogo poznanija.* Leningrad: LGU.

Suppes, P., Zinnes, G. (1967). Osnovy teorii izmerenij. In: *Psichologičeskie izmerenija:* 9-100. Moskva: Mir.
Suslov, I.P. (1978). Verojatnost' v sisteme naučnych kategorij. In: *Dinamičeskaja i verojatnostnaja optimizacija ėkonomiki:* 43-58. Novosibirsk.
Swadesh, M. (1960). Leksikostatističeskoe datirovanie doistoričeskich ėtničeskich kontaktov. In: *Novoe v lingvistike Bd.* 1, 23-52. Moskva: Izdatel'stvo inostrannoj literatury.
Těšitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics* 3, 103-120.
Tichonov, A.N. (1983). Sistema russkogo slovoobrazovanija v svete količestvennych dannych. In: *Issledovanie derivacionnoj podsistemy lokičestvennym metodom:* 61-73. Vladivostok.
Tiščenko, V. (1970). Častota častii movi v riznich funkcional'nych stiljach sučasnoj ukrains'koj movi. In: *Pitanija strukturnoi leksikologii.* Kiiv.
Tuldava, J. (1971). Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov. *Linguistica* 4, 199-220.
Tuldava, J. (1974). O statističeskoj strukture teksta. In: *Sovetsksja pedadodika i škola Bd.* 9, 5-33. Tartu.
Tuldava, J. (1976). Opyt kvantitativnogo analiza chudožestvennogo stilja. *Učennye zapiski TGU* 396, 122-141.
Tuldava, J. (1977a). O kvantitativnych charakteristikach bogatstva leksičeskogo sostava chudožestvennych tekstov. *Učennye zapiski TGU* 437, 159-175.
Tuldava, J. (1977b). Quantitative relations between the size of text and the size of vocabulary. *SMIL Quarterly, Journal of Linguistic Calculus* 4, 28-35.
Tuldava, J. (1977c). Sagedussônastik leksikostatistilise uurimise objektina. *TRÜ Toimetised* 413, 141-171.
Tuldava, J. (1978). Kvantitativnoe issledovanie struktury odnosložnogo slova v ėstonskom jazyke. *Učenye zapiski TGU 453,* 115-135.
Tuldava, J. (1979). O nekotorych kvantitativno-sistemnych charakteristikach polisemii. *Učenye zapiski TGU* 502, 107-141.
Tuldava, J. (1980). K voprosu ob analitičeskom vyraženii svjazi meždu ob'emom slovarja i ob'emom teksta. *Učenye zapiski TGU* 549, 113-144.
Tuldava, J. (1981). Opyt klassifikacii tekstov s pomoščju klaster-analiza. *Učenye zapiski TGU* 591, 136-157.
Tuldava, J. (1982). Kvantitativnoe issledovanie genetičeskogo sostava leksiki ėstonskogo jazyka. *Učenye zapiski TGU* 628, 136-166.
Tuldava, J. (1983). Social'naja differenciacija leksiki ėstonskogo jazyka s kvantitativnoj točki zrenija. *Učenye zapiski TGU* 658, 149-177.
Tuldava, J. (1983a). Kvantitativnoe issledovanie leksiko-semantičeskich grupp v ėstonskom jazyke. *Učenye zapiski TGU* 656, 123-152.
Tuldava, J. (1984). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki (na materiale ėstonskogo jazyka).* Tartu: TGU.
Tuldava, J. (1984a). Razvitie leksiki ėstonskogo jazyka po dannym slovarej XVII-XX vekov. *Učenye zapiski TGU* 684, 115-126.
Tuldava, J. (1985). Častotnaja struktura teksta i zakon Cipfa. *Učenye zapiski TGU* 711, 93-116.
Tuldava, J. (1986). Dlina slova i raspredelenie slov po dline v tekste i slovare. *Učenye zapiski TGU* 736, 150-166.
Tuldava, J. (1986a). O častotnom spektre leksiki teksta. *Učenye zapiski TGU* 745, 139-162.
Tuldava, J. (1990). Symmetrie und Asymmetrie in der Sprache. In: Saukkonen, P. (Ed.), (1992), *What is language synergetics?:* 40-43. Oulu: Universitas Ouluensis.
Tuldava, J. (1995). *Methods in quantitative linguistics.* Trier: WVT.
Tuldava, J. (1996). The frequency spectrum of text and dictionary. *Journal of Quantitative Linguistics* 3, 38-50.
Ufimceva, A.A. (1968). *Slovo v leksiko-semantičeskoj sisteme jazyka.* Moskva: Nauka.
Vannikov, Ju.V. (1979). *Sintaksis reči i sintaksičeskie osobennosti russkoj reči.* Moskva: Russkij jazyk.
Veneckij, I.G., Kil'dišev, G.S. (1975[3]). *Teorija verojatnostej i matematičeskaja statistika.* Moskva: Statistika.
Ventcel', E.S. (1976). *Issledovanie operacij.* Moskva: Znanie.
Vertel', V.A., Vertel', E.V. (1970). Algoritmy polučenija častotnogo slovarja s učetom dliny slovoform. In: *Statistika teksta. Vol. 2: Avtomatičeskaja pererabotka teksta:* 290-311. Minsk.
Viks, U. (1980). *Klassifikatoorne morfoloogia. Verb.* Tallinn.
Villup, A. (1978). A.H. Tammsaare romaani "Tôde ja ôigus" I köite autori - ja tegelaskône sagedussônastik. *TRÜ Toimetised* 446, 5-106.
Vinogradov, V.V. (1938). *Sovremennyj russkij jazyk.* Moskva.
Vinogradov, V.V. (1947). *Russkij jazyk (Grammatičeskoe učenie o slove).* Moskva-Leningrad: Učpedgiz.
Vinogradov, V.V. (1967). *Stilistika. Teorija poėtičeskoj reči. Poėtika.* Moskva: Izdatel'stvo AN SSSR.
Višnjakova, S.M. (1976). Vydelenie suščestvitel'nych i prilagatel'nych pri avtomatičeskom analize teksta. *Naučno-techničeskaja informacija* 2, Nr. 3, 15-18.
Weibull, W. (1939). *A statistical theory of the strength of materials.* Stockholm.
Wiener, N. (1964). *Ja - matematik.* Moskva: Nauka.
Williams, C.B. (9170). *Style and vocabulary: numerical studies.* London: Griffin.
Woronczak, J. (1972). Metody vyčislenija pokazatelej leksičeskogo bogatstva tekstov. In: *Semiotika i isskustvo:* 232-249. Moskva.
Zacharova, A.B. (1967). Opyt statističeskogo issledovanija ustnoj reči rebenka. In:

Issledovanija po jazyku i folkloru, Vol. 2: 16-38. Novosibirsk.
Zadeh, L.A. (1965). Fuzzy sets. *Information and Control* 8, 338-353.
Zadeh, L.A. (1976). Ponjatie lingvističeskoj peremennoj i ego primenenie k prinjatiju približennych rešenij. Moskva: Mir.
Zadeh, L.A. (1980). Razmytye množestva i ich primenenie v raspoznavanii obrazov v klaster-analize. In: *Klassifikacija i klaster:* 208-247. Moskva: Mir.
Zaplatkina, N.I. (1975). *Grafemnaja struktura odnosložnych slov v slavjanskych jazykach.* Kiev: Diss.
Zaplatkina, N.I. (1982). Sistemnyj podchod k izučeniju jazykovych javlenij. In: *Sistema i struktura v svete marksistsko-leninskoj metodologii.* Kiev: Naukova dumka.
Zasorina, L.N. (1966). *Avtomatizacija i statistika v leksikografii.* Leningrad: LGU.
Zasorina, L.N. (1967). Opyt statističeskogo issledovanija ustnoj reči rebenka. In: *Issledovanija po jazyku i folkloru, Bd.* 2: 16-38. Novosibirsk.
Zasorina, L.N. (Hrsg.) (1977). *Častotnyj slovar' russkogo jazyka.* Moskva: Russkij jazyk.
Zipf, G.K. (1935). *The psycho-biology of language.* Cambridge, Mass.: Houghton Mifflin.
Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley.
Zsilka, T. (1974). *Stilisztika és statisztika.* Budapest.
Zubov, A.V. (1980). O verojatnostno-algoritmičeskom podchode k porožděniju teksta. In: *Éksperimental'naja fonetika i prikladnaja lingvistika:* 178-183. Minsk.

Namensregister

Sachregister