QUANTITATIVE LINGUISTICS

Vol. 15

Sprache, Text, Kunst

Quantitative Analysen

von

Ju. K. Orlov

M. G. Boroda

I. Š. Nadarejšvili

Studienverlag Dr. N. Brockmeyer

Bochum 1982

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Orlov, Jurij K.:
Sprache, Text, Kunst: quantitative Analysen / von Ju. K. Orlov; M. G. Boroda; I. Š. Nadarejšvili. – Bochum: Studienverlag Brockmeyer, 1982.
(Quantitative linguistics; Vol. 15)
ISBN 3-88339-243-X

NE: Boroda, M. G.:; Nadarejšvili, I. Š.:; GT

ISBN 3-88339-243-X

Querenburger Höhe 281, 4630 Bochum 1
Druck Thiebes GmbH & Co Kommanditgesellschaft Hagen

We would like to express our gratitude

to the

STIFTUNG VOLKSWAGENWERK

a generous grant from which made possible the translation of the Russian articles in this volume.

VORWORT DER REDAKTION

Der vorliegende Sammelband stellt eine Auswahl aus den Ergebnissen der Untersuchungen eines Kollektivs georgischer Wissenschaftler dar, das von dem Mathematiker Ju.K. Orlov geleitet wird. Die hier präsentierte Entwicklung der Textanalyse blieb im Westen bisher fast völlig unbekannt. Die Gründe sind leicht festzustellen: Erstens erscheinen diese Arbeiten in äußerst schwer beschaffbaren Zeitschriften und zweitens ergäben sich Sprachschwierigkeiten, auch wenn man die Zeitschriften zur Verfügung hätte. Die Aufgabe des vorliegenden Bandes besteht darin, einige Arbeiten des Kollektivs, die die Autoren selbst ausgewählt haben, im Westen bekannt zu machen. Es sind diejenigen Werke, die die Arbeit des Teams und ihre Perspektiven am besten charakterisieren. Das grandiose Spektrum wird trotz dieser Beschränkung von Linguisten kaum im vollen Umfang gewürdigt werden können.

Die zahlreichen Aspekte und Konsequenzen dieser Arbeiten lassen sich in einem Vorwort nur selektiv und stichwortartig darstellen. Wir beschränken uns daher auf die allgemeinen Züge, ohne die einzelnen Arbeiten zu referieren.

Gleich am Anfang wird die Illusion zerstört, daß es eine Textgrundgesamtheit in der Sprache gibt, die homogen genug wäre, um konstante Parameter zu haben. Jeder Text ist eine Individualität für sich. Dies ist völlig konform mit den Ansichten der Literaturwissenschaftler, die sich eben mit der Analyse der Einzeltexte beschäftigen. Unmittelbar werden sie aber belehrt, daß trotz der schöpferischen Freiheit jeder Text - wie jedes Ding der Realität - Gesetzen unterliegt, die sich mathematisch erfassen lassen. Im Rahmen dieser Gesetze können die Parameter variieren, jenach den Randbedingungen der Texterzeugung (z.B. Genre, Stil, Raum, Zeit usw.). Dies hat für die Literaturwissenschaftler die äußerst unangenehme Konsequenz: Sie müssen die Randbedingungen metrisieren, um die Parameter in den Gesetzesformeln bestimmen zu können. Damit stößt man an das größte Problem der Literaturwissenschaft, die Abneigung gegen Mathematik. Bezeichnet man

Textkonstruktion = Gesetze +(Sprach)Regeln + Randbedingungen, so kann man etwas bedauernd konstatieren, daß die orthodoxe Literaturwissenschaft sich nur mit den Randbedingungen beschäftigt und die Stufe einer Protowissenschaft erreicht hat. Auch die "Standardlinguistik" hat erst die Stufe der Regeln erreicht.

Mathematisierung ist kein Problem der Begabung, sondern der wissenschaftlichen Redlichkeit. Man kann z.B. Stilistik, Texttheorie usw. auch in orthodoxen Bahnen weiterführen, bis sie irgendwie "von alleine" zum Anachronismus werden. Echter wissenschaftlicher Fortschritt läßt sich aber kaum verdrängen, auch wenn irgendwann später die Resultate dieses Bandes und die Entdeckungen von Zipf, Yule, Simon, Mandelbrot, Arapov, Šrejder u.a. nur als Spezialfälle einer sehr allgemeinen Texttheorie bezeichnet werden.

Ein weiterer Aspekt dieses Bandes ist die Feststellung, daß es Gesetze gibt, die vielleicht in der gesamten schöpferischen Tätigkeit des Menschen einheitlich wirken. Das verallgemeinerte Zipf-Mandelbrotsche Gesetz, dem ein großer Teil des Bandes gewidmet ist, gilt nicht nur in der Sprache, sondern auch in der Musik und in der Malerei. Es ist anzunehmen, daß das gesamte menschliche Verhalten (nicht nur das schöpferische) einheitlich ist und nach denselben Gesetzen verläuft. Es ist die Aufgabe der Wissenschaften, eben diese Gesetze zu entdecken. Dieser Entdeckungsweg ist äußerst mühsam und es ist oft gerade die Mathematisierung, die uns zur Formulierung gewisser Entitäten zwingt, durch die das Gesetz wirkt. Ein gutes Beispiel sind Borodas musikalische Motive und Phrasen, deren "Konstruktion" eine Bedingung zur Entdeckung der Gesetze der musikalischen Komposition ist. Intuitive Segmentierungen des Textes, der Sprache, der Melodie, des Bildes u.a. führen nicht weiter. Intuition (auch literaturwissenschaftliche) ist nur bei den Anfängen der Forschung von heuristischem Wert; in der reifen Wissenschaft, in der man Theorien aus gesetzesartigen Aussagen aufstellt, hat sie kaum Platz. An mehreren Stellen des Bandes werden aus der Intuition stammende Aussagen von Dichtern (die sich solche Aussagen leisten können) exakt überprüft und im Lichte der Gesetze interpretiert. Es ist faszinierend zu erfahren, welche Rolle der sogenannte Zipfsche Umfang

bei der Verfassung eines dichterischen Werkes spielt und welche psychologischen Zusammenhänge sich hinter ihm verbergen. Alle Geisteswissenschaften leiden daran, daß sie in dem ziemlich vagen Gebilde der nichtmateriellen Erscheinungen keinen festen Fuß fassen können. Die ersten strikt definierten Einheiten und Eigenschaften, die ersten Messungen, das erste entdeckte Gesetz sind die größten Pioniertaten, die eine Entdeckungslavine in Bewegung bringen. In diesem Sinne ist der vorgelgte Band eine Pioniertat und man kann erwarten, daß sich in den nächsten Jahren diese Richtung stark entwickeln wird.

Gesetze der Sprache, die für bestimmte Entitäten (z.B. Wörter) gelten, sind die einzigen objektiven Kriterien der Sprachanalyse. Alle anderen Kriterien sind lediglich Desiderata, konventionelle Festsetzungen oder ad hoc Hilfsmittel. Wenn z.B. eine bestimmte Wortidentifizierung zu einer Zählung führt, die von dem Zipf-Mandelbrotschen Gesetz abweicht, so muß man wohl das Wort anders identifizieren. Die beste Identifizierung ist diejenige, die bei der Textauszählung zu der besten Übereinstimmung mit dem Gesetz führt.

Die Forschung schaukelt sich "von alleine" nach vorne: Fruchtbare, korrekt metrisierte Begriffe ermöglichen Gesetzeshypothesen zu formulieren; aus den Gesetzen leitet man neue Hypothesen ab, in denen weitere Begriffe erscheinen, die metrisiert und gemessen werden müssen; es werden neue Zusammenhänge zwischen den Begriffen entdeckt usw. Der schwierigste Schritt ist immer der erste. In diesem Band wurden gleichzeitig mehrere Schritte getan, die vollkommen ausreichen, um ein Forschungsprogramm im Sinne von Lakatos zu gründen.

Ein Gesetz ist neben anderem eine gut bestätigte Hypothese. In dieser Hinsicht haben sich die Verfasser sehr viel Mühe gegeben. Zur Überprüfung des Zipf-Mandelbrotschen Gesetzes wurde Literatur von einfachen Formen bis zu den modernen literarischen Texten, von der klassischen bis zur heutigen Musik, ein kleineres Repertoire von Bildern und eine Reihe von Frequenzwörterbüchern herangezogen. Bedenkt man, daß es sich in jedem Fall um große Materialmassen handelt, die nur mit einem Computer bearbeitet werden können, so sieht man, daß die Epoche der romantischen Individual-

arbeit in der Textanalyse zu Ende geht. Die quantitativ ausgerichtete "content analysis" mit der quantitativen Analyse der Häufigkeitsstruktur setzt Maßstäbe, liefert Grundlagen, stellt ein neues Paradigma auf. Die zögernde Entwicklung der Textanalyse kann man nicht nur durch die mangelhafte mathematische Ausbildung der Linguisten, sondern auch durch die erdrückende Menge des Materials entschuldigen. Die Menge der literarischen Daten (in jedem einzelnen Werk) ist wie Bäume, die uns daran hindern, den Wald zu sehen. Der "Wald" ist eine nach ökologischen Gesetzen gewachsene Entität, der Text ist eine nach psychologischen und sprachlichen (musikalischen usw.) Gesetzen erzeugte Entität, bei der die schöpferische Freiheit zwar alle Regeln, aber keine Gesetze überschreiten kann.

Man kann ruhig annehmen, daß das Zipf-Mandelbrotsche Gesetz nicht das einzige Gesetz der Textbildung ist. Es ist die erste Schwalbe, nach der hoffentlich der Frühling kommt. Das komplizierte Gewebe der menschlichen Schöpfung wird ohne Zweifel von einer Reihe von Gesetzen reglementiert, die möglicherweise (aber nur möglicherweise) aufeinander abgestimmt sind. Zu denen gesellen sich die Gesetze der menschlichen Rezeption, allgemeine Kommunikationsgesetze, die Sender und Empfänger zu Kompromissen zwingen usw. Einige Gesetze fangen in dem Augenblick an zu wirken, in dem der Verfasser den künftigen Umfang des Werkes bestimmt, andere regulieren die Plazierung bestimmter Wörter (z.B. in Gedichten), noch andere steuern die meßbaren Eigenschaften eines zu erzeugenden Satzes in Abhängigkeit davon, wie der letzte oder mehrere vorhergehende Sätze gestaltet sind usw. Man soll sich vergegenwärtigen, daß man es hier nicht mit (syntaktischen, melodischen u.a.) Regeln, sondern mit Zufallsvariablen zu tun hat, die man nur probabilistisch modellieren kann. Es öffnet sich hier eine neue Welt, die uns wegen unserer starren Fixierung auf die oberflächlichen Texterscheinungen bisher verborgen blieb.

Die Autoren argumentieren vorsichtig, aber in breiter Perspektive und untermauern jede Behauptung mit Experimenten und reichlichen Belegen. Da in den hier vorgelegten Artikeln zahlreiche Verweise auf andere praktisch unzugängliche Arbeiten des Kollektivs zu finden sind, haben die Autoren mehrere Artikel mit

zusätzlichen Erklärungen und Anmerkungen versehen, die die Lücken zu überbrücken helfen.

Die einzelnen Artikel sind aus folgenden Quellen übersetzt worden:

Nr. 1 wurde von Ju.K. Orlov direkt für diesen Band geschrieben.

Nr. 2 ist die Übersetzung von Nadarejšvili & Orlov (1978), (siehe Literatur).

Nr. 3 ist die Übersetzung von Orlov (1976).

Nr. 4 ist die Übersetzung von Orlov (1978a).

Nr. 5 ist die Übersetzung von Nadarejšvili (1978).

Nr. 6 ist die Übersetzung von Boroda (1977a).

Nr. 7 ist die Übersetzung von Boroda, M.G., Ob opredelenii informacionnoj melodičeskoj edinici tipa frazy v muzyke. Soobščenija ANGSSR 89, 1978, 57-60.

Nr. 8 ist die Übersetzung von Boroda (1977).

Nr. 9 ist die Übersetzung von Vološin & Orlov (1972).

Nr. 10 ist die Übersetzung von Nadarejšvili & Orlov (1969).

Nr. 11 ist die Übersetzung von Boroda & Orlov (1970).

Nr. 12 ist die Übersetzung von Boroda & Nadarejšvili & Orlov & Čitašvili (1977).

Nr. 13 ist die Übersetzung von Nadarejšvili & Nadarejšvili & Orlov (1977).

Nr. 14 ist die Übersetzung von Boroda & Orlov (1978).

Es ist uns nicht gelungen, die bibliographischen Angaben im Literaturverzeichnis nach unseren Gepflogenheiten zu vervollständigen. Schuld daran ist die unterschiedliche Zitationsart in russisch geschriebenen Arbeiten. Die wichtigsten Angaben sind aber überall vorhanden.

Die Redaktion von QuL bedankt sich herzlichst bei der Allunionsagentur für Urheberrechte der UdSSR für die freundliche Erlaubnis, die vorliegenden Artikel ins Deutsche zu übersetzen und in dieser Reihe zu veröffentlichen. Weiter bedanken wir uns bei der Stiftung Volkswagenwerk, die die Übersetzung im Rahmen eines Projekts finanziert hat und schließlich bei H.J. Kemper, A. Falk, H. Sterner, V. von Brünning, I. Fickermann und

einigen Mitgliedern des Redaktionsrates, die die Übersetzungen, die Abbildungen, die Register, die Vorbereitung des repro-reifen Manuskripts und die vielen Korrekturen besorgt haben.

G. Altmann

INHALT

LINGUOSTATISTIK: AUFSTELLUNG VON SPRACHNORMEN ODER ANALYSE DES REDEPROZESSES? (DIE ANTINOMIE "SPRACHE-REDE" IN DER STATISTISCHEN LINGUISTIK)

Ju.K. Orlov

Die Idee der statistischen Analyse ist äußerst einfach: man zählt eine relativ kleine Stichprobe aus und bekommt eine Vorstellung über etwas, das unermeßlich viel größer ist: über die statistische Grundgesamtheit, aus der die Stichprobe erhoben wurde. Beispielsweise, wenn man eine Vorstellung über die Verteilung des Gewichts oder der Größe der Weizenkörner erhalten möchte, so braucht man keineswegs die ganze Ernte kornweise durchzumessen. Es reicht, eine zufällige Stichprobe zu erheben (in der Praxis reichen einige hundert Körner), um die Gewißheit zu erlangen, daß sich die Verteilung in der gesamten Ernte nicht allzu stark von der Verteilung in der Stichprobe unterscheidet.

Es muß lediglich die Zufälligkeit der Stichprobe gesichert werden. Wenn wir beispielsweise aus dem Lastwagen, der den Weizen vom Feld zum Elevator bringt, Körner von der obersten Weizenschicht entnehmen, so können wir sicher sein, daß dort durchschnittlich größere Körner sein werden, weil sich durch das Schütteln oben die größten Körner sammeln. Man muß die ganze Ladung ausschütten, ordentlich umschaufeln und für die Analyse ein - zwei Tassen entnehmen.

Als man diese Lehrbuchideen in den Bereich der Sprach- und Redeanalyse übertrug, erhielt man eine bestechend einfache linguistische Interpretation der statistischen Kategorien: durch die stichprobenartige Untersuchung dieser oder jener Datenfelder der Rede erforscht man die Sprachnormen (oder vielleicht den Usus; vgl. Bektaev, Delocerkovskaja, Piotrowski 1977). Um individuelle Abweichungen, die durch das Thema, den Stil des Autors usw. hervorgerufen werden, zu eliminieren, muß man in einer Stichprobe unterschiedliche Texte kombinieren, wodurch indi-

viduelle Eigenarten verwischt werden und die Durchschnittscharakteristiken, die die "sprachliche Grundgesamtheit" charakterisieren, deutlich hervortreten (s. Herdan 1966).

Diese Vorgehensweise folgt unmittelbar, gleichsam automatisch aus F. de Saussures Konzeption, nach der das höchste Ziel der Linguistik die Erforschung des Sprachsystems ist. Die durch dieses System erzeugte Sprechwirklichkeit müßte demnach irgendein Halbprodukt sein, das (vom Standpunkt des Linguisten) ausschließlich dazu bestimmt ist, daß aus ihm die Metarealität der "reinen Beziehungen" extrahiert wird.

Und nun kommt die Statistik und bietet etwas an, das wie die Automatisierung dieses schwierigen Prozesses aussieht. Jedoch "fürchtet die Danaer, die Geschenke bringen"! Die Sprache als Objekt der statistischen Analyse zeigte sich wesentlich komplizierter als das Weizenfeld. Niemand vermutete, daß auf den Forscher an diesem Weg Gefahren lauern, die sowohl mathematischer als auch linguistischer Natur sind. Die Mißachtung dieser Gefahren rief die heutige Krisensituation in der Linguostatistik hervor, in der jede einzelne Untersuchung einer lexikalischen Stichprobe mit einem absolut konkreten Anwendungszweck wie "Aufstellung eines Minimum-Vokabulars", "Zusammenstellung eines Wörterbuches für maschinelle Übersetzung", "Konstruktion eines Informationssuchenden Systems" (automatisches Steuerungssystem, Datenbank usw.)", gerechtfertigt werden muß. Und in der Tat lassen sich die Ergebnisse der Berechnungen in der Regel nur für den gegebenen konkreten Zweck verwenden, für den sie durchgeführt wurden (und dies auch nicht immer). Der Wunschtraum, gesamtsprachliche Normen festzulegen, ist heutzutage genauso unerfüllbar wie zu Kaedings Zeiten - ja, man bemüht sich sogar, nicht mehr an ihn zu denken.

Und die Methodologie statistischer Untersuchungen ist trotzdem so, als ob dieser Wunschtraum eine nahe Realität wäre. Zu einer selbständigen wissenschaftlichen Disziplin innerhalb der Linguistik, die nicht nur anwendungsbezogene, sondern auch grundlagenorientierte Bedeutung hat und zu den Nachbardisziplinen hin offen ist, können die quantitativen Methoden nach Überzeugung

des Verfassers erst dann werden, wenn die Forschungsziele und die Vorgehensweisen einen entsprechenden Wandel erfahren haben. Dazu ist ein tieferes Verständnis der untersuchten Erscheinungen nötig, die sowohl systemlinguistischer als auch mathematischer, psychologischer und physiologischer Natur sind. Ohne die Lösung aller dieser Probleme in Angriff zu nehmen, versuchen wir in dieser Arbeit diejenigen mathematischen und linguistischen Gefahren zu umreißen, deren Mißachtung die linguistische Statistik in die heutige deprimierende Lage gebracht hat, und skizzieren einige Möglichkeiten des Auswegs aus dieser Sackgasse.

1. DIE MATHEMATISCHEN GEFAHREN

Um die mathematischen Probleme sozusagen in reiner Form klarzulegen, nehmen wir an, daß man die Ergebnisse der Sprechtätigkeit der Sprecher oder der Schreiber in der gegebenen Sprache als die statistische Grundgesamtheit betrachten kann, aus der man echt zufällige, repräsentative Stichproben erheben kann. Wodurch unterscheidet sich die Sprache vom Weizenfeld bei diesen, die Situation bewußt vereinfachenden Voraussetzungen?

Im Grunde durch nichts, wenn man sich lediglich auf die Analyse der elementarsten linguistischen Einheiten und Kategorien wie z.B. Phoneme, Grapheme, Redeteile usw. beschränkt. Wenn das Inventar der untersuchten Entitäten klein ist und wenn sogar die seltensten mehr als 10 mal vorkommen (s. unten), so kann man mit hinreichender Sicherheit annehmen, daß eine weitere Vergrößerung des Stichprobenumfangs ihre Kenngrößen nicht wesentlich verändert, und man kann sie ohne die Gefahr eines groben Fehlers auf die "sprachliche Grundgesamtheit" übertragen.

Wenn wir aber zu komplexeren linguistischen Einheiten, wie z.B. Wörtern, übergehen, dann ändert sich das Bild wesentlich: für lexikalische Stichproben ist es charakteristisch, daß sie im Verhältnis zu der Grundgesamtheit, die man mit Hilfe des

Stichprobenverfahrens untersuchen möchte, hoffnungslos unzureichend sind.

Berechnen wir die absolute Häufigkeit F des Wortes in der Stichprobe, die, dividiert durch den Stichprobenumfang, einen einigermaßen zuverlässigen Wert der relativen Häufigkeit ergeben würde. Wenn die relativen Häufigkeiten klein sind, so kann man die Standardabweichung der absoluten Häufigkeit dieser Häufigkeit gleichsetzen und die Grenzen des Konfidenzintervalls der absoluten Häugigkeit, F_1 und F_2, als

$$F_{1,2} = F \pm t_\alpha \sqrt{F}$$

schreiben.

Wir werden eine Häufigkeit als zuverlässig geschätzt betrachten, wenn die halbe Länge $t_\alpha \sqrt{F}$ ihres Konfidenzintervalls gleich der Hälfte der Häufigkeit F selbst ist (mit anderen Worten: wir lassen einen dreifachen Fehler bei der Schätzung der Häufigkeit zu); das bedeutet, daß die Häufigkeit F die Gleichung $t_\alpha \sqrt{F} = \frac{1}{2} F$ erfüllen muß. Aus dieser Gleichung folgt $F = 4t_\alpha^2$. Wählt man verschiedene Werte des Konfidenzkoeffizienten α, so findet man die entsprechenden absoluten Häufigkeiten in der Tabelle 1.

Tabelle 1

α	t_α	$F = 4t_\alpha^2$
0.80	1.282	6.574
0.85	1.439	8.283
0.90	1.643	10.798
0.95	1.960	15.366
0.99	2.576	26.443
0.9973	3.000	36.000

Daraus ist ersichtlich, daß die relative Häufigkeit der Wörter, deren absolute Häufigkeit in der Stichprobe über 10 - 15 liegt, einigermaßen zuverlässig bestimmt wird. Hier sind Daten aus einigen Häufigkeitswörterbüchern (vgl. Tabelle 2).

Tabelle 2

Stichprobe Quelle, Zähleinheit	Umfang N	Vokabular V	Anzahl lex. Einheiten mit Häufigkeit F > 10	Anteil % lex. Einheiten mit Häufigkeit F > 10
Russische Elektronik (Kalinina 1968) Wortformen	50000	9464	783	8.27
Dasselbe	200894	21468	2862	13.33
Dasselbe, Lexeme	200388	6826	2030	29.74
Häufigkeitswörterbuch der Autorensprache estnischer Prosa (Tuldava 1977), Lexeme	99898	14654	1169	7.97
Dasselbe, Wortformen	99898	30733	1034	3.36
Häufigkeitswörterbuch urkainischer Prosa (V.S. Perebejnos, pers. Mitteilung), Lexeme	470560	33356	4566	13.69
Häufigkeitswörterbuch der Sprache Puškins (Frumkina 1963), Lexeme	544777	21197	5322	25.11
Häufigkeitswörterbuch des Russischen (Zasorina 1977), Lexeme	1066382	39268	8429	21.52
Häufigkeitswörterbuch französischer künstlerischer Prosa, Lexeme	71000000	71415	30392	42.56

Dies zeigt, daß nur für einen absolut kleineren Anteil der im Häufigkeitswörterbuch erfaßten Einheiten die Häufigkeit einigermaßen zuverlässig geschätzt wird, wobei sich dieser Anteil beim Anwach-

sen des Stichprobenumfangs nur äußerst langsam vergrößert. Außerdem ist dieser Anteil bei Wortformen kleiner als bei Lexemen auch im Falle eines lexikalisch reicheren Materials (vgl. die Häufigkeitswörterbücher der Sprache Puškins und der ukrainischen künstlerischen Prosa).

Man muß jedoch in Betracht ziehen, daß das Problem der Mangelhaftigkeit lexikalischer Stichproben nicht nur darauf zurückzuführen ist, daß man die Häufigkeiten der meisten Wörter nur äußerst unzuverlässig feststellen kann. Neben der sozusagen individuellen Dispersion der Häufigkeiten der Wörter von Stichprobe zu Stichprobe findet in unzureichenden Stichproben eine systematische Verschiebung aller ihrer quantitativen Kenngrößen in bezug auf dieselben Kenngrößen der Grundgesamtheit statt.

Am evidentesten ist die Verschiebung des Stichprobenwörterbuches. Der "Umfang des "allgemeinen Wortschatzes" der Sprache, der, wie bekannt, nicht mit statistisch-lexikographischen Methoden geschätzt wird, beläuft sich auf einige Hunderttausende Lexeme. Die bis jetzt größte lexikalische Stichprobe (s. Dictionnaire des . . . 1971) von etwa 70 Millionen Wortverwendungen enthält aber nur 70 Tausend Lexeme, d.h. um eine Ordnung weniger, als es im Wortvorrat der Sprache gibt. Diese Tatsache ist allgemein bekannt, aber bei weitem nicht alle Linguisten sind sich darüber im Klaren, daß die Reduktion des Stichprobenvokabulars auch andere Stichprobencharakteristika verschiebt.

Sei

$$\Pi_1, \Pi_2, \ldots, \Pi_i, \ldots, \Pi_V$$

die Menge der Lexemwahrscheinlichkeiten in einer lexikalischen Gesamtheit von V Lexemen. Die Summe dieser Wahrscheinlichkeiten ergibt 1. Dividiert man diese Summe durch die Zahl der Lexeme V, so erhält man die mittlere Lexemwahrscheinlichkeit in der Gesamtheit als 1/V. Die Menge der relativen Stichprobenhäufigkeiten

$$p_1, p_2, \ldots, p_i, \ldots, p_v$$

ergibt auch insgesamt 1, aber wenn das Stichprobenvokabular v

viel kleiner als V ist, dann ist auch die mittlere relative Lexemhäufigkeit in der Stichprobe 1/v viel größer als die mittlere Lexemwahrscheinlichkeit in der Gesamtheit 1/V. Das bedeutet aber, daß die ganze Menge der Häufigkeiten (wenn man sie wie üblich nach abnehmender Größe ordnet und in eine Graphik einträgt) höher liegt als die analog geordnete Menge der Lexemwahrscheinlichkeiten.

Um zu illustrieren, wie dies zustande kommt, nehmen wir ein kleines Beispiel, das alle Züge einer realen Situation hat. Nehmen wir eine für die Bestimmung der Häufigkeiten aller Buchstaben absichtlich unzureichende "Stichprobe", z.B. die erste Phrase aus dem "Held unserer Zeit" (Lermontov)

Я ЕХАЛ НА ПЕРЕКЛАДНЫХ ИЗ ТИФЛИСА

(Ich reiste mit Postpferden aus Tiflis).

Diese Phrase enthält 27 Buchstaben (Wortzwischenräume wurden nicht gezählt), d.h. eine Zahl, die dem Umfang des russischen Alphabets nahekommt (ein ähnliches Verhältnis kann man gewöhnlich auch bei lexikalischen Stichproben beobachten, deren Umfänge als groß betrachtet werden - Umfänge der Ordnung 10^5 und mehr entsprechen ungefähr den Schätzungen des Gesamt-Wortschatzes). In der Phrase werden insgesamt 16 unterschiedliche Buchstaben verwendet (d.h. die Hälfte des Alphabets; im Vergleich mit lexikalischen Stichproben ist das sehr günstig, da das Vokabular lexikalischer Stichproben um eine ganze Ordnung kleiner ist als die Schätzungen des Gesamtwortschatzes). Ordnet man die Buchstaben nach abnehmender Häufigkeit (bei gleicher Häufigkeit werden sie alphabetisch geordnet, wie man es gewöhnlich bei lexikalischen Stichproben macht), so erhält man die Tabelle 3.

Die Schätzungen der Wahrscheinlichkeiten in der Tabelle 3 stammen von Lebedev und Garmaš (1958) (die ursprünglichen Angaben wurden für die Zählung ohne Wortzwischenraum umgerechnet). Man kann leicht sehen, daß nur drei Buchstaben, P, C, T, hier eine niedrigere Häufigkeit als in der Grundgesamtheit haben. Alle anderen Buchstaben sind "die Glücklichen", denen es "ge-

Tabelle 3.

Rang	Buchstabe	Absolute Häufigkeit	Relative Häufigkeit	Wahrscheinlichkeit nach Lebedev & Garmaš (1958)
1	А	4	0.148	0.075
2	Е	3	0.111	0.088
3	И	3	0.111	0.076
4	Л	3	0.111	0.042
5	Н	2	0.074	0.064
6	Х	2	0.074	0.011
7	Д	1	0.037	0.030
8	З	1	0.037	0.019
9	К	1	0.037	0.034
10	П	1	0.037	0.028
11	Р*	1	0.037	0.048
12	С*	1	0.037	0.064
13	Т*	1	0.037	0.064
14	Ф	1	0.037	0.002
15	Ы	1	0.037	0.016
16	Я	1	0.037	0.022

lungen ist" (auf Kosten derer, die in der Stichprobe überhaupt nicht vorkommen) ihre Häufigkeit über die mittlere Häufigkeit zu erhöhen. Die Stärke dieses "Glücks" ist so groß, daß die Stichprobenhäufigkeiten auch in dem Falle größer bleiben, wenn man nicht die Häufigkeit und die Wahrscheinlichkeit ein- und desselben Buchstabens vergleicht, sondern den Vergleich nach der Größe durchführt; die größte Häufigkeit mit der größten Wahrscheinlichkeit usw. (diese Art des Vergleichs ist gemeint, wenn man auf die Graphik die nach abnehmender Größe geordnete Menge der Häufigkeiten aufträgt). Das Resultat eines solchen Vergleichs wird auf der Abb. 1 dargestellt, wo die Punkte die Häufigkeiten und die dünnen vertikalen Linien die Wahrscheinlichkeiten darstellen.

Wie leicht zu sehen ist, liegen hier von 10 Punkten 12 über den Buchstabenwahrscheinlichkeiten mit denselben Rangzahlen und nur 3 Buchstaben (Е, И, Н) nehmen denselben Platz in beiden geordneten Listen (nach Häufigkeit und nach Wahrscheinlichkeit) ein. Aber im Falle der lexikalischen Analyse stehen uns keine "allgemeinsprachlichen" Wortwahrscheinlichkeiten zur Verfügung. Im Gegenteil, wir wollen sie ja schätzen, indem wir die geord-

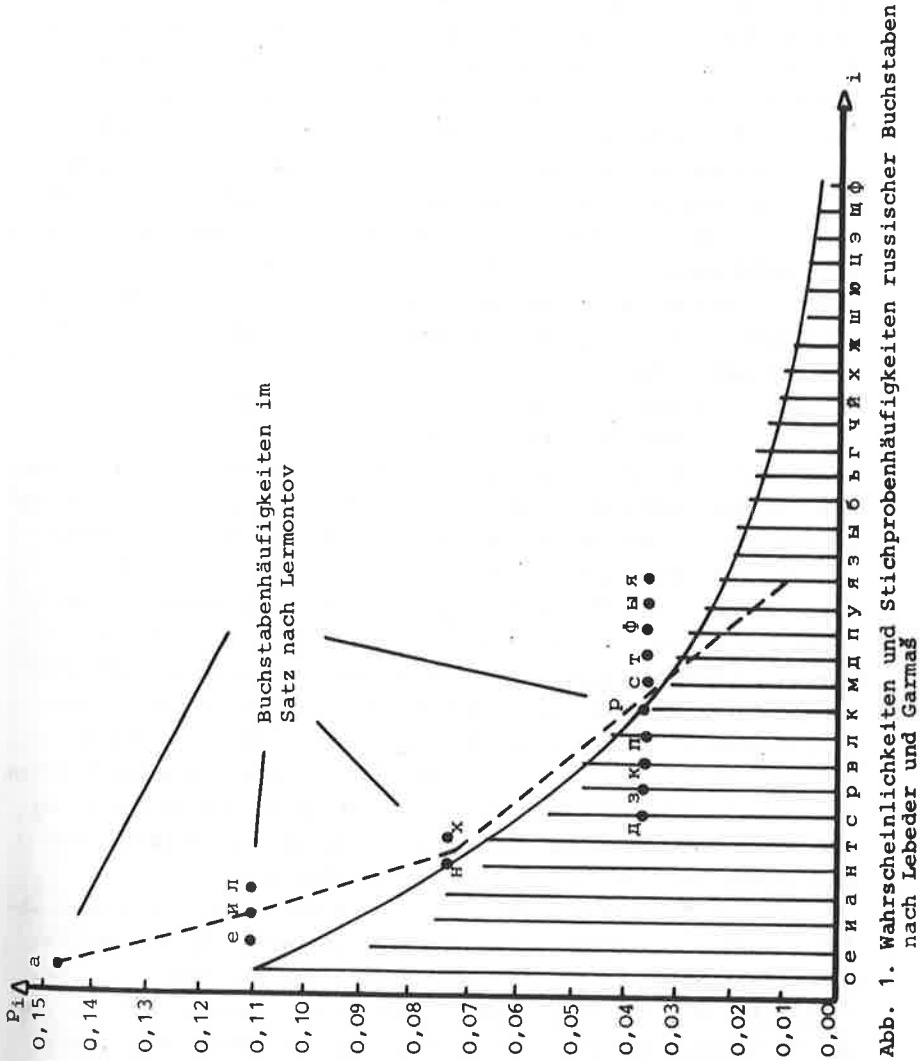


Abb. 1. Wahrscheinlichkeiten und Stichprobenhäufigkeiten russischer Buchstaben nach Lebeder und Garmaš

nete Menge auf eine Graphik eintragen und sie mit einer analytischen Kurve, z.B. mit der Zipfschen Hyperbel ausgleichen. Auf der Abb. 1 wird eine derartige Operation mit der unterbrochenen Linie dargestellt, die durch die Mitten der von den Buchstaben mit gleicher Häufigkeit gebildeten "Flächen" läuft (ungefähr auf diese Art werden empirische Häufigkeitsmengen ausgeglichen). Es ist zu betonen, daß bei Wörtern die Diskrepanz zwischen Häufigkeiten und Wahrscheinlichkeiten noch größer ist, da die mittlere Worthäufigkeit in der Stichprobe die mittlere Wahrscheinlichkeit um das Zehn- und Mehrfache überschreitet, während im untersuchten Fall der Buchstaben dieser Unterschied höchstens das Doppelte beträgt.

Ein pedantischer Leser könnte bemerken, daß die Phrase

"Я ехал на перекладных из Тифлиса"

keine zufällige Stichprobe ist. In der Tat kommt es hier zu einem für das Russische ziemlich seltenen Ereignis: 27 mal hintereinander kommt der häufigste russische Buchstabe O kein einziges mal vor (Wahrscheinlichkeit 0,1091). Die Wahrscheinlichkeit, daß bei 27 Zufallsexperimenten O kein einziges mal vorkommt ist gleich $(1 - 0.1091)^{27} = 0.044$. Dies ist kleiner als die 5%, die man traditionell als Konfidenzkoeffizient bei statistischen Musteraufgaben ansetzt. Es ist daher vernünftiger, Stichproben zu untersuchen, die den zufälligen mehr ähneln. In der Tabelle 4 findet man verkürzte Angaben aus 5 quasizufälligen (systematischen) Stichproben (es wurde jeder zehnte Buchstabe auf jeder zehnten Seite des Buches gezählt); der Umfang jeder Stichprobe beträgt 30 "Buchstabenverwendungen".

Wie aus Tabelle 4 ersichtlich, sind die Charakteristiken aller dieser Stichproben denen der Lermontovschen Phrase ähnlich und hinreichend stabil. In allen Fällen ist die Häufigkeit des häufigsten Buchstaben in der Stichprobe bedeutend höher als die der "Schätzung für die Sprache als ganze"(0.1091); alle Stichproben enthalten ungefähr die gleiche Zahl unterschiedlicher Buchstaben (von 13 bis 17). Wie in lexikalischen Stichproben, so bildet auch hier die Zahl der einmal vorkommenden Buchstaben im Durchschnitt etwa die Hälfte des Inventars (7.6/15.4 = 0.49); es

Tabelle 4

	Häufigster Buchstabe	Häufigkeit	Rel. Häufigkeit	Zahl unterschiedl. Buchstaben	Zahl der Buchstaben mit Häufigkeit 1
A.S. Puškin, Gesammelte Werke, Bd. 2 Moskva, Chud. lit. 1967	O	5	0.167	17	10
L. Nikulin, Rossii vernye syny. Moskva, Pravda 1958	O	6	0.200	13	5
I.I. Jakovlev, Korabli i verfi. Leningrad, Sudostroenie 1970	E	4	0.133	17	9
V. Koneckij, Povesti i rasskazy. Leningrad, Detskaja literatura 1978	O	6	0.200	15	7
	H	5	0.167	15	7
Mittelwert	-	-	0.173	15,4	7,6

ist zu bemerken, daß ein größerer Anteil einmaliger Buchstaben (Wörter oder anderer Einheiten) eine Art "rotes Signal" darstellt, das die Mangelhaftigkeit der Stichprobe signalisiert (leider leuchtet dieses Signal seit der Geburt der lexikalischen Statistik vergeblich!). Der einzige Unterschied zwischen den Stichproben in Tabelle 2 und der Lermontovschen Phrase ist die Tatsache, daß alle den Buchstaben O enthalten.

Die Verschiebung der empirischen Graphik einer kleinen Stichprobe in bezug auf die Menge der Wahrscheinlichkeiten, wie in Abb. 1 dargestellt, hat also keinen zufälligen, sondern einen systematischen Charakter. Die zufälligen Häufigkeits-Schwankungen jedes einzelnen Buchstabens (oder Wortes) führen bei der Rangordnung nach Häufigkeiten infolge unzureichenden Stichprobenumfanges zu einer systematischen Verschiebung, deren Größe vom Stichprobenum-

fang abhängt. Es ist klar, daß man sich bei unbeschränkter Vergrößerung der Stichprobe unbeschränkt der Menge der Wahrscheinlichkeiten nähert; folglich verringert sich die Verschiebung empirischer Beobachtungen in dem Maße, wie sich die Stichprobe vergrößert. Dies ist der Grund, warum die "Rang-Häufigkeitskurven" nicht stabil sind und für unterschiedliche Umfänge unterschiedliche Abhängigkeiten angesetzt werden müssen (Piotrowski 1975, Alekseev 1977). Wenn man naiv annimmt, daß die Häufigkeiten eine Approximation an die entsprechenden Wahrscheinlichkeiten darstellen (wie es in der klassischen Statistik, die annimmt, daß der Stichprobenumfang für die Bestimmung der Häufigkeit ausreicht, der Fall ist), dann kommt man zu dem absurden Schluß, daß die Form und die Parameter der lexikalischen Grundgesamtheit von dem Umfang der erhobenen Stichprobe abhängen. Beim Ausgleich der empirischen Häufigkeitsmenge beschreiben wir jedoch tatsächlich nichts anderes als die gegebene konkrete Stichprobe (bestenfalls Stichprobenmengen aus einer Grundgesamtheit, die jeweils einen fixierten Umfang haben).

Das Gesagte bezieht sich selbstverständlich nicht nur auf die lexikalische Statistik. Das Problem der Mangelhaftigkeit der Stichprobe entsteht bei der Analyse beliebiger linguistischer Einheiten, die über ein größeres Inventar verfügen, von den Silben an aufwärts. In allen diesen Fällen leuchtet das "rote Signal" auf in Form von umfangreichen Klassen seltener Einheiten mit derselben Häufigkeit. Dieses Signal bezeugt nicht nur den Umstand, daß im Bereich seltener Einheiten eine große zufällige Streuung vorliegt (dies könnte man noch einigermaßen ertragen!), sondern auch die Tatsache, daß in bezug auf die Charakteristiken der Grundgesamtheit alle Charakteristiken der Stichprobe systematisch verschoben sind.[1)] Diese Verschiebung werden wir im weiteren als Verschiebung erster Art bezeichnen.

Also zeigen schon rein mathematische Überlegungen die Schwäche der Grundidee der statistischen Analyse eines Objekts wie, sagen wir, Wortschatz - mittels einer relativ kleinen Stichprobe eine Vorstellung über etwas bedeutend Größeres als diese Stichprobe zu vermitteln.

2. LINGUISTISCHE GEFAHREN

Die Situation ist in der Tat noch schlimmer als oben beschrieben. Dort wurde angenommen, daß man die Resultate menschlicher Sprechtätigkeit betrachten kann, und vorausgesetzt, daß man aus ihr (repräsentative) Zufallsstichproben erheben kann. Es wird aber immer offensichtlicher, daß alle Versuche, eine repräsentative Stichprobe zu erstellen, auf Sand gebaut sind. Auch wenn das gelingen sollte, dann erzeugt die relative statistische Diversität der Texte, die unter den Bedingungen einer unzureichenden Stichprobe zu der Gesamtstichprobe vereinigt wurden, anstelle des Durchschnitts neben der im vorigen Paragraph beschriebenen Verschiebung eine zusätzliche Verschiebung der Schätzwerte. Diese Verschiebung, die dadurch entsteht, daß reale Texte oder ihre Abschnitte nicht als Zufallsstichproben aus einer einzigen Grundgesamtheit betrachtet werden können, bezeichnen wir als Verschiebung zweiter Art.

Die Auswirkungen der Verschiebung erster Art sind bekannt (Veränderung der Form der Häufigkeitskurve mit Anwachsen der Stichprobe); ebenso realisiert sich die Verschiebung zweiter Art in gut bekannten Erscheinungen. Es wurde schon oft beobachtet, daß die Vereinigung vieler Texte zu einer Stichprobe lexikalisch proportional reicher ist als jeder der sie konstituierenden Texte. Der Mechanismus dieses Phänomens ist qualitativ gesehen klar: jeder der Texte enthält einen bedeutenden Anteil von Wörtern, die nur für ihn charakteristisch sind, da sie mit dem Thema des Textes in Zusammenhang stehen. Je mehr Texte man vereinigt, desto mehr sinken die relativen Häufigkeiten dieser Wörter. Dies führt dazu, daß mit dem Anwachsen des Umfangs die Gruppe von Wörtern, welche einen vorgegebenen Anteil der Wortverwendungen in der Stichprobe abdeckt, ebenfalls anwächst, und schließlich zu einer globalen Veränderung des Verlaufs der Häufigkeitskurve führt. Im Endergebnis führt dies zu einer Vergrößerung des relativen Vokabularreichtums der zusammengesetzten Stichproben, auch wenn diese aus Texten von nominell gleichem Charakter erstellt werden. Im Unterschied zu der Verschiebung erster Art, deren Wirkungen auf den Bereich der

seltenen Wörter beschränkt bleiben, erstreckt sich die Wirkung der Verschiebung zweiter Art auf alle Häufigkeitsbereiche, einschließlich der häufigsten Wörter.

Als Beispiel führen wir aus I.Š. Nadarejšvilis unveröffentlichter Dissertation die Resultate des Vergleichs der Textabdeckung in Griboedovs Stück "Gore ot uma" (Verstand schafft Leiden) (nach Angaben von Kunickij 1896) und in modernen russischen Theaterstücken (nach Angaben des Häufigkeitswörterbuches des Russischen, Zasorina 1977) durch die fünf häufigsten Wörter an. In "Gore ot uma" decken die fünf häufigsten Wörter [ne, ja, i, on (-a, -o), v] [nicht, ich, und, er (sie, es), in] 14,4% des Textes ab. In modernen sowjetischen Theaterstücken decken die fünf häufigsten Wörter (ja, ne, i, v, byt') (ich, nicht, und, in, sein) 11,6% des Textes ab. Durch Vergleich dieser Zahlen könnte man schließen, daß in den letzten hundert Jahren die häufigsten Wörter von den dramatis personae seltener wiederholt werden und die szenische Sprache reicher geworden ist. Dieser Schluß wäre aber voreilig, da die Analyse der einzelnen Theaterstücke zeigte, daß in jedem von ihnen die fünf häufigsten Wörter einen bedeutend größeren Textanteil abdecken als die fünf häufigsten Wörter der aus ihnen zusammengesetzten Stichprobe. So sind es

in Alešins Stück "Vse ostaetsja ljudjam"	15.1%
in Arbuzovs Stücken "Gody stranstvyj"	13.2%
"Tanja"	13.9%
in Afinogenovs Stücken "Mašen'ka"	15.7%
"Strach"	13.5%

usw. Das heißt, die untersuchte Charakteristik hat sich in der russischen Dramaturgie seit Griboedov nicht verändert, obwohl die Angaben aus den zusammengesetzten Stichproben zu einem Fehlschluß verleiten könnten.

Die übliche Interpretation solcher Erscheinungen ist geradezu von einschläfernder Trivialität: jeder separate Text ist natürlich lexikalisch verhältnismäßig ärmer als "die Sprache als Ganze"! Das Problem liegt aber nicht nur darin, daß wir aufgrund des mangelhaften Umfangs unserer Stichproben die Charakteristika

der "Sprache als Ganzes" nicht ermitteln können. Denn auch wenn es möglich wäre, so erhielten wir Normen der Lexikonverwendung, die mit keinem sinnvollen Text der gegebenen Sprache etwas Gemeinsames haben. Es wären die Normen irgendeines ungeheuerlichen "Quasitextes", den man sich etwa so vorstellen kann, daß man den Inhalt einer großen Bibliothek auf Karten schreibt und dann die Karten in zufälliger Reihenfolge zusammenstellt.

Etwas Ähnliches (obwohl in kleinerem Umfang) wurde seinerzeit von Kaeding (1898) unternommen, der seinen Korpus aus den unterschiedlichsten Texten zusammengestellt hat. Soweit mir bekannt, hat seitdem niemand seine Experimente in einer solchen "reinen" Form wiederholt. Die unklar empfundene Unsinnigkeit eines derartigen Unternehmens führte zum Begriff der "Fachsprache" (Subsprache), der Gesamtheit der Texte zu einem gegebenen Thema. Dies ist im Grunde nichts anderes als ein Versuch, die Verschiebung zweiter Art, die infolge der Inhomogenität reeller Texte entsteht, zu verringern. Es wird a priori angenommen, daß die Texte in einer fachsprachlichen Stichprobe "statistisch homogen" sind, und daher alles in Ordnung ist, d.h. daß man die üblichen Stichprobenverfahren verwenden kann.

Der Übergang zur Fachsprache als Objekt der statistischen Analyse war zweifellos ein Schritt in die richtige Richtung. Ungeachtet dessen, daß sich dieser Schritt als unzureichend erwiesen hat, wie unten gezeigt wird, ist es schwer, seine grundlegende Bedeutung hoch genug einzuschätzen: mit dem Übergang zur Untersuchung der Fachsprachen wurde de facto die Unmöglichkeit (sowie die Unnötigkeit) der statistischen Analyse der Sprache als Ganzes zugegeben. Die schnell nacheinander erfolgten Anwendungen (Erstellung spezifischer Minimum-Wörterbücher, Computer-Wörterbücher, usw.) bestätigten die praktische Bedeutung dieser Richtung.

Die "Linderung", die diese Forschungsrichtung gebracht hat, führte jedoch zu dem heutigen paradoxen Zustand: die Fachsprachen vermehren sich wie Kaninchen, aber unser Wissen vermehrt sich durch ihre Untersuchung nicht. Es gelingt uns weder auf dem komparativ-typologischen noch auf dem ontologischen Feld, irgendwelche Generalisierungen durchzuführen. Jede Fachsprache ist ein

"Ding an sich", eine "Monade" ohne jegliche Wechselwirkung mit ihren Artgenossen. Die Untersuchung jeder einzelnen Fachsprache muß ausschließlich pragmatisch begründet werden, und die Ergebnisse dieser Untersuchung bilden keinen Bestandteil eines Gesamtbildes.

Zu einem kleineren Teil wird dieser Zustand hervorgerufen durch den verschobenen, trugbildähnlichen Charakter der ermittelten quantitativen Charakteristiken. Erstens, gibt es eine Verschiebung erster Art, beschrieben im vorigen Paragraphen, die es nicht erlaubt, Daten aus zwei Stichproben von unterschiedlichem Umfang zu vergleichen. Diesem Problem weicht man gewöhnlich so aus, daß man einen "runden" Umfang zum Standardumfang erklärt (z.B. die Forschungsgruppe "Statistika reči" benutzt die Standardumfänge von 50000, 100000, 200000 und manchmal 400000 Wortverwendungen). Zweitens, abgesehen von der nominellen "Homogenität" des gewählten Materials, gibt es in den Fachsprachenstichproben eine Verschiebung zweiter Art, deren Ausmaß von dem unkontrollierten Faktor der Bestimmung der "Fachsprachenmenge" abhängt.

Vergleichen wir beispielsweise die Daten von V.A. Bukovič (1969) aus dem Bereich der Computertechnik und die Daten von R.S. Melik-Gusejnova (1971) aus dem Bereich der Festkörperphysik. Beim Umfang von ungefähr 100000 Wortverwendungen ergab sich in der ersten Stichprobe ein Wortschatz von 10185 Wortformen und in der zweiten von 5542 Wortformen, d.h. reichlich die Hälfte. Welcher Faktor ruft diesen Unterschied hervor? Haben computertechnische Texte tatsächlich einen reicheren Wortschatz oder unterscheiden sie sich untereinander einfach stärker je nach der von Text zu Text wechselnden Thematik? Sind die Texte der Festkörperphysik tatsächlich lexikalisch ärmer oder hat die Forscherin einfach einen engeren thematischen Bereich gewählt? Fragen dieser Arten bleiben im Rahmen der gegenwärtig geläufigen Methodologie unbeantwortet. Es sind aber grundsätzliche Fragen: untersuchen wir reale, unabhängig von uns verlaufende Prozesse oder unsere A-priori-Vorstellungen über diese Prozesse?

Man kann sich natürlich auf den Standpunkt stellen, daß dies in der Tat ein wichtiges und kompliziertes Problem ist, aber für viele praktische Anwendungen unwesentlich, da man die Identifizierung der Fachsprache als gegeben betrachten kann. Man "nimmt die Fachsprache aus der Klammer", und alles wird gut.

Doch selbst wenn wir im Rahmen der angenommenen Definition der Fachsprache bleiben, haben wir noch immer mit der unkontrollierten Verschiebung zweiter Art zu tun, deren Größe sowohl von dem Ausmaß der "internen Inhomogenität" des erhobenen Textkorpus als auch vom Umfang dieses Korpus abhängt. Zur Illustration analysieren wir die Daten aus der "Fachsprache über Schiffstriebwerke" von K.V. Luk'janenkov (1969).

Luk'janenkovs Stichprobe von 400000 Wortverwendungen besteht aus 8 gleichgroßen Teilstichproben von jeweils 50000 Wortverwendungen. Diese Teilstichproben wurden zu Gruppen von 100000, 200000 und 300000 Wortverwendungen zusammengesetzt, so daß man den Zuwachs des Wortschatzes mit dem Anwachsen des Stichprobenumfangs beobachten kann. Beim Umfang von 50000 Wortverwendungen fand man einen Wortschatz von 4871 Lexemen; bei Vergrößerung des Umfangs auf 100000 Wortverwendungen ergab sich ein Wortschatz von 6856 Lexemen. Kann man nun annehmen, daß der Wortschatz sich nur infolge der Stichprobenzunahme vergrößert (durch Verringerung der Verschiebung erster Art) oder wurde ein Teil der erhaltenen Zunahme durch das Erscheinen völlig neuer Wörter und Termini, die in der ersten Teilstichprobe nicht erscheinen konnten, hervorgerufen (d.h. hervorgerufen durch eine Verschiebung zweiter Art)?

Betrachten wir die Stichprobe mit 100000 Wortverwendungen, als ob sie eine endliche Grundgesamtheit wäre und erheben aus ihr zufällig 50000 Wortverwendungen (ohne Zurücklegung und Mischung). Das Resultat einer derartigen Prozedur wird zuverlässig mit Hilfe der Formel von V.M. Kalinin (vgl. Formel 7 im Anhang zu dieser Arbeit) berechnet. Setzt man in diese Formel N_O = 100000, N = = 50000 und für $v_j(N_O)$ die Anzahl der bei Umfang N_O= 100000 j-malig auftretenden Wörter, so erhält man den beim Umfang von 50000 zu erwartenden Umfang des Wortschatzes; er beträgt 5271 Lexeme. Das sind 400 Lexeme mehr, als die tatsächlich in der 50000-er

Stichprobe beobachteten 4871 Lexeme [die halbe Länge des Konfidenzintervalls laut Formel 4 in Orlov (1978) ist in diesem Fall gleich 78]. D.h., die einfache Vergrößerung des Stichprobenumfangs führt zu einer spürbaren Vergrößerung des relativen Vokabularreichtums ohne Rücksicht auf den nominell gleichen Charakter der hinzugefügten Texte. Führt man diese Umrechnung auf den 50000-er Umfang an den 200000 Wortverwendungen in Luk'janenkovs Stichprobe durch, so erhält man den erwarteten Wortschatz von 5457 Lexemen, d.h. noch um 186 Lexeme mehr; die Stichprobe von 400000 Wortverwendungen ergibt für eine Stichprobe von 50000 die Erwartung von 5534 Lexemen. Das heißt, die Vergrößerung des Stichprobenumfangs erhöht den relativen Vokabularreichtum ununterbrochen. Dies würde nicht geschehen, wenn die Fachsprachentexte tatsächlich aus einer einzigen lexikalischen Grundgesamtheit stammten. Eine Umrechnung auf einen kleineren Umfang mit Kalinins Formel würde zu identischen Zahlen führen (selbstverständlich im Rahmen einer kleinen zufälligen Streuung).

Man kann sich hier fragen, was die erhaltenen Zahlen bedeuten. Wenn dieser Zuwachs des relativen Vokabularreichtums aus der Vereinigung der thematisch gleichen Texte herrührt, dann ist jeder einzelne Text offensichtlich lexikalisch ärmer als ein "Quasitext" desselben Umfangs, der zufällig aus dem gesamten Textkorpus erhoben wurde. Jedoch nicht einmal der "Quasitext" gibt die mittleren Eigenschaften des gesamten Literaturflusses über Schiffstriebwerke wieder; wir haben deutlich gesehen, daß sein "anteiliger Wortschatz" in dem Maße wächst, wie sich der Textkorpus, aus dem der "Quasitext" als eine Zufallsstichprobe erhoben wurde, vergrößert. Vergrößert man die Stichprobe 2, 10, 100 mal, so erhält man noch größere Zahlen (bei der Umrechung auf einen Umfang von z.B. 50000), und es ist überhaupt nicht bekannt, wo man aufhören soll.

Das bedeutet, daß sogar im Rahmen einer thematisch streng abgegrenzten Fachsprache ein "Abgleiten" der numerischen Charakteristiken zustande kommt, da sie von den Bedingungen der Beobachtung (vom Umfang und der Erhebungsart der Stichprobe) abhängen. Mischt man eine Menge von Texten (oder Textabschnitten) zu einer

Stichprobe, so entfernt man sich unvermeidlich ziemlich weit von einem Einzeltext. Jedoch, weh!, man nähert sich an nichts, was man als "allgemeine" oder "durchschnittliche" oder "fachsprachenspezifische" Charakteristiken bezeichnen könnte. Der Weggang vom Einzeltext erzeugt nur eine Illusion der Verallgemeinerung, tatsächlich aber untersucht man ein künstlich konstruiertes Objekt. Auch wenn wir eine Stichprobe so erheben, daß wir aus verschiedenen Texten zufällig jeweils ein Wort entnehmen, erhalten wir etwas, das zur Linguistik keine Beziehung hat: einen "Quasitext", in dem die grammatischen und die semantischen Beziehungen völlig zerstört sind und der gleichzeitig bekanntlich lexikalisch reicher ist als die jeweiligen ursprünglichen Einzeltexte. Diese Erhöhung des relativen Vokabularreichtums wird desto größer, je größer oder heterogener die erhobene Stichprobe ist (Orlov 1978). Es ist ein sehr schwacher Trost, daß die Sequenzen solcher zufällig erhobenen Wörter dem Bernoullischen oder dem Gaußschen Gesetz folgen und daß die Formel von Kalinin ideal arbeitet.

Mit einem Wort, die Erfüllung aller statistischen Maximen führt zu einem linguistischen Absurdum; die Aufrechterhaltung (auch wenn nur eine partielle) der linguistischen Realität führt zu einer Verschiebung der Schätzungen und einer größeren Ungewißheit darüber, wie man die beobachteten Zahlen auf etwas größeres als die Stichprobe extrapolieren soll. Was soll man tun?

3. DIE ALTERNATIVE: ANALYSE DER REDEPROZESSE

Man muß der Wirklichkeit ins Auge schauen und zugeben, daß eine lexikalische Stichprobe an sich keine "automatischen Verallgemeinerungen" ermöglicht. Obwohl N.S. Trubetzkojs Aussage, daß "Sprache außerhalb von Maß und Zahl" liegt, kaum richtig ist, so muß man doch heute anerkennen, daß es, von kleinen Ausnahmen ab-

gesehen, die Rede ist, die durch quantitative Kenngrößen charakterisiert wird. Es wurde zwar längst erkannt, daß "Sprache gleichzeitig Instrument und Produkt der Rede ist" (F. de Saussure 1933), aber eben die auf de Saussure basierende Tradition der Untersuchung von Redephänomenen wird aus dem Rahmen der Linguistik entfernt und der Philologie, der Psychologie, der Soziolinguistik, der Kontextologie usw. usw. überlassen. Die Natur nimmt jedoch keine Rücksicht auf die künstlichen Trennwände, die wir mit Leidenschaft aufbauen, um uns im Grunde genommen, vor ihrer Komplexität zu verstecken.

Wenn uns die Sprache durch die Realität der Rede gegeben wird, so soll man diese Realität ehrlich untersuchen und nicht versuchen, sie in Übereinstimmung mit unseren A-priori-Konzeptionen auszubessern. Wenn wir sinnvolle Zahlen erhalten wollen, dann müssen diese Zahlen zuallererst **linguistisch sinnvolle Objekte** charakterisieren. Wir müssen reale, **außerhalb** von uns verlaufende Redeprozesse untersuchen, ohne uns in ihren natürlichen Verlauf einzumischen.

Mit quantitativen Methoden soll man vor allen Dingen den **individuellen Text** untersuchen, d.h. ein solches Gebilde, das durch einen einzigen Akt der "Redeschöpfung" erzeugt wurde und für einen einzigen Akt der Rezeption bestimmt ist. Eben auf diesem Wege kann die Linguistik Verbindungen zur Psychologie und anderen Disziplinen finden, die die Phänomene der Kommunikation und der Informationsverarbeitung untersuchen. Dies schließt durchaus nicht aus, daß man sowohl Textmengen als auch Textabschnitte und das, was man heute Informationsfluß nennt, untersucht; aber die Methoden, der Sinn und die Ziele solcher Untersuchungen müssen anders, von den heutigen verschieden sein.

Der Verfasser hat, begreiflicherweise, keine Absicht, im Rahmen eines kleinen Aufsatzes ein grandioses Programm der Veränderung der Ziele, der Methoden und des Sinns einer ganzen wissenschaftlichen Disziplin aufzustellen. Sein Minimalprogramm ist die Erkennung der Notwendigkeit eines derartigen Umbaus. Die Kritik soll aber konstruktiv sein und man soll zumindest umreißen, was man als Ersatz anbietet.

Die quantitative Analyse linguistischer Erscheinungen stellt im Grunde nur die ersten Schritte dar und die exakte Formulierung allumfassender "Konzeptionen" ist offensichtlich noch vorzeitig. Daher ist es vernünftig, die Probleme an Beispielen zu verdeutlichen, die möglichen Verallgemeinerungen aufzuzeigen und ihre Realisierung der Zukunft zu überlassen.

Betrachten wir nur eine einzige, jedoch äußerst wichtige linguostatistische Charakteristik, den "Zipfschen Umfang" Z (vgl. Orlov 1976, 1978b) in den Texten des "Neuen Testaments" (NT), wie sie von R. Morgenthaler (1958) analysiert wurden. Kennt man Z (und eine, auch wenn nur grobe Schätzung des häufigsten Wortes p_1), dann kann man den Umfang des Wortschatzes v(N,Z) beim Textumfang N berechnen als

$$v(N,Z) = v(Z)\,\frac{\ln X}{X-1} \qquad (2)$$

wo $v(Z) = \frac{Z}{\ln(Zp_1)}$ und $X = \frac{Z}{N}$ ist.

Die Kenntnis von Z erlaubt außerdem, die Konstanten K und B in der Zipf-Mandelbrotschen Formel

$$p_i = \frac{K}{B+i} \qquad (3)$$

zu berechnen, wo

$$K = \frac{1}{\ln(Zp_1)} \quad \text{und} \quad B = \frac{K}{p_1} - 1$$

ist. Diese Formel beschreibt die Häufigkeiten der häufigen und mittelhäufigen Wörter wie auch den Bereich seltener Wörter in Form des Häufigkeitsspektrums $v_m(N,Z)$, d.h. der Zahl unterschiedlicher Wörter, von denen jedes beim Umfang N jeweils m-mal vorkommt:

$$v_1(N,Z) = \frac{v(Z) - Xv(N,Z)}{1 - X} \qquad (4a)$$

$$v_{m+1}(N,Z) = \frac{v_m(N,Z) - v_m(Z)}{1 - X} .$$

(Bei $X \approx 1$ ist die rekurrente Berechnung laut (4a) bei großem m nicht stabil, daher empfiehlt sich, in diesem Fall die Darstellung dieser Rechenprozedur in Form der Reihe

$$v_m(N,Z) = v(Z) \sum_{j=m} \frac{(1 - X)^{j-m}}{j(j + 1)} \qquad (4b)$$

zu benutzen. Bei $X \approx 1$ konvergiert diese Reihe sehr schnell).

Kennt man also den "Zipfschen Umfang" des Textes, so kann man sowohl die Statik als auch die Dynamik seiner quantitativen Organisation beschreiben. Diese Fragen und die Methode der Berechnung des Wertes Z behandeln wir im Anhang und in Orlov (1976, 1978 beide in diesem Band), hier erwähnen wir nur eine unmittelbar linguistische Interpretation der Größe Z: Man kann sie als Kenngröße des relativen Vokabularreichtums betrachten (Orlov 1978; Nadarejšvili & Orlov 1978). Erhebt man aus zwei Texten mit unterschiedlichem Wert von Z **gleiche** Stichproben, dann findet man einen **größeren** Wortschatz in der Stichprobe aus dem Text, dessen Z größer ist (unter der Bedingung, daß p_1 in beiden Texten ungefähr gleich ist, was für Texte desselben Genres in einer Sprache fast immer der Fall ist).

Das "Neue Testament" haben wir deswegen gewählt, weil MORGENTHALERS Arbeit einen relativ seltenen Fall einer Materialauszählung darstellt, der eine sinnvolle quantitative Analyse ermöglicht[2]; außerdem ermöglicht die allgemeine Kenntnis dieses Literaturdenkmals und seine relativ gute Erforschung, sich nicht nur auf rein formale Schlußfolgerungen zu beschränken, sondern auch eine inhaltliche Interpretation der numerischen Charakteristiken zu geben.

In Tabelle 5 findet man die grundlegenden Daten des Neuen Testaments: den Textumfang N, den Wortschatzumfang v und den daraus berechneten Wert des "Zipfschen Umfangs" Z (der für die Berechnung notwendige Wert von p_1 wurde aufgrund des gesamten NT als 0.1186 bestimmt). In der vorletzten Spalte findet man außerdem das Verhältnis des "Zipfschen Umfangs" des gegebenen Textes zu seinem tatsächlichen Umfang X = Z/N; wegen der Anschaulichkeit wird in der letzten Spalte der reale Wortschatz des Textes mit Hilfe von (2) auf den Standard von 10000 Wortverwendungen umgerechnet.

Man kann leicht sehen, daß nur zwei Texte (dazu noch sehr kurze) einen höheren relativen Vokabularreichtum haben als das NT als ganzes (Judasbrief, Titusbrief). Der mittlere Wert von Z für alle Texte ist gleich 8966 (es wurden nur separate Texte, die keine anderen in sich enthalten, in Betracht gezogen, d.h. alle Texte unter der horizontalen Trennungslinie in Tab. 5). Da diese Texte unterschiedlich lang sind, kann man annehmen, daß der einfache Mittelwert im gegebenen Fall keine so gute Kenngröße ist; daher wurde der gewichtete Mittelwert

$$\frac{\Sigma ZN}{137389}$$

berechnet. Er ergab 8173, d.h. er stand praktisch nah beim arithmetischen Mittelwert.

Berechnet man laut (2) den Wortschatz eines Textes von 10000 Wortverwendungen, dessen Z = 8173 ist, so erhält man $v(10^4,8173) = 1312$, was dem mittleren Wert der Zahlen in der letzten Spalte, 1271, sehr nah steht. Ein Teil der Differenz wird durch die nichtlineare Beziehung zwischen N und v(N,Z) hervorgerufen. Gleichzeitig prognostiziert Z, das man für den ganzen Textkorpus des NT bestimmt hat (17660), beim Umfang von 10^4 Wortverwendungen einen wesentlich größeren Wortschatz von $v(10^4,17660) = 1714$.

So bekommen wir gewisse "Mittelwerte" für den relativen Vokabularreichtum im NT, die auf unterschiedliche Weisen berechnet wurden und miteinander nicht übereinstimmen. Welche Realität verbirgt sich hinter ihnen?

Tabelle 5

Texte des NT		Trad. Autor oder Zugehörigkeit	Gruppe nach Šajkevič	Text oder Stichprobenumfang N	Wortschatz v	Zipfscher Umfang Z	$x = \frac{Z}{N}$	$v(10^4,Z)$ nach (2)	$v(10^4)$ nach (7)
Gesamttexte	NT	-	-	137389	5436	17660	0,129	1714	1640
	Paulusbriefe	Paulus	-	32349	2648	12452	0,385	1526	1530
	Alle Briefe	B	-	7583	1271	11224	1,480	1472	-
	3 Johannesbriefe	B	3	2601	302	835	0,321	492	-
Einzeltexte	Johannes	Ev.	1	15416	1011	2855	0,185	860	855
	an Philemon	Paulus	4	335	141	3023	9,024	881	-
	an die Thessalonicker 2	Paulus	3	821	250	3062	3,730	886	-
	Apokalypse	Johannes	1	9834	917	3357	0,341	921	922
	an die Thessalonicker 1	Paulus	3	1475	366	3489	2,365	936	-
	an die Epheser	Paulus	3	2418	529	4905	2,029	1077	-
	an die Kolosser	Paulus	3	1575	431	5731	3,639	1146	-
	an die Galater	Paulus	3	2229	526	5836	2,618	1154	-
	an die Korinther 1	Paulus	3	6811	967	6009	0,882	1157	1141
	an die Philipper	Paulus	3	1629	448	6031	3,702	1169	-
	an die Korinther 2	Paulus	3	4469	792	6157	1,378	1178	-
	an die Römer	Paulus	3	7105	1068	7407	1,043	1265	1283
	Markus	Ev.	1	11242	1345	7564	0,673	1275	1272
	Mathäus	Ev.	1	18305	1691	7672	0,419	1282	1308
	Lucas	Ev.	1	19428	2055	11633	0,599	1490	1473
	Apostelgeschichte	Lukas?	2	18382	2038	12237	0,666	1517	1515
	an die Hebräer	Paulus?	3	4951	1038	12500	2,525	1528	-
	Petrusbrief 2	B	3	1098	401	12527	11,41	1529	-
	Petrusbrief 1	B	3	1678	545	13836	8,246	1582	-
	Johannesbrief	B	3	1749	560	13968	7,986	1587	-
	an Timotheus 2	Paulus?	4	1236	458	16986	13,74	1693	-
	an Timotheus 1	Paulus?	4	1588	541	17294	10,89	1703	-
	Judasbrief	B	3	457	227	18462	40,40	1739	-
	an Titus	Paulus	4	658	303	20794	31,60	1805	-

Anmerkung: In der letzten Spalte $v(10^4)$ stehen die nach (7) berechneten erwarteten Werte des Vokabularreichtums bei N = = 10000. Diese Zahlen stellen die exakten mathematischen Erwartungen des Vokabularreichtums für Zufallsstichproben aus neutestamentlichen Texten dar. Während die Rechnungen in der vorletzten Spalte, $v(10^4,Z)$, nach (2) unter der Annahme, daß beim Umfang Z der Text exakt dem Zipf-Mandelbrotschen Gesetz folgt, durchgeführt wurden, so erfolgten sie in der letzten Spalte aufgrund des beobachteten Häufigkeitsspektrums $\hat{v}_j(N_o)$. Da die Formel (7) von Kalinin höchstens eine zweifache Prognose "nach vorne" ermöglicht, kann man für einen großen Teil der Texte diese Zahlen nicht berechnen. Der Unterschied zwischen $v(10^4)$ und $v(10^4,Z)$ zeigt den Genauigkeitsgrad des theoretischen Modells. Wie man in der Tabelle sieht, ist dieser Unterschied sehr klein und folglich kann man die theoretischen Berechnungen nach (2) als zufriedenstellend betrachten.

Z = 8173 drückt die Realität einer Menge von einzelnen, voneinander isolierten Texten aus. Das heißt, wenn wir jeden einzelnen neutestamentlichen Text bis zum Umfang von 10^4 bringen würden (abgesehen von der physischen Unmöglichkeit, dies für kürzere Texte zu tun), so würden wir Wortschätze, die um 1300 schwanken, erhalten (vgl. die letzte Spalte der Tab. 5). Z = = 17660 stellt eine ganz andere Realität dar: die Realität der Texte, die miteinander in eine gewisse "chemische Reaktion" der wechselseitigen Durchdringung getreten sind. Wir erhalten einen Wortschatz von ungefähr 1700 Wörtern, wenn wir eine zufällige Stichprobe von 10000 Wortverwendungen aus dem gesamten Text des NT erheben, den wir auf Zetteln ausschreiben und vermischen. Es ist klar, daß die Ausschreibung von Zetteln (sowie die Zahl 1700) an sich keinen philologischen oder linguistischen Sinn haben wird. Sie bekommt aber diesen Sinn, wenn es die Möglichkeit gibt, die Charakteristiken der gesamten Stichprobe mit den Charakteristiken einzelner Texte, die in die Stichprobe eingingen, zu vergleichen. Das Anwachsen (in unserem Fall) des relativen Vokabularreichtums von Z = 8173 auf Z = 17660 ist eine Charakteristik der "gegenseitigen Inhomogenität" der neutestamentlichen Texte. Dies ist der Effekt des Anwachsens des relativen Vokabularreichtums, vor dem sich die von K.F. Luk'janenkov analysierten Stichproben "nicht retten" konnten. Wenn man alle Stichproben (oder Texte) als aus einer lexikalischen Grundgesamt-

heit (in der Art eines Häufigkeitswörterbuches) stammend betrachten könnte, dann gäbe es keinen Zuwachs des relativen Vokabularreichtums in Abhängigkeit von der Art der Stichprobenerhebung. Je mehr sich die für die gemeinsame Stichprobe gewählten Texte untereinander unterscheiden, desto mehr wächst der relative Vokabularreichtum in seiner Gesamtheit verglichen mit den einzelnen Texten.

Ein sehr großes Anwachsen dieser Art findet man im Häufigkeitswörterbuch des Čechischen (HC) (vgl. Orlov 1978). Der mittlere Wert von Z für 66 Texte, die für dieses Wörterbuch gewählt wurden, beträgt etwa 52500 [dem entspricht der Wortschatz einer zehntausender Stichprobe mit $v(10^4, 52500) = 2622$]. Der Z-Wert für den ganzen Korpus (N = 1623527, $p_1 = 0.0413$, v = 54486) ist gleich 210000 (dem entspricht der Wortschatz einer Zehntausender-Stichprobe mit 3525 Wörtern). Man kann also zuverlässig nachweisen, daß der hohe relative Vokabularreichtum im HC keinen hohen relativen Vokabularreichtum in čechischen Texten, sondern einen generell hohen Grad der "gegenseitigen Inhomogenität" der im Korpus vereinigten Teilstichproben widerspiegelt. Eine quantitative Kontrolle dieser Inhomogenität ist nur post factum möglich und zwar unter der Bedingung, daß man nicht nur Daten über die Gesamtstichprobe sondern auch für die einzelnen sie konstituierenden Texte hat[3)].

So bleiben die Daten der Stichprobe von K.F. Luk'janenkov oder die Daten des "Häufigkeitswörterbuchs des Russischen" (HR) "Dinge an sich". Der aus dem HR erhaltene Wert Z = 150000 [Wortschatz $v(10^4, 150000) = 3330$] ist zweifellos größer als der Mittelwert der einzelnen unvermischten russischen Texte. Man kann aber nur raten, um wieviel er größer ist, d.h. der relative Vokabularreichtum des HR ist eine Charakteristik nur des konkreten Textkorpus, der zur Erstellung des HR gewählt wurde.

Die Berechnungen aufgrund einer vermischten (vereinigten) Stichprobe erhalten also einen Sinn nur dann, wenn die Charakteristiken einzelner zusammenhängender Texte (und nicht "repräsentativer" Teilstichproben) aus denen diese Stichprobe besteht, bekannt sind. Dadurch unterscheidet sich die lexikalische

Statistik prinzipiell von der üblichen Statistik, in der der Mittelwert eines inhomogenen Materials unter der Bedingung der Erhebung einer sogenannten proportionalen Stichprobe stabil sein kann. Die statistische Inhomogenität linguistischer Objekte erzeugt eine zusätzliche Verschiebung (zweiter Art), deren Berechnung und Kontrolle nur dann möglich ist, wenn die Daten über die Texte, aus denen die Stichprobe besteht, erhalten bleiben.

Beim Übergang zu einer inhaltlichen Analyse der numerischen Charakteristiken von neutestamentlichen Texten möchten wir darauf aufmerksam machen, daß alle Einzeltexte in der Tab. 5 nach der wachsenden Größe des "Zipfschen Umfangs" Z, d.h. nach der Größe seines relativen Vokabularreichtums geordnet sind. Man kann leicht sehen, daß diese rein formale Anordnung gleichzeitig eine überaus sinnvolle Anordnung ist.

Man findet nacheinander die sogenannten synoptischen Evangelien (Markus, Matthäus, Lukas), die durch gemeinsame Inhalte, Stile und Quellen verbunden sind; ihnen schließt sich unmittelbar die Apostelgeschichte an, deren Verfasser angeblich Lukas war. Zusammen steht der größte Teil der katholischen Briefe und ein erheblicher Teil der Paulusbriefe sowie auch Texte, die mit dem Namen von Johannes verbunden sind: Evangelium, Apokalypse und drei Briefe, die am Anfang der Liste, d.h. im Bereich des niedrigen relativen Vokabularreichtums stehen. Gleichzeitig aber weicht eine Anzahl von Texten von diesen Hauptgruppierungen ab.

Der hohe relative Vokabularreichtum des Briefes an die Hebräer und die extrem hohe Konzentration der Briefe an Timotheus und Titus (der sogenannten Pastoralbriefe) unterscheiden diese Texte scharf von der Hauptgruppe der Paulusbriefe. Die beiden Briefe an Timotheus und der Brief an Titus sind jedoch nach allgemeiner Ansicht (einschließlich der der Theologen) offensichtlich spätere Imitationen; begründete Zweifel gibt es auch bezüglich des Briefes an die Hebräer (vgl. z.B. Vrede 1908; Robertson 1959). Die starke Korrelation der häufigsten Wörter in den Pastoralbriefen veranlaßte A.J. Šajkevič (1979), diese Texte

als eine separate Gruppe abzuheben (konventionell bezeichnet mit der Zahl 4), was offensichtlich ihre besondere Herkunft bezeugt. Šajkevičs Gruppierung der Texte nach der Korrelation häufiger Wörter untereinander stimmt mit unserer Gruppierung gut überein, obwohl sie die Unterschiede zwischen der Hauptgruppe der Paulusbriefe und den Katholischen Briefen nicht feststellt. In einigen Fällen kann also die abstrakte und verallgemeinerte Kenngröße, der relative Vokabularreichtum, zu einem feineren Differenzierungsmerkmal als die ihrer Natur nach inhaltliche Korrelation der Verwendungshäufigkeiten der häufigsten Wörter werden (in diesem Fall geht es nicht um die Konkurrenz verschiedener Methoden, sondern um ihre gegenseitige Ergänzung und Bereicherung).

Es ist auch interessant die Beziehung des "Zipfschen Umfangs" des Textes zu seiner tatsächlichen Länge, X = Z/N, zu untersuchen. Die Ähnlichkeit der Häufigkeitsstruktur einzelner literarischer Texte mit der kanonischen Form des Zipf-Mandelbrotschen Gesetzes wurde schon öfters erwähnt (vgl. z.B. Orlov 1971, 1974, 1976; Arapov & Efimova & Šrejder 1975); die Nähe von X an 1 kann als das Maß der Übereinstimmung der Häufigkeitsgraphik des Textes mit der Hyperbel angesehen werden. Die Gruppe der neutestamentlichen Texte, deren X im Intervall <0,5; 2> liegt (dies ist der Bereich, wo das Zipf-Mandelbrotsche Gesetz gut erfüllt wird, vgl. Abb. 2) besteht aus Kor. 1, Kor. 2, Röm., Mk., Luk., Apostelgeschichte. Es sind die großen Texte, die 49% des NT ausmachen. Erweitert man das Intervall von X auf <0,25; 4> (die Grenzen dieses Bereichs entsprechen verhältnismäßig geringen Abweichungen der Häufigkeitsgraphik von der Hyperbel (3)), so schließen sich dieser Gruppe die drei Briefe von Johannes, Thess. 2,1, Apo., Ephes., Koloss., Gal., Phil., Matth., Hebr. an. Zusammen mit der vorigen Gruppe bilden sie über 82% des Textes. Außerhalb dieser Gruppe bleibt nur ein großer Text, nämlich das Evangelium von Johannes (11% des NT), die restlichen 8 Texte bilden weniger als 7% des NT. Diese Zahlen stimmen gut überein mit der Verteilung der Größe X_0 für unterschiedliche literarische Texte, die in Orlov (1978, Abb. 3) analysiert wur-

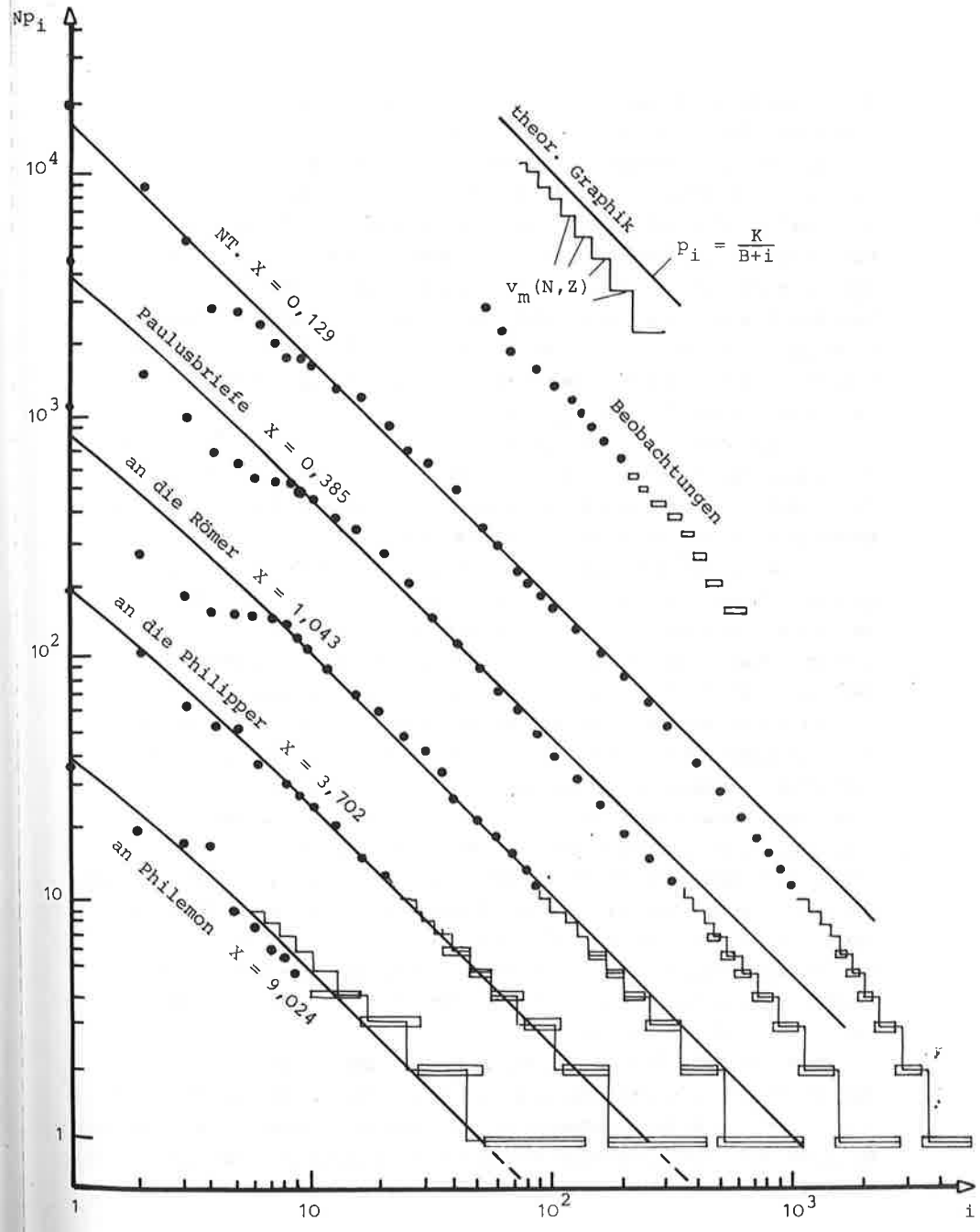


Abb. 2

den; ebenda wurde auch die Existenz eines "Schweifes" bemerkt, den kurze Texte mit hohem X_0 bilden.

Die hauptsächliche Textmasse des NT, die die maximale künstlerisch-bildliche und ideologisch-propagandistische Ladung trägt, erfüllt also zufriedenstellend mit jedem einzelnen Text das Zipf-Mandelbrotsche Gesetz. Besonders eindrucksvoll ist es bei den Briefen von Paulus: die Ordnung nach der Textlänge und die Ordnung nach dem "Zipfschen Umfang" fallen praktisch zusammen. Wenn man die unechten Briefe an Timotheus und Titus ausschließt, dann ergibt sich der Spearmansche Rangkorrelationskoeffizient $r_s = 0.89$.

Der allgemeine Charakter des relativen Vokabularreichtums in neutestamentlichen Texten stimmt gut mit dem von Texten, die für auditive Wahrnehmung bestimmt sind, nämlich von dramatischen Werken, überein. Beispielsweise in Corneilles Stücken (nach Ch. Muller 1968) schwankt Z zwischen 4000 und 9920 (bei der Längenschwankung von 13807 bis 20268; das entsprechend K schwankt zwischen 0.248 und 0.642). In Shakespeares Werken ist Z etwas größer, aber hier wirkt sich, offensichtlich, die Zählungsart aus: In Spevack (1968), aus dem wir die Zahlen entnommen haben, wurden nicht Lexeme, sondern Wortformen gezählt. Bei Schwankungen der Länge von 14369 bis 29551 bewegt sich Z zwischen 20000 und 53000; entsprechend läuft X von 0.96 bis 2.58. Die beträchtliche Spannweite der Textlängen (im Unterschied zu den Stücken von Corneille) erlaubte es, den Korrelationskoeffizienten zwischen logZ und logN zu berechnen: Es ergab sich r = 0.42 mit dem 95-%-Konfidenzintervall von 0.11 bis 0.65 (wenn der Nullpunkt nicht in das Intervall fällt, so kann man die Korrelation als statistisch signifikant betrachten). Der analoge Korrelationskoeffizient einer größeren Gruppe unterschiedlicher Texte ergab 0.77 (vgl. Orlov 1978).

Die Existenz der Beziehung zwischen dem "Zipfschen Umfang" des künstlerischen Textes und seiner vollen Länge stellt offensichtlich die rätselhafteste Erscheinung der linguostatistischen Analyse dar. Das annähernd parallele Anwachsen der Textlänge und seines relativen Vokabularreichtums kann man keineswegs z.B.

durch das oben analysierte Anwachsen des relativen Vokabularreichtums der Stichprobe mit dem Anwachsen ihres Umfangs erklären, denn dieses Anwachsen bleibt hinter dem Anwachsen des Umfangs wesentlich zurück. Schon aus den Daten des NT sieht man, daß die Vereinigung aller seiner Texte zur Erhöhung von Z auf 17660 führt, während sich der Stichprobenumfang zehnfach auf 137389 vergrößert; in Orlov (1978: 118, Tab. 5) wird auf das Fehlen einer signifikanten Korrelation zwischen dem "Zipfschen Umfang" einer Stichprobe und ihrer tatsächlichen Länge hingewiesen.

Die Tatsache, daß die "innere Organisation" eines literarischen Werkes (die in seinem "Zipfschen Umfang" fixiert ist) mit seiner Länge zusammenhängt, kann man (vorläufig) nur mit der besonderen Relevanz dieses Zusammenhanges für die "optimale Rezeption" jener einmaligen Informationsmenge, die der literarische Text darstellt, erklären. Um diesen Zusammenhang zustandezubringen, muß der Autor die gesamte Häufigkeitsstruktur des Textes unter quantitativer Kontrolle halten, er muß ununterbrochen registrieren, wie viele Wörter er schon gebraucht hat, wie viele davon unterschiedlich sind, wie viele von den unterschiedlichen er 1 mal, 2 mal usw. verwendet hat usw. Es ist nicht schwer zu beweisen, daß eine solche Arbeit nicht auf der bewußten Ebene verläuft. Aber die Tatsache, daß sie überhaupt auf der unterbewußten Ebene verlaufen muß, kann man nicht einfach nur durch die Wichtigkeit, ja die Notwendigkeit, sondern eigentlich erst durch die tiefe organische Natur dieses Prozesses erklären. Dies gibt uns den Schlüssel zum Verständnis derjenigen Aspekte der Rezeption und der Verarbeitung der Information durch den Menschen, die bisher hinter sieben Siegeln verschlossen waren und von uns einfach nicht erfaßt wurden (sondern unserem Bewußtsein vielleicht lediglich als die Empfindung einer "Vollständigkeit", "Abgeschlossenheit", "Harmonie" des künstlerischen Ganzen vermittelt wurden; ausführlicher vgl. Orlov 1974, 1978; Boroda, Orlov 1978).

Die ermittelte Anordnung neutestamentlicher Texte nach wachsendem relativen Vokabularreichtum in Tab. 5 hat schließlich noch

einen anderen Aspekt. Es ist nicht schwer zu erkennen, daß diese Ordnung in groben Zügen mit den heutigen Vorstellungen über die relative Chronologie dieser Texte übereinstimmt. Am Anfang der Liste stehen Texte, die mit dem Namen von Johannes (einschließlich 3 Briefe) verbunden sind - offensichtlich die ältesten Texte[4]. Dann folgen die Briefe von Paulus, die den synoptischen Evangelien und der Apostelgeschichte zeitlich vorangehen; die Liste endet mit den Katholischen Briefen und den Briefen an Timotheus und Titus, die zu der späteren Epoche der Entwicklung des Christentums, in der die kirchliche Organisation zu funktionieren anfing, gehören.

Dieses Anwachsen des relativen Vokabularreichtums mit der Zeit wurde bereits in Orlov (1978) gezeigt, wo es im Zeitraum der zwei letzten Jahrtausende untersucht wurde. Der Wert von Z wächst von einigen Hunderten und wenigen Tausenden (russische Bylinen, biblische Texte) zu Zehntausenden (die klassische Prosa des 19. Jhdt.) und Hunderttausenden (Šolochov, Joyce). Man kann über diese Erscheinung unterschiedliche Hypothesen aufstellen; wir führen die plausibelsten auf und besprechen sie kurz.

A. Die Entwicklung der literarischen Form?

Da zwischen dem Textumfang und dem "Zipfschen Umfang" eine Korrelation besteht, wird der Zuwachs des relativen Vokabularreichtums von dem parallelen Anwachsen der Textumfänge begleitet, und dies kann als ein Argument zugunsten der Hypothese A betrachtet werden. Es gibt hier aber Ausnahmen: "zyklische" Texte, die vom Autor in kleinere oder größere Teile aufgeteilt wurden (Kapitel, Bücher, Bände usw.) (Orlov 1971; 1978). Wenn der Umfang eines großen Textes seinen "Zipfschen Umfang" stark übersteigt, dann haben seine vom Autor bestimmten Teile in der Regel Umfänge derselben Ordnung wie Z oder etwas kleinere. Gerade an solchen kann man das Anwachsen des relativen Vokabularreichtums sozusagen in reiner Form verfolgen. Beispielsweise Cervantes' "Don Quixote" hat den Umfang N = 357255 und ist in 99 Kapitel aufgeteilt (nach Gomez 1962), d.h. der mittlere Kapitelumfang beträgt etwa 3600 Wortverwendungen. Der "Zipfsche Umfang" des ganzen Romans ist 20750. In Bezug auf jedes einzelne Kapitel ist dieser Wert wegen ihrer Heterogenität sicherlich zu hoch (von Kapitel zu Kapitel ändern sich mindestens die Orte und die zweitrangigen Personen), so daß man erwarten kann, daß bei der Bestimmung von Z für einzelne Kapitel die übliche Übereinstimmung zwischen Z und N vorgefunden wird und die beiden sich voneinander höchstens um das 2-3-fache unterscheiden werden. In Tolstojs "Auferstehung" ist N = 145000 und Z = 53000 (der Roman hat 3 Teile, vgl. Orlov 1978), in Joyce' "Ulysses" ist N = = 260430 und Z = 341400 (χ = 1.31). Wenn wir (zwecks Anschaulichkeit, wie oben) diese abstrakten Kenngrößen auf den Wortschatz einer Stichprobe mit Umfang 10000 übertragen, so erhalten wir: "Don Quixote" 1774, "Auferstehung" 2565, "Ulysses" 3682. Obwohl man Texte in unterschiedlichen Sprachen ausgezählt hat und die Lexeme wohl nicht mit äquivalenten Methoden identifiziert wurden, besteht kein Zweifel an der allgemeinen Tendenz des Anwachsens des relativen Vokabularreichtums in "dicken Romanen".

B. Bereicherung der Sprache um neue Wörter und Begriffe? (D.h. das Anwachsen des relativen Vokabularreichtums spiegelt die allgemeine Entwicklung des menschlichen Denkens wider)?

Auf den ersten Blick scheint die Hypothese verlockend zu sein, aber es reicht, das in den achtziger Jahren des vorigen Jahrhunderts herausgegebene Wörterbuch von Vladimir Dall (von etwa 200000 Wörtern) durchzublättern, um sich von der Existenz umfangreicher Schichten archaischer Lexik zu überzeugen; der heutige russische Muttersprachler kennt davon höchstens 1/5-1/4; auch in Sprachen, die sich im Anfangsstadium der Entwicklung befinden, findet man ein "detailliertes" lexikalisches System, das mit einer geringen Polysemie und Homonymie in diesen Sprachen verbunden ist (vgl. z.B. Polikarpov 1976).

C. Veränderung des Kanals der Sprachrezeption?

Die frühesten literarischen Gebilde, so wie russische Bylinen oder biblische Texte sind entweder Niederschriften münd-

lichen Schaffens, oder sie waren zum mündlichen Vortrag vor analphabetischem Auditorium bestimmt (vgl. z.B. Robertson 1959). Mit der kulturellen Entwicklung der menschlichen Gesellschaft werden jedoch immer mehr geschriebene Texte durch den visuellen Kanal durch Lesen "für sich" rezipiert. Da der visuelle Kanal eine wesentlich größere Durchlaßkapazität als der auditive hat, konnte sich auch die informationelle Belastung einfach als Folge sich eröffnender Möglichkeiten vergrößern. Indirekte Bestätigungen dieser Hypothese sind die Abwesenheit einer ähnlichen Erscheinung in der Musik (Boroda 1979), die ausschließlich für auditive Rezeption bestimmt ist, und die niedrigen Kenngrößen des relativen Vokabularreichtums in gesprochener Sprache und in Theaterstücken. Zugunsten dieser Hypothese spricht auch der Umstand, daß der relative Vokabularreichtum offensichtlich nicht von der absoluten Zeit der Textentstehung abhängt, sondern sozusagen von der "kulturellen Zeit" der Gemeinschaft, in der und für die der Text geschaffen ist. So haben beispielsweise Senecas Traktate (Seneque 1968) ein Z der Ordnung von 20000 - 30000, was zehnmal so groß ist wie in zeitgenössischen frühchristlichen Texten.

Eine Antwort auf alle diese Fragen wird erst dann möglich, wenn wir über Daten aus einer großen Menge von Texten verfügen, die einzeln, aber nach einheitlicher Methodik, nach einer Standarddefinition der Lexeme durchgezählt werden können (bisher machte sich zwischen verschiedenen Arbeiten eine beträchtliche Uneinigkeit bemerkbar; noch besser ist es wohl, einen und denselben Text mit Hilfe mehrerer verschiedener Einheiten, z.B. Lexeme und Wortformen durchzuzählen. Die Veränderung des "Zipfschen Umfangs" in Abhängigkeit von der Zähleinheit kann als Maß der Analytizität (Synthetizität) der Sprache dienen. Es wäre interessant, die Beziehung[5] $Z_{Wortformen}/Z_{Lexeme}$ für unterschiedliche Texte über einen großen Zeitraum hin zu untersuchen. Dies gäbe eine ergänzende Information über die Richtung der Sprachevolution).

Dies ist, freilich, eine sehr umfangreiche Arbeit, aber sie würde uns (neben Ergebnissen, die man heute noch nicht voraus-

sehen kann) eine rein quantitative glottochronologische Skala liefern. Genauer, sie würde eine ganze Reihe solcher Skalen für verschiedene Kulturen und (innerhalb dieser) für unterschiedliche Stile, Genres usw.[6] liefern. Der letzte Umstand erweckt die Hoffnung, daß man diese Aufgabe "in Teile" zerlegen kann, was die Chancen ihrer Lösung zweifellos vergrößert.

Man kann die Bedeutung dieses Problems für die Linguistik im allgemeinen und für die Soziolinguistik und Psycholinguistik kaum überschätzen. Es eröffnen sich Möglichkeiten auch für die Lösung angewandter Probleme wie beispielsweise die Datierung der Verfasserschaft (man muß dabei aber berücksichtigen, daß erstklassige Schriftsteller wie Puškin, Tolstoj, Joyce den relativen Vokabularreichtum ihrer Texte in Übereinstimmung mit dem Textumfang und mit ihren künstlerischen Zielsetzungen zuverlässig regulieren; vgl. Orlov 1971, 1974, 1978).

Es können sich offensichtlich noch viele andere Wege eröffnen, jedoch nur unter der Bedingung, daß die Resultate der Berechnungen eine von dem Forscher unabhängige Realität repräsentieren. Sogar in den Fällen, wo die Aufteilung des linguistischen Materials in einzelne Texte offensichtlich sinnlos ist (beispielsweise bei der Untersuchung diverser "Redeflüsse" wie Anfragen, Anforderungen, Annotationen, Referate, Dienstgespräche usw. usw.), muß man nach der Ausklammerung eines Teils eines derartigen Flusses seine Dynamik in der natürlichen chronologischen Folge (die Methode einer solchen Analyse wurde in Nadarejšvili & Orlov (1978) vorgeschlagen) und nicht in Form von "proportionalen" Teilstichproben untersuchen.

Kurz gesagt, man muß dem Objekt der eigentlichen Untersuchung mehr Beachtung schenken, so wie es in anderen Wissenschaften einschließlich klassischer Philologie und Linguistik seit langem praktiziert wird. Kein Biologe hat jemals "Gehacktes" aus einer Menge von Fröschen untersucht, um "durchschnittliche Froschcharakteristiken" zu ermitteln. Kein Philologe wird jemals ernsthaft die Phrase "Vse sčastlivye sem'i pochoži na perekladnych iz Tiflisa" [Alle glücklichen Familien ähneln den Postpferden aus Tiflis] als Modell einer "repräsentativen Stichprobe" bezeichnen.

Der Verfasser bedankt sich, ohne sich der Verantwortung für die hier vertretenen Ansichten entziehen zu wollen, bei P.M. Alekseev, M.V. Arapov, V.S. Perebejnos und Ju.A. Tuldava für die ernsthafte Diskussion einiger Behauptungen.

ANHANG

DIE STATISTISCHE ANALYSE UND PROGNOSE BEI UNZUREICHENDEN STICHPROBEN

In diesem Anhang beabsichtigen wir, ausführlicher diejenigen Fälle mathematisch zu analysieren, wo aus einer Gesamtheit mit großem "Inventar" möglicher Ausgänge zufälliger Experimente (im linguistischen Fall sind es Zähleinheiten: Buchstaben, Silben, Morpheme, Lexeme, Wortformen usw.) eine Stichprobe erhoben wird, die notorisch unzureichend ist, so daß man aus ihr die Häufigkeiten aller möglichen Ausgänge nicht zuverlässig bestimmen kann. Der Verfasser versuchte einige früher gewonnene Resultate, die in ziemlich seltenen Veröffentlichungen verstreut sind, zusammenzufassen (Kalinin 1964; 1965; Orlov 1976; 1978). Die Details der Ableitungen wurden unterlassen, aber es werden nach Möglichkeit die Grundideen ihrer Herleitung und praktischen Anwendung dargelegt.

1.

Eine einfache Rechnung bestätigt, daß man auf Abb. 1 und in Tab. 3 und 4 genau das vorfindet, was man in einer aus etwa 30 "Buchstabenverwendungen" bestehenden Stichprobe vorfinden muß. Wenn die Menge der Buchstabenwahrscheinlichkeiten in der Grundgesamtheit gleich $\{\Pi_i\}_{i=1}^{V}$ ist, dann ist die erwartete Zahl unterschiedlicher Buchstaben $v(N)$ in einer Stichprobe von N Buchstaben (aufgrund der sogenannten Poisson-Approximation) gleich

$$v(N) = \sum_{i=1}^{V} (1 - e^{-N\Pi_i}). \qquad (5)$$

Setzt man in diese Formel N = 30 und statt Π_i die von Lebedev und Garmaš angegebenen Zahlen (gezählt ohne Zwischenraum, so daß V = 31) ein, so erhält man $v(30) = 15.8$. Das heißt, die erwartete Zahl unterschiedlicher Buchstaben des russischen Alphabets, die in eine zufällige Stichprobe von 30 "Buchstabenverwendungen" geraten, kommt ganz nah an das heran, was man in der Realität beobachtet: der Mittelwert aller fünf Stichproben (vgl. Tab. 5) ergibt 15.4 unterschiedliche Buchstaben.

Man kann auch die erwartete Anzahl der Buchstaben, die in der Stichprobe nur jeweils einmal vorkommen, berechnen. Mit derselben Poisson-Approximation kann man die erwartete Anzahl $v_m(N)$ unterschiedlicher Ereignisse, von denen jedes jeweils m-mal in N Experimenten vorkam, als

$$v_m(N) = \sum_{i=1}^{V} \frac{(N\Pi_i)^m e^{-N\Pi_i}}{m!} \qquad (6)$$

berechnen.

Setzt man in diese Formel m = 1, N = 30 und dieselben Werte von Π_i ein, so findet man, daß die Anzahl einmal benutzter Buchstaben in einer Stichprobe mit Umfang N = 30 Einheiten, gleich $v_1(30) = 8.13$ ist, was wiederum dem Mittelwert der einmaligen Buchstaben in allen Stichproben, 7.6, nahekommt.

Wenn wir also die Menge der Ereigniswahrscheinlichkeiten kennen, können wir die Stichprobencharakteristika leicht berechnen. Wir können zwar nicht voraussagen, welche Buchstaben konkret in der Stichprobe vorkommen werden, aber wir können ihre Anzahl in einer Stichprobe von beliebigem Umfang voraussagen und auch, wie viele von ihnen einmal, zweimal usw. in dieser Stichprobe vorkommen werden. Um dies zu gewährleisten, muß man aber die Wahrscheinlichkeiten dieser Ereignisse kennen. Wenn

wir jedoch die Stichprobenschätzungen dieser Wahrscheinlichkeiten benutzen (d.h. die Häufigkeiten der Ereignisse in unzureichenden Stichproben), so kommen wir zu absurden Voraussagen.

Wenn wir beispielsweise für die Voraussage der Zahl unterschiedlicher Buchstaben in einer Stichprobe mit Umfang von 30 "Buchstabenverwendungen" die Buchstabenhäufigkeiten aus der Tab. 3 benutzen, so erhalten wir laut (5): $v(30) = 12.4$ (in allen realen Stichproben, einschließlich der, die als Grundlage für die Voraussage diente, sind es mehr). Eine noch schlechtere Voraussage erhält man für die Zahl der einmaligen Buchstaben laut Formel (6), nämlich $v_1(30) = 4.55$. Hier sieht man, wohin die Verschiebung der Schätzungen in mangelhaften Stichproben führt!

Also kann der Satz "Я ЕХАЛ НА ПЕРЕКЛАДНЫХ ИЗ ТИФЛИСА" weder als ein einigermaßen vollständiges Modell des Russischen dienen (er enthält nicht einmal alle im Russischen benutzten Buchstaben), noch eignen sich seine Charakteristika für eine einigermaßen annehmbare Voraussage mit Hilfe der Methoden der klassischen Wahrscheinlichkeitstheorie.

Auf der Buchstabenebene ist dies alles klar durchschaubar. Es ist nichts einfacher, als aus dieser Situation zu entrinnen: Es reicht, eine so große Stichprobe zu nehmen, daß der seltenste Buchstabe mindestens 10 mal vorkommt. Aber eine entsprechende Forderung wäre bei der Statistik der Lexeme völlig unerfüllbar. Es ist sogar absolut irreal, auch nur die einmaligen Wörter (hapax legomena) loszuwerden, deren Anzahl in einer beliebigen Stichprobe größer ist als die Anzahl der Wörter mit einer beliebigen anderen Häufigkeit. Und dies zieht unvermeidlich nach sich, daß die in lexikalischen Stichproben beobachteten empirischen Häufigkeiten in Bezug auf die Wortwahrscheinlichkeiten in der hypothetischen lexikalischen Gesamtheit systematisch verschoben sind.

Aus dem Buchstabenbeispiel können wir noch einen Schluß ziehen. Wie man aus den Daten der Tab. 4 sehen kann, sind sogar die Charakteristika solcher "Spielzeug-Stichproben" ziemlich stabil. Aber diese Stabilität ist relativ - sie hängt von dem

Stichprobenumfang ab. In unserem Fall wurden beispielsweise die Umfänge so gewählt, daß die Zahl einmaliger Wörter ungefähr die Hälfte der unterschiedlichen Buchstaben in der Stichprobe betrug. Dieses Verhältnis bleibt aber nur solange erhalten, bis wir den Stichprobenumfang ändern. Wenn wir z.B. die erwartete Zahl unterschiedlicher und einmaliger Buchstaben in einer Stichprobe mit 60 "Buchstabenverwendungen" laut (5) und (6) berechnen, so erhalten wir $v(60) = 21.3$ und $v_1(60) = 7.26$, d.h. der Anteil einmaliger Buchstaben in dem gesamten "Stichprobenalphabet" fällt ungefähr auf 1/3 zurück. Wenn wir z.B. zwei Stichproben aus den Texten von V. Koneckij vereinigen, so stellen wir fest, daß in der vereinigten Stichprobe von 60 Einheiten 20 unterschiedliche Buchstaben vorkommen und darunter 5 nur einmal (d.h. 1/4 des "Stichprobenalphabets"). Bei einer unbegrenzten Vergrößerung der Stichprobenumfänge N muß auch die Zahl einmaliger Ereignisse (wie übrigens auch der Ereignisse mit einer anderen fixierten Häufigkeit) im Grenzwert Null erreichen, vgl. Formel (6), womit wir in den Gültigkeitsbereich der klassischen Statistik eintreten. Aber solange dies nicht geschieht, ruft die Veränderung des Stichprobenumfangs eine Veränderung aller ihrer quantitativen Kenngrößen hervor.

2.

Aber gerade die systematische Natur der Verschiebung von Charakteristika einer mangelhaften Stichprobe, die sich mit der Stabilität dieser Charakteristika beim unveränderten Stichprobenumfang assoziiert, erlaubt es, Stichprobenbeobachtungen zur Prognose des Verhaltens von Stichproben, die (hypothetisch) aus einer Gesamtheit erhoben wurden einzusetzen. Man muß dabei nur die übliche Betrachtung der Häufigkeit als einer Schätzung der Wahrscheinlichkeit eines Ereignisses aufgeben, da diese Betrachtung bei mangelhaften Stichproben zu allzu großen Fehlern führt.

Die Buchstabenstichproben sollen zum letztenmal als Beispiel dienen. An diesem Beispiel zeigen wir, wie man ein gewöhnliches

theoretisches Modell konstruiert, um den Unterschied zwischen dem traditionellen und dem vorgeschlagenen Herangehen hervorzuheben.

Die traditionelle Form einer statistischen Hypothese besteht darin, daß man über die analytische Form der Verteilung in der Grundgesamtheit eine Annahme macht. Die theoretische Kurve mit den gewählten Parametern passen wir an die empirische Verteilung an, und unter der Annahme, daß die angepaßte analytische Verteilung die wahre ist, berechnen wir Konfidenzintervalle für mögliche zufällige Abweichungen der empirischen Beobachtungen von der theoretischen Verteilung. Wenn die tatsächlich beobachteten Abweichungen die Grenzen dieser Intervalle nicht überschreiten, dann kann man annehmen, daß sie der aufgestellten Hypothese nicht widersprechen. Paßt man beispielsweise die abnehmende Exponentialfunktion

$$p_i = p_1 e^{-(i-1)(p_1-p_v)} = 0{,}11e^{-(i-1)0.108}$$

(vgl. ORLOV 1976) an die empirischen Buchstabenhäufigkeiten, wie sie von Lebedev und Garmaš ermittelt wurden, an, so erhält man die Menge der "theoretischen Wahrscheinlichkeiten", die einfachheitshalber als die stetige Kurve in Abb. 1 dargestellt ist. Setzt man diese Menge in (5) und (6) mit N = 30 und m = 1 ein, so erhält man die Voraussagen: v(30) = 15.86 und $v_1(30)$ = 7.89. Dies ist sehr ähnlich den Prognosen, die man früher aus empirischen Schätzungen der Wahrscheinlichkeiten erhielt (15.8 bzw. 8.13). Wir lassen jetzt die Aufstellung der Konfidenzintervalle beiseite und bemerken lediglich, daß die "Theorie" es ermöglichte, das für die Prognose notwendige Massiv der Daten stark zu reduzieren: anstelle von 31 Zahlen, die bei Lebedev und Garmaš aufgeführt sind, reichen zwei Zahlen; die Häufigkeit des häufigsten und des seltensten Buchstabens, p_1 und p_v.

Jedoch führt die Anpassung der empirischen Häufigkeitsmenge der Buchstaben in kleinen Stichproben mit anschließender Einsetzung in (5) und (6) nicht weiter, und zwar wegen ihrer Verschiebung, denn auch die angepaßte Menge wird verschoben sein (vgl. die punktierte Linie in Abb. 1).

Um ein Modell für Einheiten des Typs von Lexemen aufzustellen, muß man mit Hilfe eines "sechsten Sinnes" die Menge der uns unbekannten Wortwahrscheinlichkeiten in der lexikalischen Grundgesamtheit erraten. So verfuhr z.B. Carroll (1968, 1969) als er die Hypothese über die log-normale Verteilung annahm. In sehr großen lexikalischen Stichproben findet man tatsächlich etwas, was an eine sogenannte "gestützte Lognormalität" erinnert. Carrolls Modell führt zu ganz guten Voraussagen, aber es ist sehr kompliziert (Carroll selbst bringt keine Arbeitsformeln und gibt zu, daß die Schätzung der Parameter aus den Daten nicht gelöst ist) und für die linguostatistische Analyse praktisch wenig geeignet (ausführlicher vgl. Orlov 1979).

Es gibt trotzdem eine Möglichkeit, die Vorstellung von der Menge der Wahrscheinlichkeiten in der statistischen Grundgesamtheit aufzugeben und ein theoretisches Modell in engerem Kontakt mit den Beobachtungen und ihren Verallgemeinerungen zu konstruieren. Diese Möglichkeit haben die Arbeiten von Kalinin (1964, 1965) eröffnet. Es hat sich gezeigt, daß es bei Kenntnis der mathematischen Erwartung des Vokabulars $v(N_o)$ und des Häufigkeitsspektrums $v_m(N_o)$ bei einem festen Stichprobenumfang N_o möglich ist, diese Größen auf einen beliebigen anderen Stichprobenumfang N umzurechnen, nämlich durch die Formeln

$$v(N) = \sum_{j \geq 1} \left[1 - \left(1 - \frac{N}{N_o}\right)^j\right] v_j(N_o) \qquad (7)$$

$$v_m(N) = \sum_{j \geq m} C_j^m \left(\frac{N}{N_o}\right)^m \left(1 - \frac{N}{N_o}\right)^{j-m} v_j(N_o) \qquad (8)$$

wo $C_j^m = \frac{j!}{m!(j-m)!}$ die Zahl aller Kombinationen von jeweils m Elementen aus j Elementen ist.

Vergleicht man (7) und (8) mit (5) und (6), so sieht man, daß man die Größen v(N) und $v_m(N)$ berechnen kann, wenn man ent-

weder die ganze Menge der Wahrscheinlichkeiten $\{\Pi_i\}_{i=1}^{v}$ oder die Menge der mathematischen Erwartungen $\{v_j(N_o)\}$ bei einem Stichprobenumfang N_o kennt. Auf den ersten Blick scheint die neue Möglichkeit keine Erleichterung zu bringen, d.h. um die Menge $\{v_j(N_o)\}$ zu erhalten, muß man die Menge der Wahrscheinlichkeiten $\{\Pi_i\}$ kennen, und wenn man sie nicht kennt, so weiß man nicht, woher man $\{v_j(N_o)\}$ nehmen soll.

Aber ähnlich, wie man in (5) und (6) gewöhnlich eine Hypothese über die Menge der Wahrscheinlichkeiten $\{\Pi_i\}$ einsetzt, so kann man auch in (7) und (8) eine Hypothese über die Menge der mathematischen Erwartungen $\{v_j(N_o)\}$ beim Umfang N_o einsetzen.

Eine solche Hypothese kann man aufgrund der Verallgemeinerungen über vorhandene Stichprobenhäufigkeiten konstruieren. Es ist seit langem bekannt, daß die "Rang-Häufigkeit"-Abhängigkeiten mit hyperbolischen Funktionen angenähert werden können, die auf doppellogarithmischer Graphik eine Gerade bilden. Dies wird freilich nicht immer beobachtet, aber wir wissen schon, daß bei kleinen Stichproben die Umfangsveränderung alle ihre Kenngrößen systematisch verändert, so daß Ausnahmen in dieser Situation völlig unvermeidlich sind. Es ist daher logisch anzunehmen, daß die hyperbolische Abhängigkeit nur eine von vielen Formen ist und daß man sie bei einem gegebenen Text nur in Stichproben mit einem bestimmten fixierten Umfang, den wir weiterhin mit Z bezeichnen werden, beobachten kann. Wir werden weiter voraussetzen, daß bei Stichproben mit diesem Umfang die beobachtete Menge von Wahrscheinlichkeiten mithilfe eines Ausdrucks des Typs (3) angenähert werden kann, wobei die Koeffizienten K und B mithilfe der Normierung bestimmt werden können (ausführlicher vgl. Orlov 1976; hier wird auch der Fall analysiert, wenn γ in der Mandelbrotschen Formel von 1 unterschiedlich ist).

Die Formel (3) beschreibt aber nur den Bereich der häufigen und mittelhäufigen Wörter gut. Sie beschreibt den Bereich seltener Wörter unadäquat, da sie voraussetzt, daß alle Häufigkeiten unterschiedlich sind, während in der Tat im Bereich seltener Wörter umfangreiche Gruppen von Wörtern mit derselben Häufigkeit vorhanden sind. Diese geradlinigen "Treppen" kann man mit der Hyperbel (3) nur dann in Übereinstimmung bringen, wenn ihre Längen gleich

$$v_m(Z) = \frac{v(Z)}{m(m+1)} \qquad (9)$$

sind, wobei $v(Z) \approx Z/\ln(Zp_1)$, der Wortschatz in der Stichprobe vom Umfang Z ist, die nach Definition einer hyperbolischen Abhängigkeit folgt.

Ohne uns den Kopf über die unbeobachtete Menge von Wahrscheinlichkeiten in der Grundgesamtheit zu zerbrechen, nehmen wir an, daß es einen Stichprobenumfang aus dieser Gesamtheit gibt, bei dem die mathematische Erwartung der m-maligen Wörter durch (9) bestimmt ist, und der Bereich der häufigen und mittelhäufigen Wörter durch (3) approximiert wird.

Was geschieht, wenn wir aus derselben Gesamtheit eine Stichprobe mit einem von Z sich unterscheidenden Umfang N erheben? Die (relativen) Häufigkeiten häufiger und mittelhäufiger Wörter (oder anderer Einheiten, für die sich diese Theorie als geeignet erweisen kann) ändern sich nicht (bei einer Genauigkeit von der Größenordnung der üblichen statistischen Streuung) und werden nach wie vor durch (3) beschrieben. Im Bereich seltener Wörter entstehen unvermeidliche systematische Verschiebungen, die man mit den Formeln von Kalinin berechnen kann, wenn man anstelle der willkürlichen mathematischen Erwartungen $v_j(N_o)$ die rechte Seite von (9) einsetzt. Die Formeln (2) und (4a, b) sind das Resultat dieser Einsetzung.[7] Die allgemeine logische Struktur der obigen Konstruktionen führt zu folgenden Behauptungen:

(1) Es gibt eine lexikalische Grundgesamtheit, in der bei zufälliger Erhebung (dies war die einzige Voraussetzung bei der Ableitung der Formeln (5) und (6) von Kalinin) jedes Wort eine feste Wahrscheinlichkeit hat.

(2) Diese Gesamtheit ist so gestaltet, daß bei der Erhebung einer zufälligen Stichprobe mit Umfang Z aus ihr die erwartete Zahl $v_m(Z)$ der Wörter mit Häufigkeit m durch (9), die nach abnehmender Größe geordnete Menge der Häufigkeiten durch (3) ge-

geben ist, und die Zahl unterschiedlicher Wörter in der Stichprobe gleich v(Z) ist.

Aus diesen zwei Behauptungen folgt die Behauptung:

(3) Erhebt man aus dieser Gesamtheit eine Stichprobe mit einem (beliebigen) Umfang N, dann ist die erwartete Anzahl unterschiedlicher Wörter v(N,Z) in dieser Stichprobe durch (2), die nach abnehmender Größe geordnete Menge der Häufigkeiten häufiger und mittelhäufiger Wörter durch (3) und die erwartete Anzahl unterschiedlicher Wörter, die jeweils m mal vorkommen, durch (4a, b) gegeben.

Auf den ersten Blick erscheint die Überprüfbarkeit der Adäquatheit dieser Konstruktionen äußerst problematisch. In der Tat, wenn auch eine solche Gesamtheit existieren mag, derart daß beim Umfang Z die Beziehungen (3) und (9) gelten, wie kann man diesen Umfang erraten? Es zeigt sich, daß es nicht schwer ist, Z aus den Stichprobendaten abzuschätzen, wenn man annimmt, daß Z der einzige unabhängige Parameter in allen Konstruktionen ist, der alle beobachteten Größen miteinander verbindet. Es reicht, eine einzige Gleichung aufzustellen, um sich aus der Lösung bezüglich Z eine Vorstellung über seinen numerischen Wert machen zu können.

Man geht am besten von der Gleichung (2) aus, die eben die Funktion des Vokabularwachstums beim Anwachsen des Stichprobenumfangs N ist. Da diese Kurve nur von einem Parameter abhängt, reicht es, sie mit einem einzigen beobachteten Punkt, mit dem beobachteten Vokabularumfang der Stichprobe $\hat{v}(N)$ gleichzusetzen:

$$v(N,Z) = \hat{v}(N). \qquad (10)$$

Diese Funktion ist leider transzendental und läßt sich mit elementaren Funktionen nicht ausdrücken. Für die numerische Lösung vgl. Orlov (1978, 1980).

Nachdem man aus (10) den numerischen Wert von Z erhalten hat, kann man ihn zur Überprüfung der Hypothese verwenden. Wenn man außer der einzigen Stichprobe, aus deren Vokabular $\hat{v}(N)$ man Z berechnet, nichts anderes zur Verfügung hat, so berechnet man die theoretische Häufigkeitsstruktur nach (3) und (9) und vergleicht sie mit der der Stichprobe; denn diese Zahlen spielten keine Rolle bei der Bestimmung von Z. Hat man mehrere Stichproben mit unterschiedlichen Umfängen aus (hypothetisch) einer Gesamtheit, so bestimmt man Z aus einer von ihnen und berechnet es nach denselben Formeln (2), (3) und (4) für die anderen Umfänge N.

Die Resultate einer ähnlichen Analyse an sehr umfangreichem Material wurden ausführlich in Orlov (1978) dargelegt. Es hat sich gezeigt, daß sich jeder einzelne literarische Text so "verhielt", als ob er eine Stichprobe aus der Gesamtheit mit den postulierten Eigenschaften wäre. Beispielsweise war die Prognose des Vokabulars laut (2) von einem Abschnitt auf das Ganze (oder vom Ganzen auf einen Abschnitt, oder vom Abschnitt eines Umfangs auf den Abschnitt eines anderen Umfangs) nur selten mit einer Ungenauigkeit von mehr als 5 - 8% behaftet. Etwas schlechter verhielten sich zusammengesetzte Stichproben, die aus vielen Texten oder Abschnitten von ihnen bestanden. Die Fehler bei der Prognose des Wokabulars von einer Teilstichprobe auf die andere erwiesen sich als etwas größer, obwohl sie in "vernünftigen" Grenzen blieben.

Zusätzlich wurde die bereits analysierte Erscheinung beobachtet: der volle Umfang eines literarischen Textes war von derselben Ordnung wie der "Zipfsche Umfang" Z; zwischen diesen Größen zeigte sich in dem gesamten Korpus der untersuchten Texte (ungefähr 100) eine signifikante Korrelation, und ihre Häufigkeitsgraphiken erinnerten an die drei mittleren Graphiken in Abb. 2. Die Häufigkeitsgraphiken von Textabschnitten ähnelten der untersten Graphik in Abb. 2. Die Zusammensetzungen vieler Texte in eine Stichprobe wiesen in der Regel einen Umfang auf, der den "Zipfschen Umfang" überstieg, und ihre Häufigkeitsgraphiken ähnelten der obersten Graphik in Abb. 2. Eine signifikante Korrelation zwischen dem Stichprobenumfang und dem Zipfschen Umfang fehlte hier. In allen 174 Texten und Stichproben, die in der obigen Arbeit analysiert wurden, gab es keine lexikalischen Korpora, deren Häufigkeitsstruktur sich von der theoretischen Prognose deutlich unterschieden hätte.

Woran liegt diese merkwürdige Stabilität der analytischen Form von "Rang-Häufigkeitskurven"? Sie sind auch in den Fällen stabil, wenn man sehr heterogene Texte vermischt, wie z.B. im Häufigkeitswörterbuch des Tschechischen. Es ist überhaupt nicht evident, warum bei mehreren Texten, die einzeln die mit (3) und (4) beschriebene Häufigkeitsstruktur aufweisen, auch eine aus ihnen zusammengesetzte Stichprobe dieselbe Struktur hat, wenn man diese Texte als nicht aus einer und derselben Gesamtheit stammend betrachten kann.

Eine mathematische Analyse der "statistischen Heterogenität" ist im allgemeinen einfach deswegen nicht möglich, weil dieser Begriff selbst eine völlige Willkür voraussetzt. Es ist aber sehr leicht, den Grenzfall der "statistischen Heterogenität" zu untersuchen.

Stellen wir uns vor, daß wir in einer Stichprobe zwei Texte vereinigen, die jeweils durch (3) und (9) mit denselben Z und p_1 beschrieben werden, d.h., beide Texte sind nach ihrer "quantitativen Konstruktion" vollkommen gleich. Zwecks Bestimmtheit nehmen wir an, daß ihre Länge jeweils N = Z ist. Wenn diese beiden Texte Stichproben aus derselben Gesamtheit darstellen, dann wird die Häufigkeitsstruktur ihrer Vereinigung durch (3) und (4a, b) erfaßt, wobei Z und p_1 gleich bleiben und N = 2Z wird.

Seien es aber zwei Texte aus unterschiedlichen Sprachen, oder sei kein Wort des einen Textes in dem anderen enthalten. Eine größere "statistische Heterogenität" kann man sich nicht vorstellen. Man kann das Resultat ihrer Vereinigung zu einer Stichprobe sehr einfach bestimmen: der Stichprobenumfang und das Vokabular werden verdoppelt, die Zahl m-maliger Wörter für beliebiges m verdoppelt sich auch (d.h. statt Z, v(Z), $v_m(Z)$ bekommt man 2Z, 2v(Z) bzw. $2v_m(Z)$); die Häufigkeiten häufiger Wörter verdoppeln sich auch, da jede durch (3) gegebene Häufigkeit einmal wiederholt wird. Dabei verringert sich die relative Häufigkeit des häufigsten Wortes p_1 um die Hälfte, da wir jetzt einen doppelten Stichprobenumfang haben und die absoluten Worthäufigkeiten unverändert bleiben. Bleibt die ursprüngliche Häufigkeitsstruktur in so addierten Texten erhalten?

Sie bleibt offensichtlich erhalten! Wenn wir in (2) statt Z jetzt 2Z und ein um die Hälfte verringertes p_1 einsetzen, so erhalten wir v(2Z) = 2v(Z); analog erhalten wir aus (9) $v_m(2Z) = 2v_m(Z)$. Das heißt, wir bekommen genau die Resultate, die man beim einfachen Zusammensetzen der Texte zu einer Stichprobe erhält. Auch die Größe K [vgl.(3)] bleibt erhalten, da

$$K = \frac{1}{\ln(2Z\frac{1}{2}p_1)} = \frac{1}{\ln(Zp_1)}.$$

Es ändern sich lediglich B in (3) wegen der Veränderung von p_1; das Resultat ist in Abb. 3 dargestellt, wobei zwecks Anschaulichkeit auf die Ordinate die absoluten Häufigkeiten aufgetragen wurden und die relativen Häufigkeiten p_i mit dem Stichprobenumfang multipliziert wurden.

In diesem Beispiel wurde die Länge der Ausgangstexte N lediglich zur leichteren Beurteilung dem Z gleichgestellt. Jedoch im allgemeinen Fall, wo N ≠ Z, geschieht dasselbe: laut (2) und (4a,b) ist v(2N, 2Z) = 2v(N,Z) und $v_m(2N, 2Z) = 2v_m(N,Z)$ (der kundige Leser kann die Gleichungen selber überprüfen: er darf nur nicht vergessen, daß p_1 halbiert wird).

Wenn man also "absolut heterogene" Texte zu einer Stichprobe vereinigt, so vergrößert sich Z, aber die analytische Form, die "Rang-Häufigkeitskurve" bleibt erhalten. Die Häufigkeitsstrukturen des Typs (3) und (4a,b) sind genauso "reproduktiv" wie die Normalverteilung (zumindest unter bestimmten Bedingungen). Man kann annehmen, daß die reale statistische Heterogenität gewöhnlicher Texte die Reproduktivität nicht verhindert, und daß die Veränderung des "Zipfschen Umfangs" Z bei der Zusammensetzung von Texten als Kenngröße dieser Heterogenität dient.

Gerade durch diese Stabilität der Häufigkeitsstrukturen gegen statistische Heterogenität erscheint die Existenz derjenigen Gesamtheit, deren Eigenschaften wir bei der Ableitung von (4a,b) postulierten, als illusorisch - es kann sie gar nicht geben. Es stellt sich vielleicht heraus, daß, ähnlich wie die

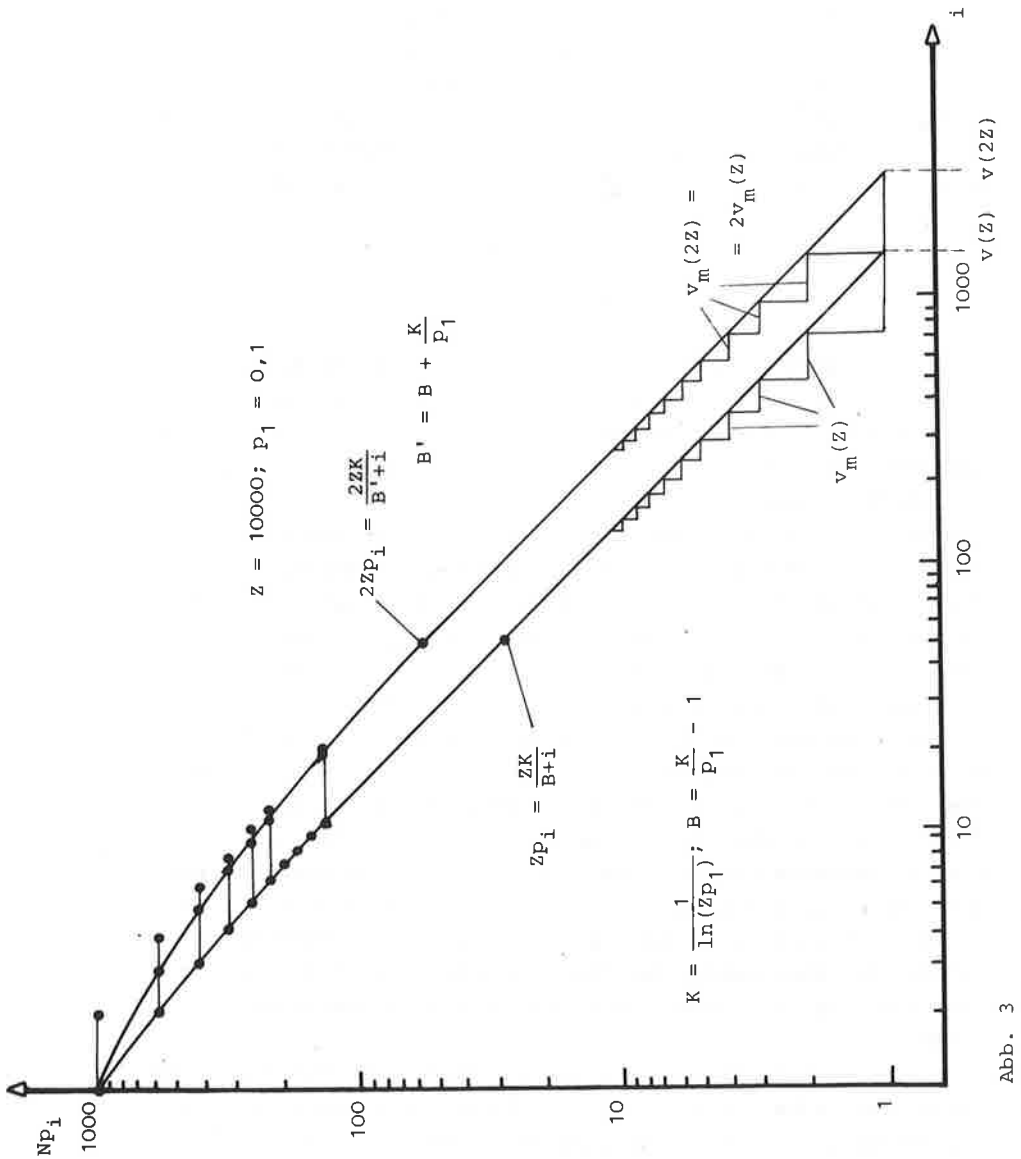


Abb. 3

Gaußsche Verteilung als Grenzverteilung bei der Summierung beliebig verteilter Zufallsvariablen entsteht, so auch die Strukturen (3) und (4a,b) bei der Vereinigung verschiedener lexikalischer "Flüsse" entstehen. Sogar einen individuellen literarischen Text kann man als Vereinigung unterschiedlicher "Flüsse" betrachten, z.B. der Autorensprache, der direkten Rede der dargestellten Personen usw.[8] Und in allen diesen Fällen kann man einen Z-Wert wählen, der die Struktur jedes einzelnen Flusses charakterisiert.

Bisher haben wir nur die Statik der Vereinigung zweier "absolut heterogener" Texte zu einer Stichprobe betrachtet. Es ist aber nicht weniger lehrreich, auch die Dynamik dieses Prozesses bei unterschiedlichen Vereinigungsweisen zu analysieren.

Wir werden zunächst so verfahren, wie man es gewöhnlich tut: wir "homogenisieren" die Stichprobe künstlich. Wir schreiben den Text auf Zettel, vermischen sie gründlich, und dann ziehen wir einen Zettel nach dem anderen (ohne Zurücklegung) und halten ihre Folge fest. So bekommen wir einen "Quasitext", dessen Vokabular bei einem Umfang N gleich v(N, 2Z) (vgl. Formel (2) und oberste Graphik in Abb. 4) sein wird. Es ist offensichtlich, daß bei dieser Analyse jegliche Information über die Eigenschaften der ursprünglichen Texte völlig verschwindet.

Beobachtet man den Vokabularzuwachs in einer linearen Folge von zwei Texten, dann ist bei Umfängen $N \leq Z$ der erwartete Vokabularumfang gleich v(N,Z), d.h. dies ist das ganz natürliche Anwachsen des Vokabulars in jedem einzelnen Text (die unterste Graphik in Abb. 4 links von der Abszisse Z). Jedoch, wenn nach dem ersten Text der zweite anfängt, so muß es zu einem schnellen Zuwachs des Vokabulars kommen (ein Text in einer anderen Sprache fängt an, und am Anfang sind alle Wörter neu). Bei $Z < N \leq 2Z$ wird das erwartete Vokabular durch v(Z) + v(N - Z, Z) bestimmt (d.h. (das Vokabular bis N = Z) + Vokabularzuwachs in neuem Text; der Teil der unteren Graphik in Abb. 4 zwischen den Abszissenpunkten Z und 2Z). Dadurch ergibt sich auf der Vokabularzuwachskurve ein scharfer Bruch, der durch die Ablösung eines Textes durch den anderen entsteht. Gerade dieser Bruch (wenn

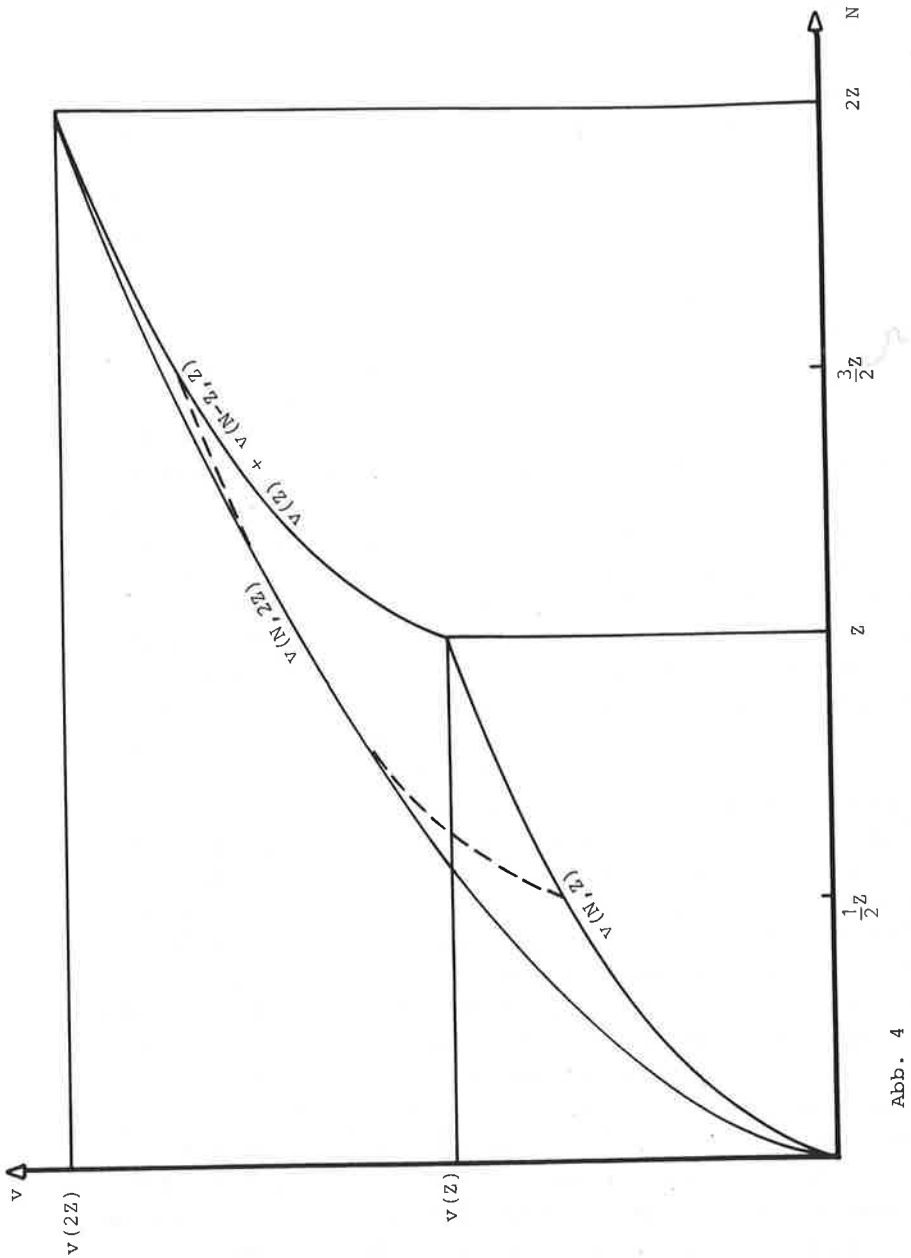


Abb. 4

wir den Charakter der vereinigten Texte vorher nicht kennen) gibt uns Auskunft über die plötzliche Veränderung der statistischen Situation. Das heißt, wenn wir einen "Fluß" des Lexikons verfolgen, dessen statistischer Homogenitätsgrad uns vorher völlig unbekannt ist, dann müssen wir ihn in seiner natürlichen Anordnung untersuchen: nur in dem Fall bewahren wir die Information über die wesentlichen, nicht von uns selbst erzeugten Eigenschaften des Prozesses.

Zum Schluß analysieren wir den der linguostatistischen Praxis nahen Fall der groben "Homogenisierung", wobei die Texte nicht wie oben beschrieben, gemischt, sondern stückweise vereinigt werden. Zerlegen wir die beiden Texte in jeweils zwei Hälften und bilden die Folge: 1. Hälfte des 1. Textes + 1. Hälfte des 2. Textes + 2. Hälfte des 1. Textes + 2. Hälfte des 2. Textes. Das Vokabularwachstum in einzelnen Teilen dieses "Textes" wird durch folgende Ausdrücke bestimmt:

bei $N \leq \frac{1}{2}Z$	$v(N,Z)$
bei $\frac{1}{2}Z < N \leq \frac{3}{2}Z$	$v(\frac{1}{2}Z, Z) + v(N - \frac{1}{2}Z, Z)$
bei $\frac{3}{2}Z < N \leq 2Z$	$v(Z) - v(N - Z, Z)$.

Die Stellen, wo diese Ausdrücke mit den vorher aufgetragenen Kurven in Abb. 4 nicht übereinstimmen, sind mit unterbrochenen Linien angegeben. Wie aus der Abbildung ersichtlich ist, nähert schon diese grobe Textmischung die Wachstumskurve des Vokabulars an die Kurve der "idealen Mischung" $v(N, 2Z)$ an: merkliche Abweichungen kommen nur am Anfang vor, im ersten Viertel des addierten "Textes". Je feiner wir die Texte zerlegen und abwechselnd ihre Teile hintereinanderfügen, desto näher kommen wir der Kurve $v(N, 2Z)$. Dies stellt aber jene künstlich erzeugte Realität dar, die die Mediziner und Biologen als Artefakt bezeichnen.

Auf den ersten Blick könnte es so scheinen, daß man hier einen zu extremen Fall analysiert, der nur theoretisch interessant ist: wer würde denn in einer Stichprobe Texte aus verschiedenen Sprachen vereinigen! Es gibt aber Daten, die zeigen, daß etwas Ähnliches sogar in zusammenhängenden literarischen Texten in einer Sprache vorkommt. In Nadarejšvili & Orlov (1971, 1978) wurde erwähnt, daß das Vokabularwachstum in Anfangsteilen von "Die Kosaken" und "Die Kreuzersonate" von L.N. Tolstoj bedeutend tiefer liegt als die theoretische Kurve v(N,Z), deren Parameter Z aus den Daten der ganzen Stichprobe ermittelt wurde. Die Übereinstimmung mit der theoretischen Situation, die in Abb. 4 für die "stückweise Textmischung" dargestellt wird, sagt an sich wenig, aber eine detailliertere von E.B. Oborneva[9] durchgeführte Analyse zeigte, daß der Tolstojsche Text erstaunlich ähnlich wie unsere "stückweise Textmischung" "konstruiert" ist. Oborneva analysierte "Die Kosaken" nach einzelnen Kapiteln, wobei sie Z für jedes Kapitel separat berechnete, anschließend vereinigte sie die Kapitel in ihrer natürlichen Reihenfolge und berechnete Z nach jedem hinzugefügten Kapitel. Es zeigte sich, daß die Z-Werte von einzelnen Kapiteln zwischen 15000 - 25000 liegen, d.h. der relative Vokabularreichtum jedes einzelnen Kapitels nicht hoch ist (mit Ausnahme von ein - zwei Kapiteln). Die anschließende Vereinigung der Kapitel erhöht Z ziemlich schnell bis auf (ungefähr) 50000[10] (wonach es sich stabilisiert), d.h. der relative Vokabularreichtum wächst ungefähr so an, als ob die Kapitel abwechselnd in zwei unterschiedlichen Sprachen geschrieben wären[11]. Offensichtlich erhöht sich Z nach der Vergrößerung des Umfangs des literarischen Werkes gerade wegen der Verflechtung unterschiedlicher Sujet- und thematischer Linien, die jeweils ihr eigenes charakteristisches Lexikon besitzen. Aber dieser Mechanismus der Erhöhung des relativen Vokabularreichtums mit dem Anwachsen des vollen Textumfangs ist nicht "automatisch": er braucht sozusagen die Kontrolle des Autors, da es nötig ist, die Menge dieser Linien und den Grad ihrer "gegenseitigen Heterogenität" mit dem vollen Textumfang in Übereinstimmung zu bringen. Es ist daher wichtig, nicht nur einen Text "als ganzen" zu untersuchen, sondern auch seine strukturellen Komponenten und ihre Interrelationen im Detail zu analysieren.

Nach der Fertigstellung dieser Arbeit lernte der Verfasser den Aufsatz von V.V. Nalimov (1979) kennen, in dem ähnliche Probleme von einem allgemeinbiologischen Standpunkt behandelt werden (wer würde bestreiten, daß Sprache und Rede Produkte lebendiger Natur sind?). Der Autor kann der Versuchung, Nalimovs grundlegende Folgerungen zu zitieren, nicht widerstehen, er erlaubt sich nur, eine Stelle durch Sperrdruck hervorzuheben:

"Die unangenehme Überraschung besteht darin, daß uns, nachdem wir erkannt haben, daß die (biologische) Variabilität zufälliger Natur ist, zu unserem größten Erstaunen die Möglichkeit, den gewöhnlichen Wahrscheinlichkeitsansatz zu benutzen, genommen wurde. Denn eine statistische Beschreibung ist dann möglich, wenn es gelingt, aus den Ergebnissen der Beobachtung einer kleinen Stichprobe eine Vorstellung über das Verhalten jeder denkbaren Folge von Erscheinungen zu gewinnen. Jedoch in Fällen mit biologischer Variation ermöglichen Beobachtungen von kurzen Folgen von Erscheinungen keine Schlüsse über das weitere Verhalten des Systems. *Im Unterschied zur Physik haben hier durchschnittliche Charakteristika keinen Sinn! Wichtig sind die einzelnen Erscheinungen in ihrer individuellen Erscheinungsform* unabhängig von ihrer Vorkommenswahrscheinlichkeit."

Anmerkungen

1 Eine ausführliche mathematische Analyse und die Probleme der Prognose unter den Bedingungen unzureichender Stichproben werden im Anhang besprochen.

2 Es ist interessant, daß Morgenthaler selbst offensichtlich nur an die philologische Sinnfälligkeit seiner Zahlen gedacht hat; zum Glück hat er nicht angefangen, das Anwachsen des Wortschatzes in der "mittleren neutestamentlichen Fachsprache" durch Vermischung von Textabschnitten in proportionale Teilstichproben zu untersuchen.

3 Der Zuwachs des relativen Vokabularreichtums beim Zusammenstellen unterschiedlicher Texte (oder zusammenhängender Abschnitte aus ihnen) kann als Indikator der lexikalischen Ähnlichkeit der Texte dienen. Diese Annahme wird z.B. in I.S. Nadarejšvili (1978) verwendet. Die Veränderung der Größe Z bei Vereinigung von Stichproben und auch innerhalb eines zusammenhängenden Textes (vgl. Tab. 3 in Orlov 1978) kann man als Maß der Verschiebung zweiter Art betrachten.

4 Nach Ansicht einiger Autoren sind es vielleicht spätere Kompilationen früherer Texte (Vrede 1908; Robertson 1959; Lencman 1960). Diese Version könnte nicht nur das niedrige Z in diesen Texten erklären, sondern auch die Tatsache, daß diese Texte ihren "Zipfschen Umfang" übersteigen, wenn man annimmt, daß jeder der Quellentexte einen kleineren Umfang hatte und folglich mit seinem "Zipfschen Umfang" besser übereinstimmte.

5 Je größer dieses Verhältnis, desto mehr wächst der relative Wortschatz beim Übergang von Lexemen zu Wortformen und folglich ist die Synthetizität der Sprache desto größer. Der Vorzug dieser Kenngröße liegt darin, daß sie nicht vom Textumfang abhängt, während sich das Verhältnis vWortformen/vLexeme mit der Veränderung des Umfangs des Textes oder der Stichproben, aus denen v berechnet wurde, selbst ändert, und dadurch einen Vergleich unterschiedlich langer Texte ausschließt.

6 In Orlov (1978) wurde gezeigt, daß der relative Vokabularreichtum in gereimter Poesie höher ist als in der zur gleichen Zeit erschienenen Prosa; die möglichen Ursachen wurden in Orlov (1974) analysiert. Die Differenzierung nach Stilen und Genres ist unbedingt nötig.

7 Ein Mathematiker würde hier bemerken, daß eine mathematische Erwartung der Anzahl m-maliger Wörter laut (9) in keiner diskreten Stichprobe möglich ist. Bei m > Z bleiben nämlich die Wörter $v_m(Z)$ von Null verschieden, obwohl die absolute Häufigkeit eines Ereignisses den Stichprobenumfang Z nicht überschreiten kann. Dies ist aber der Preis der stetigen Approximation, mit der man die Aufgabe lösen muß. Es gibt noch mehrere spezifisch mathematische Probleme, die mit der beschriebenen Form statistischer Hypothesen zusammenhängen (speziell: nicht jede funktionale Abhängigkeit zwischen m und v_m eignet sich als die hypothetische erwartete Häufigkeitsstruktur, die man in Kalinins Formeln einsetzen könnte).
Der Verfasser hofft, diese Probleme in einem anderen Aufsatz erörtern zu können, aber diese Fragen haben auf die hier dargelegten Resultate keinen Einfluß.

8 Wie Darčuk (1975) anhand mehrerer Texte ukrainischer Autoren festgestellt hat, bewegt sich das Z der Autorensprache zwischen 100000 - 200000 und das der Personenrede zwischen

15000 - 45000. Bestimmt anhand von ununterbrochenen Passagen, liegt Z zwischen 60000 - 120000; die Längen der von Darcuk untersuchten Werke sind bis auf zwei kürzere Texte von derselben Größenordnung.

9 Die Daten wurden nicht publiziert. Der Verfasser dankt E.B. Oborneva für die persönliche Mitteilung.

10 Nach unserer und I.S. Nadarejšvilis Schätzung beträgt die volle Länge von "Die Kosaken" 48000, nach der Schätzung von Oborneva 43000 Wortverwendungen.

11 Die Erscheinung stimmt gut mit der "Häufung" seltener Wörter in literarischen Texten überein (Boroda, Nadarejšvili, Orlov, Čitašvili 1977; G.Š. Nadarejšvili, I.Š. Nadarejšvili, Orlov 1977). Eben diese Erscheinung ist "schuld" an der Veränderung von Z innerhalb eines zusammenhängenden Textes. Die Tatsache, daß Z innerhalb eines zusammenhängenden Textes oft "kriecht", wird manchmal als "Unadäquatheit" des dargelegten Modells betrachtet (es versteht sich, daß es für reale Texte in dem Sinne unadäquat ist, daß es eine ideale statistische Homogenität bei der Ableitung aller Formeln voraussetzt!). Wenn aber "das Kriechen" von Z innerhalb eines zusammenhängenden Textes reale, vom Forscher unabhängige Eigenschaften dieses Textes wiederspiegelt, dann drückt die Veränderung von Z bei der Zusammensetzung mehrerer Texte in eine Stichprobe nur die Eigenschaften der "künstlichen Realität" aus, die der Forscher selbst konstruiert hat; dieses "Kriechen" kann bestenfalls für die Kontrolle des tatsächlichen Heterogenitätsgrades der vermischten Texte verwendet werden.

Die Methode der vollständigen Textfixierung durch eine linguistisch-statistische Analyse

I.Š. Nadarejšvili, Ju.K. Orlov

1.

Quantitative Methoden bei Untersuchungen von Sprache und Rede sind in der heutigen Zeit zu einer gewohnten Erscheinung geworden, eine große Menge von Daten wurde bereits veröffentlicht, eine nicht geringe Zahl von Dissertationen (darunter auch solche, die sich der Hilfe des Computers bedienten), aber trotzdem treffen die Versuche, aus dem vorhandenen Material irgendwelche Verallgemeinerungen zu ziehen oder für seine Bearbeitung mathematische Methoden anzuwenden, noch immer auf nicht geringe Schwierigkeiten. Dafür gibt es zwei Ursachen: Die Uneinheitlichkeit bei den Berechnungen und die wesentliche Unvollständigkeit der gewöhnlich veröffentlichten Angaben. Sehr oft fehlen sogar Angaben über die Häufigkeitsstruktur [für die Liste aller beobachteten Worthäufigkeiten werden ebenso die Termini "lexikalisches Spektrum" (Kalinin 1964), "linguistisches Spektrum" (Alekseev 1975) u.a. verwendet]. Über dieses dringliche Problem schreibt Alekseev (1975) folgendes: "Leider werden solche Tabellen nicht vollständig in jedem Häufigkeitswörterbuch angeführt; darüberhinaus kann man sie fast nirgendwo finden, von seltenen Ausnahmen abgesehen ... . Anscheinend reichte die Geduld des Verfassers, der eine langwierige und schwierige Arbeit

über dem Wörterbuch geleistet hat, nicht aus, noch einige Stunden an die Herstellung der Tabellen zu verwenden. Schließlich kann der Leser sich eine solche Tabelle auch selbst erstellen, wenn das Wörterbuch wenigstens die grundlegenden Angaben enthält: die Häufigkeiten und die Anzahl der Wörter mit gleichen Häufigkeiten. Jedoch ... nimmt man in ein veröffentlichtes Wörterbuch gewöhnlich nur eine begrenzte Anzahl von Einheiten auf. Wenn über die Einheiten, die nicht in den Publikationen berücksichtigt wurden, keine quantitativen Angaben gemacht werden, verliert ein solches Wörterbuch in vielen Hinsichten seinen linguistischen Wert. Um ein solches Wörterbuch mit Tabellen zu versehen, die die Anzahl der verschiedenen Einheiten jeder Häufigkeit angeben, brauchte man eine geringfügige Zeit verglichen mit der, die bei der Erstellung des Wörterbuches benötigt wurde."

Eine der Ursachen dieser Erscheinung hängt damit zusammen, daß in eine Publikation gewöhnlich nur solche Angaben hereingebracht werden, die auf irgendeine Weise in der Konzeption des Autors eine Rolle spielen (um derentwillen die Berechnungen angestellt wurden). Die Angaben, die keine Beziehung zu dieser Konzeption haben (die man jedoch im Verlauf der Berechnungen unweigerlich erhält) werden schlicht vernachlässigt, und das macht die Resultate der Berechnungen für andere Forscher unzugänglich. Andererseits sind linguistische Berechnungen außerordentlich arbeitsaufwendig (sogar wenn sie mit Hilfe von EDV ausgeführt werden) und deshalb stellen sie wertvolle Ergebnisse dar, völlig unabhängig von jener Konzeption, die der Verfasser der Berechnungen mit ihrer Hilfe zu erhärten versucht. Die statistische Linguistik kann sich nicht den Luxus der Nachahmung erlauben, zum Beispiel der Nachahmung der Physik, in der nur Messungen durchgeführt werden, die für die Überprüfung dieser oder jener Hypothese unumgänglich sind. Die statistische Linguistik (und die gesamte Computerlinguistik) hat soeben erst ihr Interesse für die Sammlung von Fakten entdeckt, und deshalb ist es sehr wichtig, daß jedes untersuchte Faktum nicht nur genau und

sorgfältig festgehalten wird, sondern es muß auch ein Vergleich mit anderen Fakten ermöglicht werden, die von anderen Forschern an anderen Orten untersucht worden sind.

Wir lassen das rein lintuistische Problem der Vereinheitlichung der Berechnungen (zum Beispiel das Problem der Wortsegmentierung oder der Unterscheidung der Homonyme) beiseite und widmen uns in der vorliegenden Arbeit dem Problem der Vollständigkeit der quantitativen Beschreibung eines linguistischen Objekts.

Was kann man unter der vollständigen Beschreibung eines bestimmten Objektes verstehen? Offensichtlich eine solche Beschreibung, anhand derer man, wenn auch nur im Prinzip, das Objekt selbst wieder erstellen kann. Eine gewöhnliche linguistisch-statistische Analyse führt aber zu einer irreversiblen Destrukturierung des Objektes: es ist offensichtlich, daß man anhand des Häufigkeitswörterbuches eines bestimmten Textes den Text selbst unmöglich wieder herstellen kann. Im Prinzip kann ein Text wieder erstellt werden durch einen Index (oder eine Konkordanz), in dem die Adresse (die Seite und Zeile in einer festgelegten Ausgabe des Textes) der Verwendung jedes Wortes angegeben ist. Demnach stellt eben der Index eine vollständige Beschreibung eines Textes dar. Vom Standpunkt der quantitativen Analyse erscheint der Index jedoch als ein außerordentlich unfertiges Produkt und es sind sehr große Anstrengungen nötig, um aus ihm die notwendigen Zahlen zu entnehmen.

Die traditionelle Form eines Häufigkeitswörterbuches besitzt noch einen wesentlichen Mangel. Sie hält nur eine bestimmte Statik eines Textes fest, wobei sie die Frage der Dynamik außer acht läßt. Wenn man zum Beispiel das Anwachsen des Wortschatzes mit dem Wachsen des Textumfanges untersuchen will, so ist man gezwungen, eine zusätzliche, langwierige und mit Fehlern behaftete Prozedur der Einteilung eines Textes in Teilstichproben vorzunehmen, die Teilstichproben auszuwählen und zum Schluß die einzelnen Wortschätze der Teilstich-

proben zu vereinigen. Die Umfänge der Teilstichproben bestimmen hierbei den Grad der Genauigkeit, mit dem wir den Prozeß des Anwachsens des Wortschatzes untersuchen können. Ist der Text zum Beispiel in Teilstichproben zu je 1000 Wortverwendungen aufgeteilt, werden wir nach der Vereinigung dieser Teilstichproben den Wortschatz des Textes nur nach jeweils 1000 Wortverwendungen kennen. Da es notwendig ist, das Anwachsen des Wortschatzes detaillierter zu untersuchen (zum Beispiel nach jeweils 100 Wortverwendungen) erweisen sich die alten Berechnungen als unbrauchbar und die Zählung muß von neuem begonnen werden. Natürlich geht praktisch niemand so vor - man muß sich mit der Aufteilung zufrieden geben, die am Anfang der Arbeit auf Grund der apriorischen Überlegungen erhalten worden war, obwohl sie möglicherweise bei weitem nicht die optimale Aufteilung für die Lösung der gestellten Aufgabe ist. Noch weniger wird diese Aufteilung anderen Forschern von Nutzen sein, die die veröffentlichten Daten verwenden wollen.

In der vorliegenden Arbeit wird die Methode der vollständigen Fixierung eines Textes durch eine linguistisch-statistische Analyse beschrieben, die es erlaubt, sowohl Information über die Statik, wie auch über die Dynamik von linguistischen Einheiten im Text mit einem beliebigen Grad der Ausführlichkeit zu erhalten. Was den allgemeinen Arbeitsaufwand betrifft, so ist diese Methode nur unbedeutend komplizierter als die gewöhnlichen Prozeduren zur Erstellung eines Häufigkeitswörterbuches. Die Methode wird in Bezug auf die gewöhnliche Arbeitstechnik mit Kärtchen beschrieben; ihre Übertragung auf maschinelle Verfahren bereitet keine Probleme.

2.

Nachdem eine Zähleinheit ausgewählt wurde (z.B. beschließt man, den Text auf der Ebene der Lexeme zu untersuchen; doch man kann auch irgendeine andere Einheit wählen), wird der Text auf vorher durchnumerierte Kärtchen übertragen. Da das Numerieren der Kärtchen arbeitsaufwendig und relevant ist, (nicht rechtzeitig bemerkte Fehler in der Numerierung lassen sich nicht korrigieren) sollte man einen zumindest halbautomatischen Numerator verwenden. Die Numerierung beginnt bei Eins, und jede Wortverwendung wir auf diese Weise mit der Nummer ihrer Position im Text verbunden. Nach der üblichen alphabetischen Anordnung stehen alle Kärtchen zusammen, auf denen ein und dasselbe Wort notiert ist. Wenn die Anordnung korrekt durchgeführt wurde, sind auch die Nummern der Positionen, an denen ein gegebenes Wort im Text auftritt, entweder in aufsteigender oder in umgekehrter Reihenfolge angeordnet (in Abhängigkeit von der Zahl der Sortierungen). Diese Kärtchen werden durchgezählt, und für die Wortkartei wird das Kärtchen mit der kleinsten Positionsnummer, die das erste Auftreten des gegebenen Wortes im Text anzeigt, ausgewählt. Auf diesem Kärtchen wird die volle Anzahl aller Vorkommen des gegebenen Wortes im Text angegeben (die Häufigkeit), und (auf der Rückseite der Kärtchen) die Nummern aller übrigen Positionen in steigender Reihenfolge, an denen das Wort erscheint (wenn ein Wort sehr häufig vorkommt, kann man einen zusätzlichen Papierstreifen ankleben und ihn in Ziehharmonikaform zusammenlegen). So erhält man ein alphabetisches Häufigkeitswörterbuch, vergleichbar mit einem gewöhnlichen Index. Der Unterschied liegt darin, daß es mit den Positionsnummern schwieriger ist, ein beliebiges Wort im Text zu finden (deshalb sollte man die Wörter auch im Text durchnumerieren). Jedoch gibt es auch einen Vorteil: während man bei der Arbeit mit dem Index unbedingt über eine ganz konkrete Textausgabe verfügen muß, ist die Arbeit mit dem Wörterbuch, in dem die Positionsnummern fixiert sind, mit jeder beliebigen

Textausgabe möglich. Es bleibt zu bemerken, daß der Unterschied im Arbeitsaufwand bei der Verwendung beider Arten des Wörterbuches fast völlig verschwinden kann, wenn während der Übertragung auf die Kärtchen die Positionsnummer des ersten Wortes auf jeder folgenden Seite festgehalten wird. Verfügt man über ein solches Verzeichnis, kann man die Seite leicht finden, der Rest macht nicht viel Arbeit. Es versteht sich, wie auch im Falle eines herkömmlichen Indexes, daß ein solches Verzeichnis nur für die Textausgabe gültig ist, nach der die Übertragung auf die Kärtchen vorgenommen wurde.

Nachdem die alphabetische Liste erstellt ist, kann die Wortkartei nach der Größe der Positionsnummern des ersten Auftretens der Wörter im Text (in zunehmender Reihenfolge) zusammengestellt werden. Dadurch werden die Wörter in der Reihenfolge ihres ersten Auftretens im Text geordnet. Es ist unschwer zu sehen, daß eine solche Liste eine äußerst ausführliche Information über das Anwachsen der Zahl der verschiedenen Wörter mit dem Anwachsen des Textumfanges enthält. Beispielsweise befindet sich das Wort HERMANN, das in A.S. Puškins "Pique Dame" zum ersten Mal auf Platz 130 erscheint, in der Wortliste, die in der Reihenfolge des ersten Auftretens geordnet ist, auf dem 98. Platz. Das bedeutet, daß die ersten 130 Wortverwendungen im Text einen Wortschatz von 98 Wörtern ausmachen. Auf diese Weise läßt sich der Prozeß des Anwachsens des Wortschatzes durch eine solche Liste mit einer Genauigkeit bis zu einem Wort verfolgen. Es bleibt zu bemerken, daß das Ordnen der Kärtchen bedeutend weniger arbeitsaufwendig und weniger mit Fehlern behaftet ist (auftretende Fehler sind sichtbar und lassen sich korrigieren) als die Prozedur der Vereinigung und des Durchzählens des Wortschatzes der Teilstichproben. Ist die umgeordnete Kartei fertiggestellt, muß man die Häufigkeit jedes Wortes und die Nummer des ersten Auftretens des Wortes aufschreiben (s. Anhang 1); die Positionsnummern aller übrigen Vorkommen des Wortes im Text (sofern dies nicht für ein bestimmtes spezielles Ziel

der Untersuchung erforderlich ist) braucht man dann nicht mehr festzuhalten, da sie in der alphabetischen Liste fixiert sind.

Als Illustration der Arbeitsmöglichkeiten mit einer solchen Liste führen wir die Analyse eines Textabschnittes aus "Pique Dame" vor, der zwei benachbarte Episoden enthält: das qualvolle Warten Hermanns nach der Abfahrt der Gräfin und ihres Pflegekindes (von"Dvercy zachlopnulis'" [Die Türflügel wurden zugeschlagen] bis "Lampa slabo osveščala ich iz perednej" [Die Lampe beleuchtete sie schwach aus dem Vorzimmer her]) und seine Eindrücke im Schlafzimmer der Gräfin (von "Germann vošel v spal'nju" [Hermann betrat das Schlafzimmer] bis "...vmeste s Mongol'f'erovym šarom i Mesmerovym magnetizmom" [...gleichzeitig mit dem Ballon des Montgolfier und dem Mesmerschen Magnetismus.]). Jede Episode enthält 101 Wortverwendungen; ihre Positionen im Text reichen von 3420 bis 3621. Im Anhang 1 ist ein Teil der Wortschatzliste, die zu diesem Textabschnitt gehört, dargestellt. Eine horizontale Linie trennt die Wörter, die zur ersten Episode gehören, von den Wörtern der zweiten Episode.

Man kann leicht sehen, daß in der ersten Episode 26 verschiedene Wörter verwendet werden, die früher nicht im Text vorgekommen sind. Achtzehn werden im ganzen Text nur einmal verwendet (die Ziffern in Klammern), das heißt, sie kommen nur in dieser Episode vor. In der zweiten Episode werden 56 verschiedene Wörter benutzt, davon 44 nur ein einziges Mal im ganzen Text. Mit anderen Worten: die lexikalische Sättigung beider Episoden ist sowohl hinsichtlich der neuen, vorher nicht verwendeten Wörter, als auch bezüglich der Wörter, die nur in dieser Episode benutzt werden, wesentlich anders. Unwillkürlich entsteht der Gedanke: ist eine derartige Dosierung des Wortschatzes nicht gewissermaßen ein Mittel der künstlerischen Ausdrucksfähigkeit, das auf den Leser unbewußt wirkt? Konnte den Puškin zuvor noch nicht verwendete Wörter für die Beschreibung der Hermann umgebenden Gegenstände finden, während dieser die vereinbarte Stunde abwartet (wie es oft

Schriftsteller geringeren Formates tun)?

Eine ähnliche Dosierung in viel größeren Maßstäben beobachtet man im Anfangsteil der "Kosaken" von L.N. Tolstoj. In der Arbeit von I.Š. Nadarejšvili und Ju. Orlov (1971) ist eine Graphik der Zunahme des Wortschatzes in den ersten 10000 Wortverwendungen dieses Textes dargestellt. Aus dieser Graphik ist ersichtlich, daß in den ersten 4000 Wortverwendungen das Anwachsen des Wortschatzes ein wenig verlangsamt ist (jedenfalls im Vergleich mit der theoretischen Kurve, die in dieser Graphik dargestellt ist). Dieser Textabschnitt beschreibt die Reise Olenins in den Kaukasus - eine einförmige weiße Ebene, abgelegene Städtchen, eines wie das andere, selbst Olenin schlummert auf dem Weg dahin. Aber dann kommt er im Kaukasus an - und die Welt flammt in grellen Farben auf. Diese ersten kaukasischen Eindrücke entsprechen einem Abschnitt steilen Ansteigens der Kurve des Wortschatzwachstums (die Wortverwendungen von 5000 bis 6000), wonach der Verlauf der empirischen Kurve flach wird und sie beginnt, kongruent zu "ihrer" theoretischen Kurve zu verlaufen (eine ausführliche quantitative Analyse dieser Situation wird im Anhang 2 betrachtet).

Auf diese Weise erlaubt die Anordnung der Wortkartei in der Reihenfolge, in der die Wörter zum ersten Mal im Text erscheinen, nicht nur mit einem festgelegten Ausführlichkeitsgrad die rein quantitative Seite des Wachstumsprozesses des Wortschatzes mit dem Wachsen des Textes zu beschreiben, sondern sie gibt auch die Möglichkeit, philologische und linguistische Probleme zu stellen und zu lösen. Ein Wörterverzeichnis solchen Typs ist außerordentlich eng mit dem Text verbunden: mit seiner Hilfe kann man leicht Probleme lösen, wie die Voraussage der Position neuer (für den gegebenen Text) Wörter im Satz oder deren Verteilung auf die Positionen eines Gedichttextes[1] und ähnliches mehr.

Die Technik des Numerierens der Kärtchen und die Herausstellung der Positionen des ersten Erscheinens eines Wortes im Text verschafft auch bei der Organisation der Häufigkeits-

listen der Lexik neue Möglichkeiten. Wenn man eine Wortkartei, die nach der Reihenfolge des ersten Erscheinens der Wörter angeordnet ist, nach abnehmenden Häufigkeiten aufstellt, ordnen sich auch alle Wörter mit gleicher Häufigkeit (wenn die Aufstellung exakt durchgeführt wurde) "von selbst" in der Reihenfolge ihres ersten Erscheinens im Text. Das heißt, anstatt der gewöhnlich praktizierten alphabetischen Anordnung der Wörter mit derselben Häufigkeit, die keinen philologischen oder linguistischen Sinn hat, entsteht eine Anordnung, die von dem untersuchten Objekt selbst diktiert wird und die Besonderheiten seiner Struktur widerspiegelt. Einer solchen Anordnung unterliegen alle seltenen Wörter, die insgesamt den Löwenanteil des Wortschatzes eines Textes bilden. Im Bereich der häufigen Wörter jedoch sind die Beziehungen zwischen der Häufigkeit eines Wortes und der Positionsnummer seines ersten Erscheinens von Interesse. Es ist klar, daß die häufigsten Wörter relativ früh zum ersten Mal erscheinen müssen (d.h., die Positionsnummern ihrer ersten Auftritte werden klein sein). Wenn aber ein häufiges Wort eine hohe Positionsnummer des ersten Erscheinens im Text hat, so wird dieses bedeuten, daß das betreffende Wort nur für einen bestimmten Teil des Textes spezifisch ist. Die vorher in der alphabetischen Liste festgehaltenen Nummern aller Positionen des Auftretens eines Wortes erlauben es, die Besonderheiten und Grenzen seiner Verwendungen genau zu bestimmen; insbesondere können solche Aufgaben leicht gelöst werden, wie sie in den Arbeiten von Tokarev & Jakubajtis (1969), Bektaev & Luk'janenkov (1971) oder Kaširina (1974) behandelt wurden; auf diese Weise ergibt sich die Möglichkeit, eine beliebige Unterteilung des Textes in Teilstichproben zu wählen und die Resultate zu vergleichen, die bei verschiedenen Untersuchungen erhalten wurden.

Vom Standpunkt einer quantitativen Analyse aus erlaubt ein auf diese Weise organisiertes Häufigkeits-Wörterverzeichnis nicht nur die Dynamik des Erscheinens der Wörter verschiedener Häufigkeiten zu untersuchen, sondern auch, Informationen

über die Häufigkeitsstrukturen eines beliebigen Textabschnittes zu entnehmen, was ein gewöhnliches Häufigkeitswörterbuch prinzipiell nicht erlaubt. Ebenso ist die Untersuchung der Positionen der Wörter mit geringer Häufigkeit (vg. Boroda & Nadarejšvili & Orlov & Čitašvili 1977; G.Š. Nadarejšvili & I.Š. Nadarejšvili & Orlov 1975) möglich.

3.

Auf den ersten Blick scheint es, daß die Technik der Kärtchennumerierung nur bei der Analyse einzelner Texte Sinn hat, wo die Position jedes Wortes nicht von der Willkür des Forschers abhängt. Jedoch erlaubt nur eine Untersuchung der Kurve des Wortschatzwachstums innerhalb eines Korpus verschiedener Texte und die Aufdeckung der Bruchstellen der Kurve an den Grenzen zwischen Texten, begründete Schlußfolgerungen über den Grad der statistischen Homogenität des untersuchten Korpus zu ziehen (s. Anhang 2 und auch Orlov 1977; Nešitoj 1976; Tuldava 1971). Wenn bei der Erstellung der Kärtchen die Positionsnummern fixiert worden sind, die den Grenzen des Überganges von Text zu Text entsprechen, wird es möglich, aus dem beschriebenen Wortschatz auch die Information zu entnehmen, die eine sogenannte Verteilungs-Wortliste enthält (vg. Andreev 1965). Besonderes Interesse gewinnt die Anwendung einer ähnlichen Technik dann, wenn man das Korpus der Dokumente chronologisch anordnen kann. In diesem Fall wird im Grunde genommen ein bestimmter Prozeß nichtlinguistscher Natur, der sich in der Zeit abspielt, mit linguistischen Methoden untersucht.[2)]

Die vorgeschlagene Form der Fixierung von Ergebnissen linguistischer Berechnungen in drei Wörterlisten:

1) der alphabetischen Liste mit Angabe der Häufigkeit und der Nummer aller - in zunehmender Reihenfolge angeordneter -

Positionen des Auftretens eines gegebenen Wortes,

2) der in der Reihenfolge des ersten Auftretens im Text angeordneten Liste der Wörter (mit Angabe der Häufigkeit) und

3) der Häufigkeitsliste (mit Angabe der Nummer des ersten Auftretens eines Wortes im Text und mit der Anordnung - in zunehmender Reihenfolge der Nummer - der Wörter, die die gleiche Häufigkeit haben),

gibt in bequemer und übersichtlicher Weise vollständige Information über einen Text. Diese Information ist für die verschiedenartigsten Analysen geeignet, sowohl für eine quantitative, als auch für eine traditionell-philologische Analyse. Es hat keinen Sinn, sich darüber zu streiten, welche Art von Wortlisten: der Index, die Häufigkeits- oder die Verteilungsliste besser ist. Bei der Erstellung einer beliebigen von ihnen ergeben sich ein und dieselben, außerordentlich arbeitsaufwendigen Arbeitsgänge. Deshalb ist es sehr wichtig, die Resultate dieser Arbeitsgänge in solcher Form zu fixieren, daß nichts verlorengeht, so daß die Resultate der Berechnungen nicht nur von ihrem Verfasser (der in der Regel ein ziemlich begrenztes Ziel verfolgt) benutzt werden können, sondern auch von anderen Forschern. Die quantitative, mathematische Linguistik macht gerade die ersten Schritte, die Zeit weitreichender Verallgemeinerungen und fundierter Hypothesen liegt noch vor ihr. Aber damit diese Zeit kommen kann, muß man Datenmaterial in sinnvoller Weise ansammeln, damit das heute bereits Erreichte nicht für die Zukunft verlorengeht und damit die Resultate der Berechnungen mit einer solchen Vollständigkeit und Sorgfältigkeit festgehalten werden, wie sie seit je her die klassische Philologie auszeichnete.

Berücksichtigt man die Schwierigkeiten bei der Veröffentlichung von drei Wörterlisten, so wäre es vernünftig, in der Praxis eine Deponierung solcher Arbeiten und die Herausgabe von Referatsammlungen einzurichten. In das Referat müssen die grundlegenden Parameter eines Textes eingehen: der Umfang eines Wortschatzes , sowie eine vollständige Liste aller Häufigkeiten (Häufigkeitsspektrum) und (wünschenswerterweise) eine Liste der häufigsten Wörter, ebenso wie Angaben über das Anwachsen des Wortschatzes mit dem Wachsen des Textumfanges (z.B. gibt man den Umfang des Wortschatzes nach allen 1000 oder 10000 Wortverwendungen an). Als Minimum sollte eine alphabetische Liste mit der Angabe der Nummern aller Positionen deponiert werden; es wäre natürlich besser, wenn alle drei Listen deponiert wären und darüber hinaus Angaben über die Positionsnummern der ersten Wörter auf jeder Textseite, über die Einteilung des Textes in Kapitel, Bücher, Bände u.ä. und, wenn ein Korpus von Dokumenten ausgezählt wird, über die Grenzen zwischen den einzelnen Dokumenten hinzugefügt wären. Eine derartige Praxis könnte die Entwicklung der statistischen Linguistik wesentlich beschleunigen.

Die Autoren danken V.S. Perebejnos und den Mitarbeitern der von ihr geleiteten Abteilung für strukturell-mathematische Linguistik des Institutes für Sprachwissenschaft der Ukrainischen Akademie der Wissenschaften für die ausführliche Besprechung der vorgeschlagenen Methode der Fixierung von Ergebnissen linguistischer Zählungen.

ANHANG 1

Wir führen einen Teil des Wortschatzes von "Pique Dame" an, der zu den untersuchten Episoden im Text gehört. Die Wörter sind in der Reihenfolge ihres ersten Auftretens (vom Textanfang an) angeordnet. Die Positionsnummer des ersten Auftretens findet man in der rechten Spalte. Die laufende Nummer des Wortes, die in der linken Spalte angegeben ist, stellt gleichzeitig den Umfang des Wortschatzes dar, der im Augenblick des Auftretens des gegebenen Wortes vorliegt. Insbesondere ist der Wortschatz, der im Text von "Pique Dame"

bis zum Anfang der ersten Episode vorliegt, gleich 1167 Wörter; bis zum Ende der zweiten Episode, im Umfang von 3621 Wortverwendungen, sind es 1250 Wörter. In den runden Klammern ist die absolute Häufigkeit eines Wortes im Text angegeben.

Lfd. Nr./ Auftreten		Pos.-Nr.
1168	zachlopnut'sja [zuschlagen] (1)	3423
1169	tjaželo [schwierig] (1)	3425
1170	pokatit'sja [rollen] (2)	3426
1171	zaperet' [abschließen] (1)	3430
1172	pomerknut' [erlischen] (1)	3433
1173	opustevšij [leer] (1)	3438
1174	časy [Uhr] (5)	3446
1175	dvadcat' [zwanzig] (2)	3448
1176	časovoj [Uhr-] (1)	3457
1177	strelka [Zeiger] (1)	3458
1178	vyžidat' [abwarten] (1)	3460
1179	ostal'noj [übrig] (1)	3461
1180	rovno [genau] (2)	3463
1181	stupit' [treten] (1)	3468
1182	vzojti [hinaufsteigen] (2)	3473
1183	jarko [hell] (1)	3475
1184	vzbežat' [hinauflaufen] (1)	3482
1185	spjaščij [schlafend] (1)	3492
1186	lampa [Lampe] (9)	3494
1187	zapačkannyj [beschmutzt] (1)	3497
1188	kreslo [Sessel] (1)	3498
1189	legkij [leicht] (1)	3499
1190	tverdyj [fest] (2)	3501
1191	šag [Schritt] (2)	3502
1192	temnyj [dunkel] (3)	3511
1193	slabo [schwach] (1)	3513
1194	osveščat' [beleuchten] (1)	3514
1195	kivot [Ikonenschein] (1)	3526
1196	napolnennyj [angefüllt](2)	3527
1197	obraz [Bild] (1)	3529
1198	teplit'sja [glimmen] (1)	3530
1199	zolotoj [golden] (1)	3531
1200	lampada [Öllämpchen] (1)	3532
1201	polinjalyj [verblichen] (1)	3533
1202	štofnyj [Stoff-] (2)	3534
1203	divan [Diwan] (3)	3537
1204	puchovyj [Daunen-] (1)	3539
1205	poduška [Kissen] (1)	3540

Lfd. Nr./ Auftreten		Pos.-Nr.
1206	sošedšij [verschossen] (1)	3542
1207	pozolota [Vergoldung] (1)	3543
1208	pečal'nyj [betrübt] (2)	3546
1209	simmetrija [Symmetrie] (1)	3547
1210	stena [Wand] (2)	3549
1211	obityj [ausgeschlagen] (1)	3550
1212	kitajskij [chinesisch] (1)	3551
1213	viset' [hängen] (1)	3555
1214	portret [Potrait] (3)	3557
1215	pisannyj [geschrieben] (1)	3558
1216	M-me [Madame] (1)	3561
1217	Lebrun [Lebrun] (1)	3562
1218	izobražat' [darstellen] (2)	3566
1219	sorok [vierzig] (3)	3569
1220	rumjanyj [gerötet] (1)	3570
1221	svetlozelenyj [hellgrün] (1)	3572
1222	mundir [Uniform] (1)	3573
1223	zvezda [Stern] (1)	3576
1224	orlinyj [Adler-] (1)	3581
1225	nos [Nase] (1)	3582
1226	začesannyj [glattgekämmt] (1)	3584
1227	visok [Schläfe] (1)	3585
1228	roza [rosa] (1)	3588
1229	pudrenyj [gepudert] (1)	3590
1230	volosy [Haare] (1)	3591
1231	torčat' [stecken] (1)	3595
1232	farforovyj [Porzellan-] (1)	3596
1233	pastušok [Hirtenjunge] (1)	3597
1234	stolovyj [Tisch-] (1)	3598
1235	slavnyj [berühmt] (3)	3601
1236	Leroy [Leroy] (1)	3602
1237	korobočka [Schächtelchen] (1)	3603
1238	ruletka [Roulett] (1)	3604
1239	veer [Fächer] (1)	3605
1240	raznyj [verschieden] (1)	3607
1241	damskij [Damen-] (1)	3608
1242	igruška [Spielzeug] (1)	3609
1243	izobretennyj [erfunden] (1)	3610
1244	konec [Ende] (1)	3612
1245	minuvšij [vergangen] (1)	3613
1246	stoletie [Jahrhundert] (1)	3614
1247	Mongol'f'erov [Montgolfiers] (1)	3617
1248	šar [Ballon] (1)	3618
1249	Mesmerov [Mesmers] (1)	3620
1250	magnetizm [Magnetismus] (1)	3621

Betrachtet man die angeführte Liste, so kann man leicht einen weiteren Vorzug einer derartigen Organisation des Wort-

schatzes bemerken. Es ist das einzige wortgetreue Wörterverzeichnis, das es erlaubt, bestimmte typische Wortverbindungen wie "tverdyj šag" [fester Schritt], "zolotoja lampada [goldenes Öllämpchen], "damskaja igruška" [Damenspielzeug], "puchovaja poduška" [Daunenkissen], "orlinyj nos" [Adlernase] u.a. zu rekonstruieren. Dabei geschieht diese Rekonstruktion unabhängig von der Häufigkeit der gegebenen Verbindung.

ANHANG 2

ANALYSE DER KURVEN DES WORTSCHATZWACHTUMS

Die Anordnung des Wortschatzes in der Reihenfolge des ersten Auftretens der Wörter im Text ermöglicht es mit äußerster Detailliertheit , buchstäblich wortgetreu, den Prozeß des Anwachsens des Wortschatzes mit dem Wachsen des Textumfangs zu untersuchen. Wenn solche Angaben systematisch veröffentlicht (oder zumindestens deponiert) werden, so eröffnet dies mit der Zeit ein breites Feld für eine vergleichende Analyse, wodurch das gegenwärtige Problem der Nicht-Vergleichbarkeit der Resultate (eine Stichprobe enthält 5000 Wortverwendungen, eine andere 25000, eine dritte 101000) vollends wegfallen. Jedoch ist auch schon heute eine fruchtbare theoretische Analyse der Abhängigkeiten des Wortschatzes von der Textlänge möglich, und zwar mit Hilfe eines Vergleiches der tatsächlich beobachteten Daten mit den vorhandenen theoretischen Modellen.

Aus allen vorgeschlagenen Methoden der Erstellung der theoretischen Kurven des Wortschatzwachstums mit dem Wachsen des Umfangs der Stichprobe muß besonders die Formel von V.M. Kalinin (1964) hervorgehoben werden:

$$Ev(N) = Ev(N_o) - \sum_{j \geq 1} \left(1 - \frac{N}{N_o}\right)^j Ev^j(N_o) \qquad (1)$$

Hierbei ist $Ev(N_o)$ der Erwartungswert der Größe des Wortschatzes in einer Zufallsstichprobe mit einem Umfang von N_o Wortverwendungen; $Ev_j(N_o)$ ist der Erwartungswert der Anzahl der j-mal vorkommenden Wörter der betreffenden Stichprobe; $Ev(N)$ ist der Erwartungswert des Wortschatzes in einer Stichprobe mit einem beliebigen Umfang N. Da die Erwartungswerte $Ev(N_o)$ und $Ev_j(N_o)$ gewöhnlich nicht bekannt sind, egal für welchen Umfang N_o (wir können nur Hypothesen über ihre Größe aufstellen; vgl. Orlov 1977), scheint es, daß die Formel (1) keinen praktischen Wert besitzt. Wenn man jedoch statt der Erwartungswerte $Ev(N_o)$ und $Ev_j(N_o)$ die tatsächlich beobachteten Werte des Wortschatzumfanges $v^*(N_o)$ und die Zahl der j-mal vorkommenden Wörter $v_j^*(N_o)$ in der Stichprobe mit dem Umfang N_o einsetzt, gibt diese Formel den genauen Wert des Erwartungswertes $Ev(N)$ für eine beliebige zufällige Teilstichprobe an[3], d.h. N muß kleiner als N_o sein. Mit anderen Worten: eine Prognose aus den beobachteten Werten des Wortschatzes und des Spektrums ist nach der Formel (1) nur "rückwärts" möglich. Im Falle $N > N_o$ ist eine Prognose "vorwärts" für die beobachteten Werte nach der Formel (1) unzuverlässig.

Die Formel Kalinins ist für die Beschreibung des Prozesses des Anwachsens der Anzahl beliebiger Einheiten geeignet, da sie nicht von deren Verteilung abhängt. Eine Abweichung der tatsächlich beobachteten Kurve des Anwachsens eines Wortschatzes von der Formel (1) muß man als Folge der Nichtzufälligkeit der Teilstichproben ansehen. Mit anderen Worten: bei einer Textanalyse kann der einzige Grund für die Divergenz zwischen dem theoretischen und dem tatsächlichen Anwachsen des Wortschatzes nur die statistische Heterogenität des Textes sein. Dies macht die Formel Kalinins zu einer Art Eichmaß für die Überprüfung der Hypothese der Homogenität eines gegebenen Textes (oder eines Korpus von Texten).

Die Formel Kalinins beschreibt das Anwachsen des Wortschatzes in einem ideal strukturlosen Text. Einen solchen "Text" kann man zum Beispiel durch sorgfältiges Mischen der

beschrifteten Kärtchen erhalten. Wenn man diese Kärtchen in zufälliger Reihenfolge durchnumeriert (nach dem Mischen), muß die Kurve des Anwachsens des Wortschatzes in einem solchen "Text" genau mit der nach Formel (1) berechneten übereinstimmen, wobei sie von ihr geringfügig nach beiden Seiten aufgrund zufälliger Streuungen abweicht.

Interessant ist, daß in der statistisch-linguistischen Praxis nicht selten Prozeduren durchgeführt werden, deren Ergebnis sich zu einem gewissen Grad dem oben beschriebenen Mischen annähert. So insbesondere bei Untersuchungen des Wortschatzes in Stichproben von unterschiedlichem Umfang aus wissentlich heterogenem Material (z.B. Zeitungssprache), wenn in den Teilstichproben die Texte aus verschiedenen Zeitungsrubriken in den gleichen Proportionen wie im gesamten Materialkorpus auftauchen. Das gilt auch für sogenannte "mechanische" Stichproben, wenn die Teilstichproben durch Zerteilung des Textes in gleiche Fragmente gebildet werden, die sich in gleichem Abstand voneinander befinden (z.B. immer das obere Drittel einer Seite). Im Grunde genommen sind diese Prozeduren vollkommen überflüssig, da sich ihre Ergebnisse mit der Formel Kalinins genau vorhersagen lassen. Wenn dennoch eine Divergenz vorliegt, dann bedeutet dies nur, daß die durchgeführte Prozedur des "Mittelns" der Teilstichprobe nicht "repräsentativ" ist. Dies ist der Fall, in dem man den theoretischen Berechnungen mehr glauben kann, als den experimentellen Angaben.

Wenn man also über das experimentell beobachtete Häufigkeitsspektrum $v_m^*(N)$; $m = 1, 2, \ldots$ in einer Stichprobe mit Umfang N_o und über die Werte der Wortschätze in seinen Teilstichproben $v^*(N)$ $(N < N_o)$ verfügt, dann kann man unter Verwendung der Formel (1) den Grad der Homogenität der Teilstichproben abschätzen (für einen zusammenhängenden Text: den Grad der Homogenität des Textes) oder, wenn die Teilstichproben oder der Text bewußt heterogen sind (zum Beispiel: der Text ist in zwei Teilstichproben geteilt, nämlich in Autorensprache und direkte Rede der Personen)

den Einfluß dieser Heterogenität auf den Wortschatz beurteilen. Es genügt, nach Formel (1) den theoretischen Wortschatz in den Umfängen der Teilstichproben zu berechnen und ihn mit dem beobachteten Wortschatz zu vergleichen.

Der Idee nach analoge Prozeduren zur Beurteilung des Grades der Heterogenität wurden auch in anderen Arbeiten vorgeschlagen (Orlov 1977, Nešitoj 1976, Tuldava 1971). Sie alle basieren auf empirisch bestätigten Kurven des Anwachsens des Wortschatzes (mit einem oder zwei Parametern), wodurch es ausreichend ist, nur die Wortschätze in unterschiedlichen Umfängen zu kennen (im Unterschied zur Formel Kalinins, für die die Werte des Häufigkeitsspektrums der Stichprobe erforderlich sind). Jedoch gibt es in allen drei Fällen keine Garantie, daß die beobachtete Divergenz zwischen den empirischen und den theoretischen Wortschätzen nur durch die Heterogenität des Textes und nicht auch durch die Unvollkommenheit der Formeln selbst hervorgerufen wurde.

Betrachten wir von diesem Standpunkt aus die Leistung der sich hierzu empfehlenden Formel des Anwachsens des Wortschatzes (Nadarejšvili & Orlov 1971; Orlov 1977):

$$v(N,Z) = v(Z)\,\frac{\ln\frac{Z}{N}}{\frac{Z}{N} - 1}, \text{ wobei } v(Z) = \frac{Z}{\ln(Zp_1)} \qquad (2)$$

Diese Formel wurde in Orlov (1976, 1977) als Ergebnis der Einsetzung eines hypothetischen Erwartungswertes des Häufigkeitsspektrums beim Stichprobenumfang Z in die Formel Kalinins erhalten; N ist der laufende Umfang; v(N,Z) ist der erwartete Umfang des Wortschatzes in einer Stichprobe mit dem Umfang N; v(Z) entspricht dem erwarteten Umfang des Wortschatzes in einer Stichprobe mit dem Umfang Z; p_1 ist die relative Häufigkeit des häufigsten Wortes im Text. Der Parameter Z kann geschätzt werden aus der Gleichung

$$v(N_o,Z) = v^*(N_o), \qquad (3)$$

wobei $v^*(N_o)$ der tatsächlich beobachtete Umfang des Wortschatzes in einer Stichprobe mit dem Umfang von N_o Wortverwendungen ist. Lösungsmöglichkeiten der transzendenten Gleichung (3) sind in Orlov (1977) gegeben; dort ist auch die Erstellung der Konfidenzintervalle für Prognosen nach Formel (2) durchgeführt.

Im Unterschied zur Formel (1) Kalinins ist die Formel (2) nicht nur für Prognosen "rückwärts", sondern auch "vorwärts" geeignet; darüberhinaus erfordert sie keine Kenntnis der Werte der Häufigkeitsstrukturen in der Ausgangsstichprobe. In Orlov (1977) ist das Beispiel einer zehnmaligen, sowohl vorwärts als auch rückwärts durchgeführten Prognose nach dieser Formel angegeben; eine dreißigfache Prognose vorwärts ist in Orlov (1976) enthalten. Selbst in den schlimmsten Fällen überschreiten die Fehler der Prognose nach der Formel (2) nicht die Grenze von 10 - 12 %. Jedoch können Divergenzen zwischen dem tatsächlich beobachteten Wortschatz und der Prognose nach der Formel (2) - im Unterschied zur Formel (1) - nicht nur infolge der Heterogenität eines Textes vorkommen, sondern auch infolge der Inadäquatheit (in Bezug auf den gegebenen Text) der bei der Ableitung der Formel (2) gewählten Hypothese über die Häufigkeitsstruktur dieses Textes. Daher charakterisieren die Divergenzen zwischen den theoretischen Prognosen, die man nach den Formeln (1) und (2) erhält, eben den Grad dieser Inadäquatheit. Praktisch geschieht es jedoch nicht selten, daß im Falle einer Divergenz zwischen den Formeln (1) und (2) die Formel (2) das tatsächliche Wachsen des Wortschatzes im Text besser beschreibt (die Approximation, die der Ableitung der Formel (2) zugrunde liegt, scheint durch die Heterogenität realer Texte kompensiert zu werden, insbesondere durch die Häufung von Wörtern, die eine Verlangsamung des Wachsens des Wortschatzes hervorruft) (s. Boroda & Nadarejšvili & Orlov & Čitašvili 1977).

Wir illustrieren die Leistung der Formeln (1) und (2) am Beispiel des Anwachsens des Wortschatzes in "Pique Dame":

Tabelle 1: Anwachsen des Wortschatzes in Puškins "Pique Dame"

Umfang der Stichprobe	Umfang des Wortschatzes		
	tatsächl. $v^*(N)$	berechnet	
		nach Formel (1)	nach Formel (2)
500	281	308	239
1000	462	523	497
2000	787	846	845
4000	1348	1371	1360
6000	1752	1778	1770
6861	1928	1928	1930*

Die Berechnungen nach der Formel (1) wurden durchgeführt durch Einsetzung des beobachteten Häufigkeitsspektrums $v_m^*(N_o)$ beim Text von $N_{o*} = 6861$. Der tatsächliche Wortschatz in diesem Umfang $v^*(N_o) = 1928$ wurde auch in die Gleichung (3) eingesetzt, um den Parameter Z zu erhalten. Bei $p_1 = 0.038$ (tatsächliche Häufigkeit des Wortes "i" [und] in "Pique Dame") erhalten wir aus dieser Gleichung $Z \sim 35000$. Der mit dem tatsächlichen Wortschatz übereinstimmende theoretische Wert des Wortschatzes bei der Lösung von Gleichung (3) ist durch ein Sternchen gekennzeichnet. Alle übrigen Prognosen nach Formel (2) sind mit dem erhaltenen Z = 35000 berechnet worden.

Wie aus der Tabelle 1 ersichtlich ist, beschreiben beide Formeln das Anwachsen des Wortschatzes in dem gegebenen Text bei fast allen Umfängen völlig zufriedenstellend. Lediglich im Anfangsteil des Textes (ungefähr die ersten zwei Kapitel, bis zum Beginn von Hermanns Intrige) ist das tatsächliche Anwachsen des Wortschatzes ein wenig verlangsamt, im Vergleich zu den Prognosen nach Formel (1). Dies kann, wie oben bemerkt, nur durch die Heterogenität des

zusammenhängenden Textes erklärt werden. Die Prognosen nach der Formel (2) kommen dem tatsächlichen Wachsen des Wortschatzes etwas näher, aber auch in diesem Fall macht sich der Einfluß der Heterogenität des Textes bemerkbar.

Ein "plastischeres" Bild der Divergenz zwischen den Formeln und dem tatsächlichen Anwachsen des Wortschatzes kann man in den ersten 10000 Wortverwendungen von "Die Kosaken" von L.N. Tolstoj beobachten (Tabelle 2):

Tabelle 2: Anwachsen des Wortschatzes in Tolstojs "Die Kosaken"

Umfang der Stichprobe	Umfang des Wortschatzes		
	tatsächl. $v^*(N)$	berechnet	
		nach Formel (1)	nach Formel (2)
500	274	337	298
1000	438	548	513
2000	732	922	862
4000	1233	1472	1395
6000	1814	1904	1840
8000	2253	2266	2225
10000	2582	2582	2600*

In diesem Falle wurden für die Berechnungen ebenfalls der Wortschatz und das Spektrum im Gesamtumfang verwendet; die Einsetzung von $v^*(10000) = 2582$ in die Gleichung (3) ergibt (bei $p_1 = 0.05$) $Z \sim 53000$. Wie aus Tabelle 2 ersichtlich wird, ist die Divergenz zwischen dem tatsächlichen Anwachsen des Wortschatzes in "Die Kosaken" und den Prognosen nach beiden Formeln wesentlich größer als im Falle von "Pique Dame". Dies erklärt sich aus einer sehr ausgeprägten Heterogenität der untersuchten Stichprobe, die aus zwei großen kontrastierenden Teilen besteht (s. zugrunde liegender Text). Wie im vorhergehenden Fall beschreibt die Formel (2) das tatsächliche Anwachsen des Wortschatzes ein wenig besser als die Formel (1). Das angeführte Beispiel ist lehrreich in dem Sinne, daß man aus

dieser Tatsache in keiner Weise den Schluß ziehen kann, daß die Formel Kalinins schlechter ist als die Formel (2).

Liegt bei zufälliger Wortwahl aus einem Text mit gegebener Häufigkeitsstruktur der tatsächliche Wortschatz deutlich unter dem erwarteten, so zeugt dies davon, daß bei aufeinanderfolgenden Stichproben vom Anfang des Textes an der Platz vieler neuer Wörter von "überflüssigen" Wiederholungen bereits verwandter Wörter besetzt ist (würde man dagegen beobachten, daß der tatsächliche Wortschatz im Anfangsteil eines Textes über dem erwarteten liegt, so hieße dies, daß der Autor absichtlich die Verwendung bereits am Anfang des Textes benutzter Wörter vermeidet, mit dem Ziel, sie am Ende zu wiederholen). Das heißt, der "Eichmaß"-Charakter der Formel Kalinins erlaubt es, im Falle von Divergenzen positive Aussagen über die Struktur eines Textes zu machen.

Die Formel (2) weicht in diesem Falle ein wenig von der Formel Kalinins ab. Das bestätigt, daß die Häufigkeitsstruktur des Auszuges aus "Die Kosaken" sich geringfügig von der unterscheidet, die bei der Ableitung der Formel (2) gefordert worden war. Man kann vermuten, daß der Anfangsteil des Textes auf die Häufigkeitsstruktur des untersuchten Auszuges verzerrend wirkt und daß sich bei weiterer Vergrößerung des Stichprobenumfangs der Unterschied zwischen den Prognosen nach den Formeln (1) und (2) verringern wird. Aber auch auf Grund eines Vergleichs des tatsächlichen Anwachsens des Wortschatzes mit der Prognose nach der Formel (2) kann man auf die grundsätzliche Heterogenität des Textes schließen.

Außerordentlich interessant ist es, die Tabelle 2 mit den entsprechenden Angaben über "Die Auferstehung" von L.N. Tolstoj (ebenfalls die ersten 10000 Wortverwendungen vom Textanfang an, Tabelle 3) zu vergleichen. Der Verlauf der Prognosen nach der Formel (2) änderte sich nicht, da die der Berechnung des Parameters zugrunde liegenden Aus-

gangsangaben praktisch gleich sind (p_1 = 0.05, v^*(10000) = 2587). Ein wenig veränderte sich der Verlauf der Prognosen nach der Formel (1), wodurch eine sehr gute Übereinstimmung zwischen den Prognosen nach der Formel (1) und (2) entstand: ebenso gut stimmen mit diesen Prognosen die tatsächlichen Werte des Wortschatzes überein, was sowohl eine hohe Homogenität des Textes als auch eine Übereinstimmung der Häufigkeitsstruktur des Textes mit der Hypothese, die bei der Ableitung der Formel (2) angenommen worden war, bescheinigt.

Tabelle 3: Anwachsen des Wortschatzes in Tolstojs "Die Auferstehung"

Umfang der Stichprobe	Umfang des Wortschatzes		
	tatsächl. v^*(N)	berechnet	
		nach Formel (1)	nach Formel (2)
500	282	319	298
1000	489	545	513
2000	824	906	862
4000	1409	1380	1395
6000	1863	1893	1840
8000	2237	2262	2225
10000	2587	2587	2590*

Vom inhaltlichen Standpunkt aus ist der hohe Grad der Homogenität von "Die Auferstehung" ein wenig paradox, zumindest auf den ersten Blick. Der Anfangsteil des Romans besteht aus kontrastierenden, lexikalisch sehr unterschiedlichen Episoden, die das alltägliche Leben von Katjuša Maslova im Gefängnis und das Leben des Fürsten Nechljudov beschreibt. Jedoch wechseln sich diese Episoden in kleinen Abständen ab (mit anderen Worten: es scheint so, als ob L.N. Tolstoj selbst eine Mischung der Lexik durchgeführt hätte), daher auch die gute Übereinstimmung des tatsächlichen Anwachsens des Wortschatzes mit den Prognosen nach der Formel (1).

Die Verwendung der theoretischen Formeln erlaubt es auch, die Möglichkeiten einer vergleichenden Analyse von Texten zu bereichern. Wenn man zum Beispiel nur Umfang und Wortschatz von "Pique Dame" (N = 6861, v^* = 1928) und von den Auszügen aus "Die Kosaken" (N = 10000, v^* = 2582) kennt, kann man keine Folgerungen über den relativen Wortschatz (Sättigung) in diesen Texten ziehen. Die Kalininsche Kurve für "Die Kosaken" läuft sicherlich oberhalb der Kurve für "Pique Dame" (vgl. Tab. 1 und 2). Das bedeutet, daß sich in zufälligen Teilstichproben gleichen Umfangs aus beiden Texten der größere Wortschatz in den Teilstichproben aus "Die Kosaken" befinden wird, und daher eine bestimmte mittlere lexikalische Vielfalt, die wir als relativen Vokabularreichtum (lexikalische Sättigung, lexikalische Konzentration usw.) bezeichnen wollen, in "Die Kosaken" höher ist als in "Pique Dame". Die auf diese Weise erhaltene Schlußfolgerung ist sogar zuverlässiger als der direkte Vergleich gleicher, zusammenhängender Teilstichproben aus beiden Texten, da (z.B. in unserem Falle) beim Umfang von N = 2000 der tatsächliche Wortschatz von "Pique Dame" (787 Wörter) höher ist als der Wortschatz von "Die Kosaken" (732 Wörter). Dieselbe Folgerung über die höhere lexikalische Sättigung in "Die Kosaken" kann man auch aus dem Vergleich der theoretischen Prognosen nach der Formel (2) für beide Texte ziehen (bei der Arbeit mit der Formel (2) kann der Wert des Parameters Z als vereinbartes Maß des relativen Vokabularreichtums dienen. Je größer Z ist, umso größer ist die Zahl der verschiedenen Wörter, die in der zufälligen Stichprobe einer festgelegten Länge vorkommen. Genaueres siehe Orlov 1976, 1977).

Die angeführten Beispiele demonstrieren die Möglichkeiten einer statistisch-linguistischen Analyse, die bisher noch nicht verwertet wurden, obwohl die Arbeiten Kalinins schon vor mehr als zehn Jahren veröffentlichet worden sind.

Die Autoren danken R.Ja. Čitašvili (der die Formel (1a)

abgeleitet hat) und M.G. Boroda für ihre Teilnahme an der Beurteilung des Sinnes und der Möglichkeiten der Anwendung der Formeln (1) und (2).

ANMERKUNGEN

[1] In diesem Fall bedeutet die Position die Rangzahl des Wortes im Satz oder im Vers. Ein Beispiel der Untersuchung der Verteilung von Wörtern unterschiedlicher Häufigkeit in Verspositionen findet man in Nadarejšvili & Orlov 1969.

[2] Dieser Gedanke über die Möglichkeiten der beschriebenen Methode für eine Untersuchung der Dynamik der Informationsflüsse entstand während eines Gespräches eines der Autoren der vorliegenden Arbeit mit G.F. Krajčinskaja (Patenabteilung UKRNIIPLASTMAŠ, Kiev) über das Problem der Analyse von Texten, die Erfindungen beschreiben.

[3] Genau gesagt, wird eine Zufallsstichprobe aus einer endlichen Grundgesamtheit (ohne zurücklegung und Vermischung) durch die hypergeometrische Verteilung beschrieben (s. z.B. Bol'šev, Smirnov 1965:114), die zu folgenden Abhängigkeiten für das Häufigkeitsspektrum und den Wortschatz führt:

$$Ev_m(N) = \sum_{k=m}^{N_o-N-m} v_k^*(N_o) \frac{\binom{k}{k-m}\binom{N_o-k}{N_o-m}}{\binom{N_o}{N}}$$

$$Ev(N) = \sum_{k\geq 1} Ev_k(N) = \sum_{k\geq 1} v_k^*(N_o) \sum_{i=\max(k-N,0)}^{\min(N_o-N,k-1)} \frac{\binom{k}{i}\binom{N_o-k}{N_o-N-i}}{\binom{N_o}{N}}$$

$$= \sum_{k=1}^{N_o} v_k^*(N_o)\left(1-\frac{\binom{N_o-k}{N_o-N-k}}{\binom{N_o}{N}}\right) - \sum_{k\geq N+1} v_k^*(N_o) \sum_{i=0}^{k-N-1} \frac{\binom{k}{i}\binom{N_o-k}{N_o-N-k}}{\binom{N_o}{N}} \quad (1a)$$

Wenn N die absolute Häufigkeit des häufigsten Wortes beim Umfang N_o übersteigt, wird die letzte Summe zu Null und

$$Ev(N) = v^*(N_o) - \sum_{k\geq 1} v_k^*(N_o) \frac{(N_o-n-k+1)\ (N_o-N-k+2)\ldots(N_o-N)}{(N_o-k+1)\quad (N_o-k+2)\quad \ldots\ N_o} \quad (1b)$$

Jedoch ist in dem für die statistische Linguistik interessanten Bereich der Werte N und N_o der Unterschied zwischen den angeführten Formeln und der Formel (1), die man bei der Annahme der Richtigkeit der Poisson-Verteilung erhält, völlig unwesentlich. Er wird nur bei sehr kleinen N (der Ordnung Eins und einiger Zehner) wegen des Anwachsens der doppelten Summe im rechten Teil der Formel (1a) bemerkbar. Die Formeln (1a) und (1b) kann man in Zweifelsfällen zur Kontrolle der Formel (1) verwenden.

DYNAMIK DER HÄUFIGKEITSSTRUKTUREN

Ju. K. Orlov

Der Begriff der "Rangverteilung" ist auf dem Weg, zu einem allgemein akzeptierten Begriff zu werden. Er bezieht sich auf eine bestimmte Konstanz der Form einer geordneten Menge von Häufigkeiten der Elemente in einer gegebenen Stichprobenklasse. Eine vollständige Übersicht der gegenwärtigen Vorstellungen über diesen Begriff findet man in Arapov & Efimova & Srejder (1975).

Der Begriff der "Rangverteilung" ruft aber aus wenigstens zwei Gründen Einwände hervor. Erstens, das Wort "Verteilung" erinnert an die übliche statistische Vorstellung, für die die Sätze über Konvergenz, über statistische Stabilität usw. gelten, während in der Wirklichkeit mit der Vergrößerung des Stichprobenumfangs keine "Verbesserung" der beobachteten Häufigkeitssequenz und keine Konvergenz gegen eine "ideale" Form zustandekommt (vgl. Arapov & Efimova & Šrejder, 1975: 13). Zweitens hängt das Wort "Rang" nur mit einer gewissen Darstellungsweise für Relationen in einer Zahlenmenge zusammen. Im Grunde genommen ist der "Rang" eine Hilfsgröße ohne organischen Zusammenhang mit der Zahlenmenge. Außerdem ruft dieses Wort unerwünschte Assoziationen mit der Rangordnungsstatistik, mit Rangordnungskriterien usw. hervor. Aus diesen Gründen wird in dieser Arbeit der Begriff der "Häufigkeitsstruktur" verwendet, der die Mängel des Begriffs der Rangverteilung nicht aufweist.

Im ersten Teil der Arbeit erklären wir die stetige Approximation von diskreten Häufigkeitsstrukturen, bei der der Begriff des Ranges an sich entbehrlich ist und die es ermöglicht, die Zusammenhänge zwischen den Parametern der Struktur leicht zu finden. Diese Methode eignet sich im Grunde für die Beschreibung beliebiger Häufigkeitsstrukturen. Gleichzeitig zeigen wir den Übergang zur traditionellen Vorstellung der Häufigkeitsstruktur mit Hilfe einer rangierten Sequenz.

Im zweiten Teil werden mit dem eingeführten Formalismus die Häufigkeitsstrukturen des Zipf-Mandelbrotschen Typs analysiert.

Es wird eine allgemeine Form der Beschreibung ähnlicher Strukturen aufgestellt, ferner werden die Beziehungen zwischen den Parametern der Mandelbrotschen Formel

$$p_i = \frac{K}{(B+i)^\gamma}; \quad i = 1,2,..v \qquad (1)$$

sowie zwischen anderen beobachteten Parametern der Stichprobe, deren Häufigkeitsstruktur der Beziehung (1) folgt, dargestellt.

Im dritten Teil wird der interessante und theoretisch sowie praktisch äußerst wichtige Fall der kleinen Stichproben, zu denen beispielsweise alle lexikalischen Stichproben gehören, analysiert. Für diese Stichproben ist charakteristisch, daß ihr Vokabular das der Grundgesamtheit, aus der sie stammen, nur zu einem geringen Teil ausschöpft (als die erste Approximation an eine solche Grundgesamtheit kann man die Sprache als ganze oder eine Teilsprache betrachten). Solange wir nur über dürftige Kenntnisse der Grundgesamtheit verfügen, muß das obligatorische Vorkommen von hapax legomena (Wörter, die nur einmal vorkommen) als charakteristisches Attribut solcher Stichproben betrachtet werden. Es sind nämlich eben die hapax legomena, die die Zugehörigkeit dieser Stichproben zur Klasse der kleinen Stichproben bestimmen.

Im vierten Teil wird gezeigt, daß die Häufigkeitsstruktur kleiner Stichproben dynamisch ist und daß die Veränderungen dieser Struktur mit den Veränderungen ihres Umfangs gesetzmäßig verbunden sind. Die Dynamik der Häufigkeitsstrukturen, allgemein untersucht von Kalinin (1965), wurde bis jetzt beim Vergleich empirischer Daten mit theoretischen Modellen überhaupt nicht in Betracht gezogen. Die bekannten "Abweichungen" im Bereich der niedrigen Häufigkeiten führten zu dem Schluß, daß das Zipfsche Gesetz nur eine grobe Annäherung sei (vgl. Frumkina 1961), und stimulierten zur Ableitung komplizierter Formeln, die eine "bessere Approximation" an die empirischen Häufigkeitskurven darstellen sollten. Die im vierten Teil abgeleiteten Beziehungen beschreiben die Häufigkeitsstruktur eines statistisch homogenen Textes, der das Zipf-Mandelbrotsche Gesetz bei eindeutig bestimmtem Text-

umfang erfüllt. Es wird gezeigt, daß bei anderen Umfängen "Abweichungen an den Schweifen" üblicherweise unvermeidlich sind.

Im fünften, letzten Teil bringen wir ein Beispiel zur Berechnung des Vokabularwachstums und der Zahl der hapax legomena, sowie einen Vergleich mit empirischen Daten und mit Kalinins (1965) Berechnungen.

Die Problematik wird am linguistischen Material dargelegt, aber das konstruierte allgemeine Modell der Häufigkeitsstruktur des Zipfschen Typs kann auch bei der Lösung solcher Probleme der Informationsverarbeitungssysteme, der automatischen Steuerungssysteme, der Ökonomie, der Soziologie, der Biologie usw., die auf "Rangverteilungen" des Zipfschen Typs führen, Verwendung finden. Weitere Entwicklungen der dargelegten Methode ermöglichen es, auch zu grundsätzlich nicht-zipfschen Strukturen überzugehen, d.h. zu solchen, die nicht aus dem Zipf-Mandelbrotschen Gesetz abgeleitet werden (analog den Strukturen im vierten Teil der vorliegenden Arbeit).

1

Wir werden annehmen, daß die Häufigkeitsstruktur (≡ statistische Struktur) einer Stichprobe (Grundgesamtheit) durch die gesamte Menge der Häufigkeiten (Wahrscheinlichkeiten) der sie konstituierenden Elemente gegeben ist, ohne jegliche Anordnung der Elemente und ohne die Angabe, welche Häufigkeit zu welchem Element gehört. Diese Auffassung der Häufigkeitsstruktur unterscheidet sich wesentlich von der einer diskreten Verteilung. Sie ist in dem Sinne allgemeiner, daß unterschiedliche Verteilungen dieselbe Struktur haben können, während das Umgekehrte nicht gilt (stellt man beispielsweise im Frequenzwörterbuch die Wörter in Bezug zu den Häufigkeiten willkürlich um, so bekommt man eine andere Verteilung, während die Häufigkeitsstruktur erhalten bleibt).

Es interessierten uns im Grunde einige Beziehungen in einer Menge von Zahlen, deren Summe Eins ergibt. Diese Zahlen werden

wir bequemlichkeitshalber im weiteren als Häufigkeiten bezeichnen. Wir nehmen an, daß die Häufigkeitsstruktur einer Stichprobe bekannt ist, wenn man für ein beliebiges Intervall zeigen kann, mit welcher Wahrscheinlichkeit ein Element, dessen Häufigkeit in dieses Intervall fällt, zufällig gewählt werden kann. Um diese Aufgabe zu lösen, muß man eine stetige Approximation an die Häufigkeitsstruktur in den Fällen machen, wo die Zahl der Häufigkeitswerte groß ist und hinreichend dicht das ganze Intervall der Häufigkeitswerte der gegebenen Stichprobe ausfüllt.

Gegeben sei eine Stichprobe mit dem Umfang N aus v untereinander unterschiedlichen Elementen mit den Häufigkeiten $p_i^N = \frac{m_i}{N}$, wobei m_i die absolute Häufigkeit des i-ten Elements für i = 1,2,... ...v ist. Um weitere Überlegungen und auch den Übergang zu traditionellen Arten der Analyse der Häufigkeitsstrukturen zu ermöglichen, setzten wir voraus, daß die Menge $\{p_i^N\}$ abnehmend geordnet ist. Es interessiert uns nur das Problem der Approximation von $\{p_i^N\}$, sowie einige andere Charakteristika der Stichprobe unter der Bedingung, daß N und v so groß sind, daß die Summe der Menge der Häufigkeiten, die auf kleine Teilintervalle innerhalb des Intervalls $[p_v^N, p_1^N]$ entfallen, klein ist.

Sei $v_m(N)$ die Anzahl unterschiedlicher Elemente, von denen jedes die absolute Häufigkeit m besitzt. Entsprechend unserer Forderung bedeutet dies, daß die Größe

$$\sum_{p<\frac{m}{N}\leq p+\Delta p} v_m(N)\frac{m}{N} \qquad (2)$$

bei kleinem Δp klein ist.

Wir führen folgende Funktionen ein:

$$v^N(p) = \sum_{\frac{m}{N}\leq p} v_m(N) \qquad (3)$$

$$F^N(p) = \sum_{\frac{m}{N}<p} v_m(N)\frac{m}{N}. \qquad (4)$$

Bei Anwendungen auf lexikalische Stichproben stellt Funktion (3) die Zahl unterschiedlicher Wörter, deren relative Häufigkeiten in der gegebenen Stichprobe p nicht übersteigen, und Funktion (4) die Wahrscheinlichkeit der zufälligen Wahl eines beliebigen Wortes aus dieser Stichprobe, dessen relative Häufigkeit p nicht übersteigt, dar. Bei den gegebenen Annahmen ist es natürlich, diese Treppenfunktionen mit glatten stetigen Funktionen v(p) und F(p) zu approximieren, ähnlich wie man eine diskrete empirische Verteilungsfunktion üblicherweise durch eine glatte Approximation ersetzt. Aus der Definition folgt unmittelbar:

$$\sum_{p'<p_i^N\le p''} \frac{F^N(p_i^N)-F^N(p_{i-1}^N)}{p_i^N} = \int_{p'}^{p''} \frac{dF^N(p)}{p} = v^N(p'')-v^N(p'), \qquad (5)$$

wonach die Beziehung zwischen p_i^N und $F^N(p)$ durch die Minimum-Lösung der Ungleichung

$$\int_{p_i^N}^{p_1^N} \frac{dF^N(p)}{p} \le i-1, \qquad i = 1,2,\ldots v, \qquad (6)$$

bestimmt wird.

Wenn F(p) monoton wachsend ist, so kann man als Approximation für p_i^N die Lösungen der Gleichungen

$$\int_{p_i}^{p_1} \frac{dF(p)}{p} = i-1, \qquad p_i \approx p_i^N \qquad (7)$$

verwenden. Diese Funktion kann man auch für die Schätzung der Entropie verwenden:

$$H = -\sum_i p_i^N \ln p_i^N = -\sum_m \frac{m}{N} v_m(N) \ln \frac{m}{N} =$$

$$= -\int_{p_v^N}^{p_1^N} \ln(p)\,dF^N(p) \approx -\int_{p_v}^{p_1} \ln(p)\,dF(p). \qquad (8)$$

Aus der Ungleichung

$$\frac{1}{p''}[F^N(p'')-F^N(p')] \le v^N(p'')-v^N(p') \le \frac{1}{p'}[F^N(p'')-F^N(p')], \qquad (9)$$

die bei $p' < p''$ offensichtlich gilt, finden wir den Zusammenhang zwischen v(p) und F(p), wenn $p' \sim p''$, als

$$\frac{dF(p)}{p} = dv(p); \qquad (10)$$

daraus erhalten wir die Beziehung zwischen der Funktion F(p) und der Anzahl unterschiedlicher Elemente, deren Häufigkeiten in das Intervall $[p',p''] \subset [p_n,p_1]$ fallen:

$$\int_{p'}^{p''} dv(p) = \int_{p'}^{p''} \frac{dF(p)}{p} = v(p'') - v(p'). \qquad (11)$$

Unter der Voraussetzung, daß F(p) in $[p_v,p_1]$ differenzierbar ist, führen wir eine Funktion f(p) so ein, daß

$$f(p)\,dp = dF(p). \qquad (12)$$

Dann ist

$$\int_{p'}^{p''} f(p)\,dp = \int_{p'}^{p''} dF(p) = F(p'') - F(p') \approx F^N(p'') - F^N(p') =$$

$$= \sum_{p'<\frac{m}{N}\le p''} v_m(N)\frac{m}{N} = P(p'<p\le p''), \qquad (13)$$

d.h. das Integral auf der linken Seite ist gleich der Wahrscheinlichkeit der Wahl eines Elements aus unserer Stichprobe, dessen Häufigkeit zwischen p' und p'' liegt. Diese Gleichung kann man als eine direkte Definition von f(p) ansehen; offensichtlich wird sie relativ umso genauer , je größer das Intervall [p',p''] ist. Die Funktion f(p) bezeichnen wir - in Analogie zu der üblichen Dichte - als die Funktion der strukturellen Dichte der Menge der Häufigkeiten $\{p_i^N\}$.

Da $f(p) \neq 0$ nur im Intervall $[p_v,p_1]$ gilt und außerhalb dieses Intervalls f(p) identisch Null ist, kann man aufgrund von (13) die übliche Bedingung der Normierung als

$$\int_{p_v}^{p_1} f(p)dp = 1 \qquad (14)$$

schreiben.

Unter Verwendung der Funktion der strukturellen Dichte kann man (8) zu

$$H = - \int_{p_v}^{p_1} f(p)\ln(p)dp, \qquad (15)$$

und (10) zu

$$dv(p) = \frac{f(p)}{p}\,dp \qquad (16)$$

umformen. Daraus erhält man leicht die Zahl unterschiedlicher Elemente $v[p',p'']$, deren Häufigkeiten zwischen p' und p'' liegen:

$$v[p',p''] - 1 = v^N(p'') - v^N(p') = \int_{p'}^{p''} \frac{f(p)}{p}\,dp. \qquad (17)$$

Insbesondere für die Anzahl unterschiedlicher Elemente in der Stichprobe (d.h., für das "Vokabular") erhalten wir

$$v = v[p_n,p_1] \approx \int_{p_v}^{p_1} \frac{f(p)}{p}\,dp + 1. \qquad (18)$$

Mit Hilfe der Funktion der strukturellen Dichte kann man auch (7) als

$$\int_{p_i}^{p_1} \frac{f(p)}{p}\,dp = i - 1; \quad i = 1,2,\ldots,v \qquad (19)$$

schreiben.

Kennt man also eine der drei Funktionen f(p), F(p) oder v(p), so kennt man auch die Häufigkeitsstruktur der Stichprobe. Obwohl man nicht direkt weiß, zu welchem Element welche Häufigkeit gehört (die Analyse der Häufigkeitsstruktur hat nicht die Intention, dieses Problem zu lösen), kann man trotzdem bei zufälliger Überprüfung eines beliebigen Elements, dessen Häufigkeit in dem gegebenen Intervall liegt, sowohl die Wahrscheinlichkeit der Stichprobe [Formel (13)] als auch die Anzahl unterschiedlicher Elemente, deren Häufigkeiten in dem gegebenen Intervall liegen [Formel (14)], abschätzen. Ebenso kann man die Entropie [Formel (8) und (15)] schätzen. Die Formeln (7) und (9) geben die Möglichkeit, zu der traditionellen Vorstellung der Häufigkeitsstruktur als einer nach abnehmenden Häufigkeiten geordneten Menge überzugehen.

Diese Darstellungsart der Häufigkeitsstruktur kann man offensichtlich immer dann anwenden, wenn die Zahl unterschiedlicher Elemente groß ist und die Bedingung (2) erfüllt ist. Im allgemeinen ist sie dann zweckmäßig, wenn die Funktionen f(p), F(p) oder v(p) als elementare Funktionen leicht integrierbar sind.

2

Wir zeigen, daß eine Funktion der strukturellen Dichte des Typs

$$f(p) = \frac{A}{p^{\alpha}} \qquad (A,\alpha \text{ sind Konstanten}) \qquad (20)$$

dieselbe Häufigkeitsstruktur wie die Mandelbrotsche Formel (1) beschreibt.

Setzt man (20) in (19) ein und löst bezüglich p_i, so erhält man

$$p_i = \frac{(A/\alpha)^{1/\alpha}}{(A/\alpha p_1^{\alpha} - 1+i)^{1/\alpha}} . \qquad (21)$$

Vergleicht man (21) mit (1), so sieht man, daß beide Formeln identisch sind, wenn man

$$K = \left(\frac{A}{\alpha}\right)^{1/\alpha} ; \quad B = \frac{A}{\alpha p_1^{\alpha}} - 1; \quad \gamma = \frac{1}{\alpha} \qquad (22)$$

setzt.

Die Konstante A wird aufgrund der Normierung (14) bestimmt:

$$A_{\alpha} = \frac{1-\alpha}{p_1^{1-\alpha} - p_v^{1-\alpha}} \qquad \text{für } \alpha \neq 1 \qquad (23a)$$

$$A_1 = \frac{1}{\ln(p_1/p_v)} \qquad \text{für } \alpha = 1 \qquad (23b)$$

Damit haben wir nicht nur gezeigt, daß (1) und (20) im Grunde dieselbe Häufigkeitsstruktur beschreiben, sondern wir haben auch die Möglichkeit erhalten, die Konstanten K und B in der Mandelbrotschen Formel als Funktionen der unmittelbar beobachteten Parameter p_1 und p_v der Stichprobe zu schätzen. Die Anzahl unterschiedlicher Elemente in der Stichprobe, berechnet nach (18), erhält man auch als Funktionen eben dieser Parameter:

$$v = \frac{A}{\alpha}\left(p_v^{-\alpha} - p_1^{-\alpha}\right) + 1 . \qquad (24)$$

Der einzige unabhängige Parameter ist also die Größe $\alpha = 1/\gamma$.

Es ist zu bemerken, daß Formel (21), die man aus (20) erhält, nur ein Spezialfall der breiteren Klasse von Häufigkeitsstrukturen ist, die durch (20) beschrieben werden. Obwohl bei $\alpha = 0$ die erhaltenen Beziehungen ihren Sinn verlieren, ist das speziell bei (20) nicht der Fall, denn setzt man sie in (19) ein und löst nach p_i, so findet man die Approximation für die Häufigkeitsfolge als

$$p_i = p_1 e^{-(i-1)(p_1-p_v)} = p_1 e^{p_1-p_v} e^{-i(p_1-p_v)} . \qquad (25)$$

In der quantitativen Linguistik waren solche Häufigkeitsfolgen bisher nicht bekannt[1]. Es hat sich aber gezeigt, daß nach der Formel (25) die Häufigkeiten elementarer Einheiten wie Buchstaben in geschriebenen Texten oder melodische Intervalle in der Musik abnehmen. Als Beispiel hierfür werden in Abb. 1 die Häufigkeiten russischer Buchstaben und in Abb. 2 die Häufigkeiten melodischer Intervalle in Chopins Nocturnen gezeigt. Beide Graphen sind im halblogarithmischen Maßstab dargestellt, wodurch eine exponentielle Funktion in eine Gerade transformiert wird.

Analoge Graphen erhält man für das Englische (mehrere Stichproben), Französische, Althebräische, Rumänische, Estnische, Ungarische und Deutsche. In allen Fällen zeigte sich bei den seltensten Buchstaben eine starke Abweichung nach unten von der Kurve (25). Eine allgemeine Häufigkeitsgrenze, unterhalb welcher Abweichungen vorkamen, kann als 0.01 ± 0.005 angesetzt werden. Einige gemeinsame Abweichungen von (25) zeigten sich beim Georgischen, Armenischen, Suaheli und Hausa. Ein charakteristisches Beispiel für die Häufigkeit georgischer Buchstaben ist auf Abb. 3 dargestellt. Auch in diesen Fällen nimmt die Häufigkeitsmasse

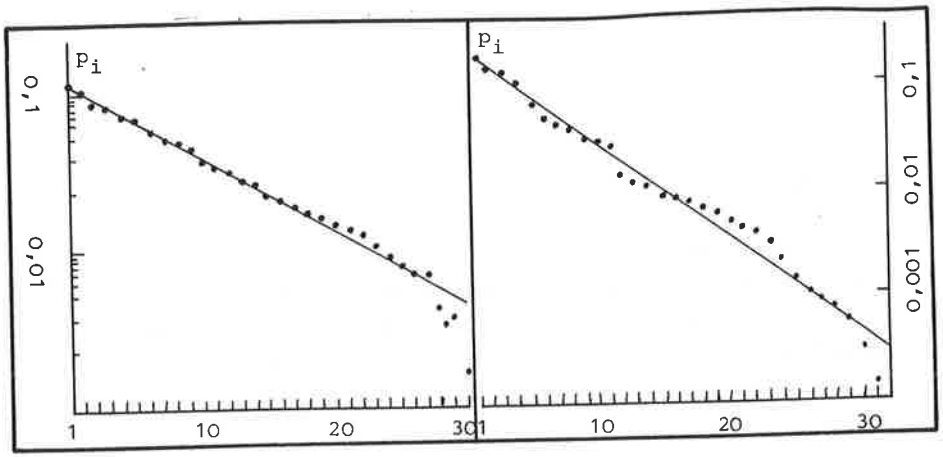


Abb. 1. Buchstabenhäufigkeiten in russischen Texten (nach Lebedev und Garmaš; Zwischenraum ausgelassen). Die unterbrochene theoretische Gerade wurde laut (25) berechnet, wobei p_1 und p_V aus den Daten eingesetzt wurden

Abb. 2. Häufigkeiten melodischer Intervalle in Nokturnen von F. Chopin (eigene Zählungen zusammen mit M.G. Boroda)

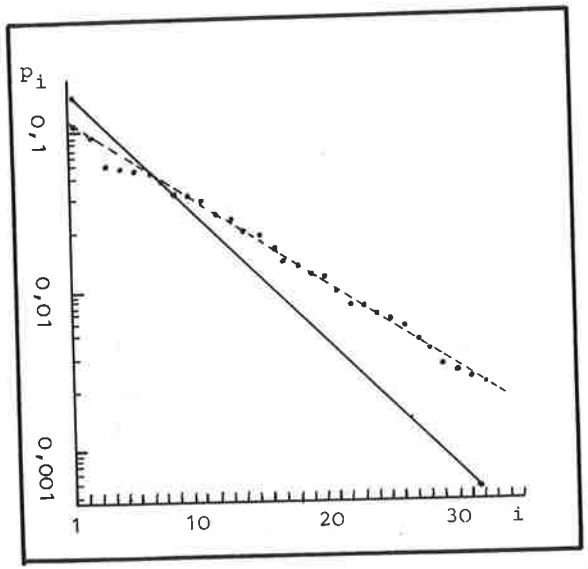


Abb. 3. Buchstabenhäufigkeiten in georgischen Texten (nach Gačečiladze & Eliasvili 1958)

exponentiell ab, jedoch unterscheiden sich die Parameter von denen in (25) (punktierte Linie in Abb. 3). Der Unterschied im Verlauf der Kurven entsteht in diesen Sprachen durch die überhöhte Frequenz der häufigsten Buchstaben im Vergleich mit dem allgemeinen Abnahmetrend der gesamten Buchstabenmasse.

Die Formel (20) verallgemeinert also die Häufigkeitsstruktur des Typs (1) auch für den Fall, wenn $\alpha \leq 0$, ähnlich wie die Mandelbrotsche Formel die Beobachtungen von Estoup, Condon und Zipf für den Fall der im Bereich großer Frequenzen gekrümmten doppellogarithmischen Häufigkeitskurven (mit Hilfe der Konstante B) verallgemeinert. Von den Stichproben, deren Häufigkeitsstruktur durch (20) beschrieben wird (oder durch eine aus (20) abgeleitete Reihe) werden wir sagen, daß sie dem verallgemeinerten Zipf-Mandelbrotschen Gesetz, unabhängig von dem Parameter α, folgen.

3

Von den allgemeinen Beziehungen gehen wir jetzt zu der Untersuchung eines spezifischen Falles über, der z.B. für lexikalische Stichproben charakteristisch ist. Eine wichtige Eigenschaft dieser Stichproben ist ein **obligatorisch** großer Anteil von hapax legomena und anderer umfangreicher Klassen von seltenen Wörtern. Dieser Umstand wurde zum ersten Mal von Yule (1944) bemerkt; auch später wurden keine Stichproben ohne hapax legomena beobachtet. Mit anderen Worten, die Stichproben schöpfen ihr "potentielles Vokabular" nicht aus, und die Wachstumskurve des Vokabulars erreicht den Sättigungsbereich nicht.

Das obligatorische Vorkommen einer großen Klasse von hapax legomena erlaubt zu schreiben

$$p_v = \frac{1}{N}. \tag{26}$$

Aufgrund dieser Gleichheit kann man die Konstanten A,K,B sowie den Wortschatz auf den Umfang derjenigen Stichprobe, deren Häufigkeitsstruktur dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgt, in Bezug setzen. Im weiteren werden wir den Umfang einer solchen Stichprobe, die laut Definition diesem Gesetz folgt, mit Z bezeichnen und den "Zipfschen Umfang" nennen.

Setzt man (26) in (23) ein, so erhält man

$$A_\alpha = \frac{1-\alpha}{p_1^{1-\alpha}-Z^{\alpha-1}}; \quad A_1 = \frac{1}{\ln(Zp_1)}. \tag{27}$$

Weitere Einsetzungen ergeben

$$K_\alpha = \left[\frac{1-\alpha}{\alpha(p_1^{1-\alpha}-Z^{\alpha-1})}\right]^{1/\alpha}; \quad K_1 = \frac{1}{\ln(Zp_1)}; \tag{28}$$

$$B_\alpha = \frac{1}{\alpha p_1^{\alpha}(p_1^{1-\alpha}-Z^{\alpha-1})} - 1; \quad B_1 = \frac{1}{p_1\ln(Zp_1)} - 1. \tag{29}$$

Der Ausdruck (21) wird in dem Fall, daß $\alpha = 1$ ist, zu

$$p_i = \frac{\frac{1}{\ln(Zp_1)}}{\frac{1}{p_1\ln(Zp_1)} - 1 + i}. \tag{30}$$

Aus (24) erhalten wir die Zahl unterschiedlicher Elemente (das "Vokabular") als die Funktion des Zipfschen Umfangs:

$$v^{(\alpha)}(Z) = \frac{A}{\alpha}(Z^{\alpha}-p_1^{-\alpha}) + 1. \tag{31a}$$

Im speziellen Fall, wenn $\alpha = 1$ ist, und die Größe

$$\frac{1}{p_1\ln(Zp_1)} - 1 = B$$

vernachlässigt wird, kann man schreiben

$$v^{(1)}(Z) \approx \frac{Z}{\ln(Zp_i)} \tag{31b}$$

Da sich im untersuchten Fall selten vorkommende Elemente in umfangreiche Klassen mit der Häufigkeit $m = 1,2,\ldots$ gruppieren, bestimmen wir aus (17) die Anzahl unterschiedlicher Elemente, die jeweils die absolute Häufigkeit m haben, als die Anzahl der Elemente, deren Häufigkeit im halboffenen Intervall der relativen Häufigkeiten $[\frac{m}{Z}, \frac{m+1}{Z})$ liegt, als

$$v_m(Z) = v[\tfrac{m}{Z}, \tfrac{m+1}{Z}) = A\int_{\frac{m}{Z}}^{\frac{m+1}{Z}} p^{\alpha+1}\,dn. \tag{32}$$

Für $\alpha \neq 1$ ergibt sich daraus

$$v_m^{(\alpha)}(Z) = v^{(\alpha)}(Z)\left(\frac{1}{m^{\alpha}} - \frac{1}{(m+1)^{\alpha}}\right). \tag{33a}$$

Für $\alpha = 1$ ergibt sich

$$v_m(Z) = \frac{v(Z)}{m(m+1)}. \tag{33b}$$

Obwohl die Formeln (33a,b) in der Literatur noch nicht vorkommen, müssen wir bemerken, daß die allgemeine Abhängigkeit der Abnahme des Klassenumfangs von der Häufigkeit der in ihr enthaltenen Elemente dieselbe ist wie in den diesbezüglichen Ausdrücken in den Arbeiten von Yule (1944), Kalinin (1965) und Booth (1967). Formel (33b) ist ein Spezialfall der Formel von Simon (1960). Auf diese Weise widersprechen die erhaltenen Beziehungen nicht den bekannten Verallgemeinerungen.

4

Bis jetzt war unsere Darstellung rein formal. Es wurden keine Voraussetzungen über statistische oder kombinatorische Mechanismen verwendet, durch die die analysierten Strukturen erzeugt werden. Obwohl wir auch im weiteren auf Hypothesen über die Struktur der Verteilung der "wahren Wahrscheinlichkeiten" verzichten werden (die Gründe hierfür werden am Schluß diskutiert), führen wir ein einfaches kombinatorisches Modell ein, das uns ermöglicht, die Dynamik der Veränderung der Häufigkeitsstruktur und die Zunahme der Anzahl unterschiedlicher Elemente in Stichproben des lexikalischen Typs, für die (26) gilt, zu beschreiben.

Es soll ein unendlich großer statistisch homogener Text gegeben sein. Unter einem homogenen Text verstehen wir laut Kalinin (1965) einen solchen Text, in dem die Wahrscheinlichkeit der Verwendung jedes Wortes von seiner Position und von den früher verwendeten Wörtern unabhängig ist (d.h. wir setzen nur die Existenz der Wahrscheinlichkeiten voraus, ohne jegliche Beschränkungen über ihre Art; dies hängt damit zusammen, daß unterschiedliche Verteilungen zu identischen mathematischen Erwartungen der selten vorkommenden Wörter führen). Genaugenommen wird diese Voraussetzung in realen Texten nicht erfüllt, aber sie erlaubt, Zusammenhänge, die mit der Realität gut übereinstimmen, abzuleiten. Unser Modelltext soll noch zwei weitere Bedingungen erfüllen:

1. In einer Stichprobe beliebiger Länge aus diesem Text soll es hapax legomena geben, d.h. im ganzen Verlauf des Textes gilt die Beziehung (26); mit anderen Worten, das Vokabular des Textes ist unendlich.

2. In Stichproben mit festem Umfang Z aus diesem Text soll das verallgemeinerte Zipf-Mandelbrotsche Gesetz gelten; d.h. die Wahrscheinlichkeit häufiger Wörter folgt der Beziehung (1) mit Koeffizienten, die durch (28)-(29) gegeben sind, und die Ausdrücke (33) betrachten wir als mathematische Erwartungen der Zahl der m-maligen Wörter in solchen Stichproben. So wird anstatt der Vorgabe einer nichtbeobachteten Verteilung im Bereich

selten vorkommender Wörter die beobachtete Struktur in einem fixierten Stichprobenumfang postuliert.

Untersuchen wir, wie die Häufigkeitsstruktur in Stichproben mit einem anderen, von Z unterschiedlichen Umfang N gestaltet wird. Zwecks Anschaulichkeit führen wir zuerst eine qualitative Analyse durch.

Wenn sich der Umfang einer Stichprobe ändert, so bleiben die Häufigkeiten häufiger Wörter unverändert (bis auf eine rein statistische Streuung, die man mit üblichen Methoden, z.B. mit dem t-Test beurteilen kann). Jedoch wird das letzte Element in der empirischen Menge der Häufigkeiten diesmal nicht 1/Z sondern 1/N. Das heißt, wenn $N < Z$, dann hat diese Menge weniger Elemente als die Menge mit Zipfschem Umfang, und wenn $N > Z$, dann ist es umgekehrt. Offensichtlich gilt also: wenn alle Häufigkeiten nach wie vor auf der Kurve (1) liegen, dann wird die Bedingung der Normiertheit verletzt.

Nehmen wir an, daß wir die ursprüngliche Stichprobe verringert haben, so daß der neue Umfang $N < Z$ ist. Da sich die Häufigkeiten häufiger Wörter nicht geändert haben, so müssen, zwecks Normierung, wegen des höheren Wertes des letzten Elements der Häufigkeitsfolge (und offensichtlich auch wegen der kleineren Zahl der Elemente), die Häufigkeiten seltener Wörter größer sein als derjenigen mit derselben Rangzahl im Zipfschen Umfang. Vergrößert man den Umfang der Stichprobe, so erhält man auf Grund derselben Überlegung das umgekehrte Bild: Der "Schweif" der Kurve im Bereich seltener Wörter muß unterhalb der ursprünglichen Zipfschen Kurve liegen, d.h. die Häufigkeitsstruktur unseres Modelltextes kann nicht stabil sein; sie scheint eine Funktion des Stichprobenumfangs zu sein.

Diese qualitative Überlegung kann durch eine kombinatorische Berechnung unterstützt werden.

Kalinin (1965) hat das Problem der Deformierung der Häufigkeitsstruktur in homogenen Stichproben gelöst. Wenn die mathematischen Erwartungen des Vokabulars $v(N_o)$ und der Zahl der m-maligen Wörter $v_m(N_o)$ in einer "Basisstichprobe" mit Umfang N_o aus demselben Text bekannt sind, dann sind in Stichproben

eines beliebigen Umfangs N das Vokabular $v(N,N_o)$ und die Zahl der m-maligen Wörter $v_m(N,N_o)$ identisch:

$$v(N,N_o) = v(N_o) - \sum_{j=1}^{\infty} \left(1 - \frac{N}{N_o}\right)^j v_j(N_o); \tag{34}$$

$$v_m(N,N_o) = (-1)^{m+1} \frac{N^m}{m!} \frac{d^m}{dN^m} v(N,N_o) \tag{35}$$

Setzt man die Häufigkeitsstruktur mit Zipfschem Umfang (31) und (33) als Charakteristikum der Basisstichproben in (34) und (35) ein, so erhält man

$$v(N,Z) = v(Z)\frac{1-x}{x} \lambda(x,\alpha); \tag{36}$$

$$v_m(N,Z) = v(Z)\frac{(-1)^{m+1}}{m!} \left(\frac{N}{Z}\right)^m \frac{d^m}{dx^m} \left(\frac{1-x}{x} \lambda(x,\alpha)\right) \tag{37}$$

wo $x = 1 - \frac{N}{Z}$

und $\lambda(x,\alpha) = \sum_{j=1}^{\infty} \frac{x^j}{j^\alpha}$.

Die Reihe $\lambda(x,\alpha)$ hat den Konvergenzradius |1|; ihre analytische Fortsetzung im Intervall $-\infty<x<1$ ist

$$\lambda(x,\alpha) = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{t^{\alpha-1} x e^{-t}}{1-xe^{-t}} dt, \tag{38}$$

wo $\Gamma(\alpha)$ die Gamma-Funktion mit dem Argument α ist.

Wenn $\alpha = 1$, so kann man $v(N,Z)$ und $v_m(N,Z)$ mit Hilfe elementarer Funktionen ausdrücken:

$$v(N,Z) = v(Z) \frac{\ln \frac{N}{Z}}{\frac{Z}{N}-1}; \tag{39}$$

$$v_m(N,Z) = v(Z) \frac{(-1)^{m+1}}{m!} \left(\frac{N}{Z}\right)^m \frac{d^m}{dx^m} \left(\frac{x-1}{x} \ln(1-x)\right). \tag{40}$$

Bei kleinem m bekommt man speziell:
für hapax legomena:

$$v_1(N,Z) = v(Z) \frac{M-1}{X-1}; \tag{41}$$

für zweimal vorkommende Wörter:

$$v_2(N,Z) = v(Z) \frac{X+1-2M}{2(X-1)^2}; \tag{42}$$

für dreimal vorkommende Wörter:

$$v_3(N,Z) = v(Z) \frac{6M+X^2-5X-2}{6(X-1)^3}, \tag{43}$$

wo $X = \frac{Z}{N}$

und $M = \frac{\ln X}{1-\frac{1}{X}}$ ist.

Wenn $X \approx 1$, dann kann man den Ausdruck

$$v_m(N,Z) = v(Z) \sum_{i=m}^{\infty} \frac{(1-X)^{i-m}}{i(i+1)} \tag{44}$$

verwenden.

Formel (40) wird bei großem m sehr umfangreich und schwer auswertbar. Daher kann man den Verlauf des Bereichs mit kleinen Häufigkeiten auf folgende Weise leichter beschreiben.

Ist der Wert der absoluten Häufigkeit m gegeben, dann kann man die letzte Rangzahl in der allgemeinen Häufigkeitsliste für die Wörter mit der Häufigkeit m + 1 [die rechte Abszisse des Rechtecks für die Wörter mit der Häufigkeit m + 1] folgendermaßen berechnen

$$i_{m+1} = v(N,Z) - \sum_{j=1}^{m} v_j(N,Z) = v(Z) \sum_{j=1}^{\infty} \frac{(1-X)^{j-1}}{j+m}. \qquad (45a)$$

Diesen Ausdruck kann man für Werte von $x \approx 1$ benutzen. Im allgemeinen Fall soll man den Ausdruck

$$i_{m+1} = \frac{v(Z)}{(1-X)^m} \left[\frac{\ln X}{X-1} - \sum_{k=1}^{m} \frac{(1-X)^{k-1}}{k} \right] \qquad (45b)$$

benutzen, woraus folgt

$$v_m(N,Z) = i_m - i_{m+1}.$$

Die linke Abszisse des Rechtecks für Wörter mit Häufigkeit m ist dann natürlich $i_{m+1}+1$. Den Ausgangspunkt für die Konstruktion der theoretischen Graphen bildet das vollständige Vokabular des Textes $v(N,Z)$, das die von hapax legomena gebildete Abszisse des untersten Rechtecks bestimmt. Wenn man den Graphen in relativen Häufigkeiten zeichnet, dann ist die Ordinate dieses Punktes gleich 1/N.

Auf Abb. 4 sieht man die Veränderung der Häufigkeitsstruktur des Textes bei der Veränderung des Stichprobenumfangs; hierbei wurde $\alpha = 1$ gesetzt.

Beim Umfang Z wird die Häufigkeitsfolge durch (21) beschrieben. Die ununterbrochene, fallende Kurve ist die Häufigkeitsstruktur beim Umfang Z = 27000 und $p_1 = 0.07$. Die "Treppen" im unteren Teil des Graphen sind laut (45a und b) für verschiedene Werte des Umfangs N berechnet worden. Bei wachsendem m nähern sich die Treppen der ursprünglichen Kurve, die für den Umfang Z

laut (30) berechnet wurde. Das vollständige Vokabular bei diesem Umfang ist gleich $v(Z)$. Die kleine Krümmung im oberen Teil des Graphen wird durch den Umstand verursacht, daß $B > 0$. Je mehr sich der Stichprobenumfang N von dem Zipfschen Umfang Z unterscheidet, desto stärker ist der "Schweif" der Kurve gekrümmt.

Beim Umfang $N_1 < Z$ liegt der größte Teil der "Rechtecke" über der ursprünglichen Kurve (21). Wenn man die Größe γ direkt für diesen Graphen bestimmen wollte, so würde sie im unteren Teil des Graphen etwas kleiner als 1 ausfallen. Die größere Länge der niedrigsten Rechtecke (m = 1,2) zeugt von der Vergrößerung des Anteils seltener Wörter im Vokabular (es ist zu bemerken, daß bei $N < Z$ der Anteil der hapax legomena über die Hälfte des Wortschatzes beträgt).

Vergrößert man die Stichprobe (Umfang N_2 ist unwesentlich kleiner als Z), so verlängert sich die Folge von Häufigkeiten (d.h. das Vokabular wächst) und die rechten Ecken der Rechtecke nähern sich der Kurve (21). Wenn N = Z, so wird (21) zur oberen Grenze dieser Rechtecke; damit der Graph nicht unübersichtlich wird, haben wir diese Situation nicht gezeichnet, man kann sie sich aber leicht durch $v(N,Z) \to v(Z)$ bei $N \to Z$ vorstellen. Die Anzahl der hapax legomena nähert sich dabei der Hälfte des Vokabulars.

Bei $N_3 > Z$ beginnt der Schweif unter der Kurve (21) zu liegen. Der Anteil seltener Wörter im Vokabular nimmt ab (obwohl ihre absolute Zahl wächst); die Anzahl der hapax legomena wird kleiner als die Hälfte des Vokabulars. Bei der empirischen Schätzung von γ einer derartigen Kurve ist es nicht leicht zu bestimmen, wo ihr geradliniger Teil endet; die dargestellte Krümmung würde zu dem Schluß führen, daß das Zipfsche Gesetz in dem gegebenen Text schlecht erfüllt ist und daß γ im unteren Teil des Graphen etwas größer als 1 ist.

Schon bei einem flüchtigen Blick auf Abb. 4 sieht man also, daß das vorgeschlagene Modell Situationen wiedergibt, die man in empirischen Graphen üblicherweise vorfindet. Aus dieser Analyse folgt, daß die "Abweichungen im Schweif" eine gesetzmäßige und unvermeidliche Erscheinung sind in dem Fall, wo sich der Stichprobenumfang vom Zipfschen Umfang Z unterscheidet. D.h.,

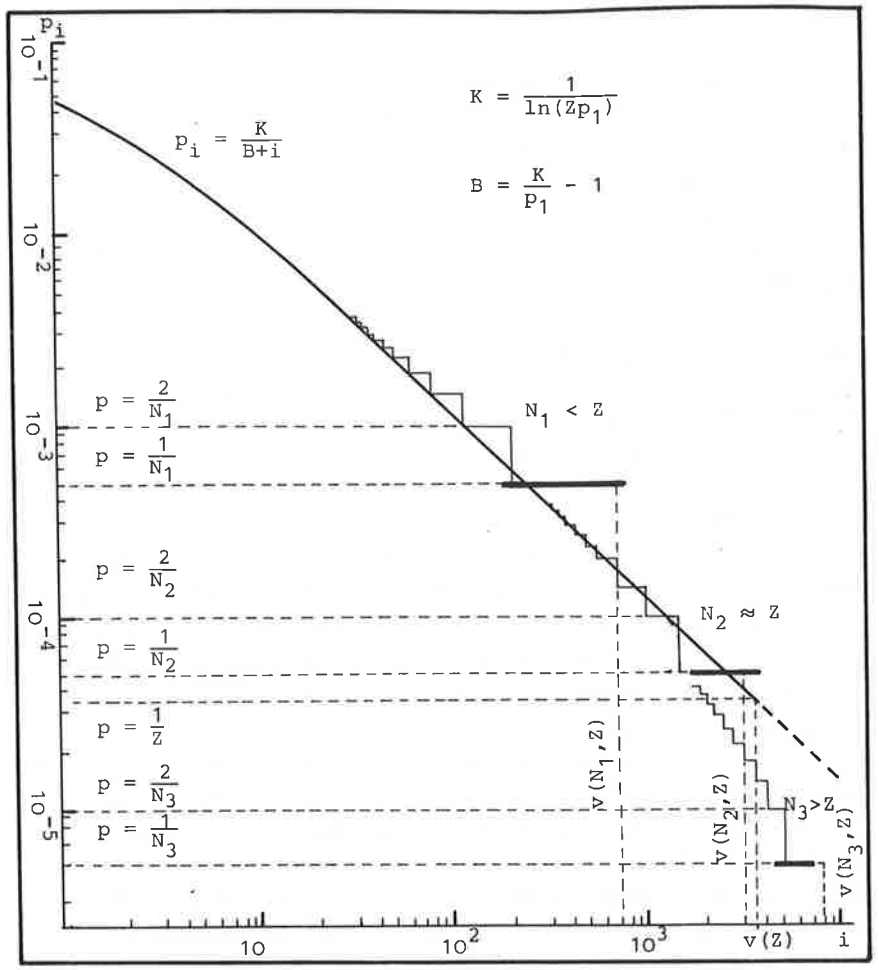


Abb. 4. Das allgemeine Bild der Dynamik der Häufigkeitsstruktur

ein beliebiger homogener Text kann dem Zipf-Mandelbrotschen Gesetz in seiner "kanonischen" Form (1) nur bei einem einzigen Wert des Stichprobenumfangs folgen (wenn er ihm überhaupt folgen kann). Bei allen anderen Umfängen entstehen gesetzmäßige Krümmungen der Treppenfunktion, ungeachtet dessen, daß die gegebenen Werte von α oder γ gleich bleiben. Die Kenntnis des Z-Wertes erlaubt es, sowohl das Wachsen des Vokabulars innerhalb des Textes [Formeln (37) und (39)] als auch seine Häufigkeitsstruktur beim beliebigen Umfang [Formeln (40-45) und (37)] zu beschreiben.

Die durch (36) - (45) beschriebenen Häufigkeitsstrukturen kann man als eine natürliche Verallgemeinerung des Zipf-Mandelbrotschen Gesetzes auf den Fall, wo der Stichprobenumfang N beliebig und ungleich Z ist (d.h. ungleich dem einzigen Umfang, bei dem die Beziehungen (27) - (33) "konstruktionsgemäß" erfüllt sind) betrachten. Diese Häufigkeitsstrukturen kann man als "Quasizipfsche" bezeichnen, da sie die Folge des Umstandes sind, daß beim Umfang Z das Zipf-Mandelbrotsche Gesetz erfüllt wird.

Klären wir die besondere Rolle des Wertes Z, die mit dem allgemeinen Verlauf des Vokabularwachstums innerhalb des Textes zusammenhängt. Betrachten wir zwei Texte mit identischen p_1 und α, aber unterschiedlichen Zipfschen Umfängen $Z_1 > Z_2$. Erhebt man aus beiden Texten gleiche Stichproben mit dem Umfang N, so sieht man, daß

$$v(N,Z_1) > v(N,Z_2). \qquad (46)$$

Das heißt, ein Vokabular wächst beim Anwachsen des Stichprobenumfangs umso schneller, je größer die Größe Z . Das bedeutet, daß die relative Sättigung des Textvokabulars nicht nur von $\alpha = 1/\gamma$, sondern auch vom Wert des Zipfschen Umfangs abhängt. Da man im Rahmen einer Sprache p_1 und α als Konstanten betrachten kann, kann man Z als ein akzeptables Maß des relativen Vokabularreichtums betrachten.

5

Als Ergebnis der Untersuchung erhielten wir statt der Konstanten K,B und γ in der Mandelbrotschen Formel die Parameter p_1, p_v und α oder (im Fall kleiner lexikalischer Stichproben, die hapax legomena enthalten) p_1, Z und α. Die Parameter p_1 und p_v kann man direkt aus den Daten der Stichprobe abschätzen; α = = 1/γ; der Zipfsche Umfang Z scheint eine neue Größe zu sein, die in bisherigen linguostatistischen Untersuchungen nicht vorhanden war. Zipf (1949) hat sie zwar unter dem Namen "optimaler Umfang" in seiner Antwort auf die Kritik von M. Joos (1936) eingeführt, (Joos behauptete, daß die von Zipf untersuchte Gesetzmäßigkeit kein universales Gesetz sein kann), aber sonst wurde diese Idee nicht weiter entwickelt. Als Zipf versuchte, den Wert dieses Umfangs empirisch zu bestimmen, hat man ihn sogar beschuldigt, daß er "das Experiment der Theorie anpaßt" (vgl. Frumkina 1961).

Der Begriff des Zipfschen Umfangs erlaubt es, bei kleinen Stichproben den Mechanismus der "Abweichung im Schweif" von "kanonischen" Kurven des Typs (1) im Bereich der kleinen Häufigkeiten zu verstehen und die Größe dieser Abweichung abzuschätzen. Dies ist das wesentliche theoretische Resultat der vorliegenden Arbeit.

Der Vergleich mit empirischem Material zeigt, daß man lexikalische Stichproben allgemein mit dem vorgelegten Modell bei α = 1 unabhängig von Stil, Genre und Sprache beschreiben kann. Dabei kommt der Umfang Z dem Umfang eines größeren literarischen Werkes sehr nahe (vgl. Orlov 1969a,b 1970), denn sowohl das Vokabular solcher Texte als auch ihre Häufigkeitsstruktur lassen sich mit Hilfe von (30), (31b) und (33b) zufriedenstellend (in der Regel mit einer Fehlergenauigkeit von ± 20%) beschreiben, wenn man für Z den reellen Textumfang einsetzt. In Orlov (1969b) wurden ergänzend die Veränderungen des Vokabulars und der Häufigkeitsstruktur innerhalb des Textes aufgrund der aus ihm erhobenen Stichprobe analysiert. Die Veränderungen entsprechen den Ausdrücken (39) und (41-43) mit derselben Genauigkeit.

In Nadarejšvili & Orlov (1971) wurde der Unterschied der Zuwachsrate des Vokabulars in Texten unterschiedlicher Länge, die von demselben Autor stammen, gezeigt; dort wurde auch gezeigt, daß sehr große Texte in ihren vom Autor bestimmten Teilen (Kapitel, Band usw.) dem Zipf-Mandelbrotschen Gesetz folgen können.

Gleichzeitig gilt aber, daß willkürliche Stichproben (Ausschnitte und Zusammenstellungen vieler Texte in eine Stichprobe, die die "Sprache als ganze" oder eine "Fachsprache" repräsentieren sollen) in der Regel dem kanonischen Zipf-Mandelbrotschen Gesetz (1) nicht folgen. Die Abweichungen des Vokabulars vom Umfang (31b) erreichen 50 - 100%, der Verlauf der Häufigkeitskurve stimmt mit der theoretischen Kurve (30) nicht überein, die Krümmung der Treppenkurve im Schweif des Graphen ist erheblich (vgl. Orlov 1974). Ausnahmen, d.h. Fälle, wo die kanonische Form genau erfüllt wird, sind selten und können durch die zufällige Nähe des Stichprobenumfangs zu dem Zipfschen Umfang erklärt werden. Jedoch erlaubt das entwickelte Modell praktisch in allen Fällen, die Parameter solcher Stichproben miteinander in Zusammenhang zu bringen (vgl. Orlov 1976).

Als Beispiel bringen wir in der Tabelle 1 die Resultate der Berechnung des Anwachsens des Vokabulars und der Zahl der hapax legomena mit Hilfe des dargestellten Modells für die Daten von Guiraud (1960). In der Tabelle werden auch die von Kalinin (1965) berechneten theoretischen Werte angegeben. Kalinin geht bei den Berechnungen von der Größe des Vokabulars in einem beliebigen Punkt N_o und von den in relativ kleinen Abständen genommenen sukzessiven Anzahlen der hapax legomena aus (deswegen konnte Kalinin die theoretische Prognose für die Umfänge N = 100000 und N = 200000 nicht berechnen). Bei unseren Berechnungen benutzten wir den Vokabularumfang v = 3040 bei einem Stichprobenumfang N = 200000. Diese Werte wurden in die Gleichung

$$v(20000,Z) = 3040$$

eingesetzt. Die approximative Lösung dieser Gleichung[2)] bezüg-

Tabelle 1. Vergleich der Prognosen

N	Daten von Guiraud: v	Daten von Guiraud: v_1	Prognose von Kalinin: Berechnet bei	Prognose von Kalinin: Theoretisches Vokabular v(N)	Prognose von Kalinin: $\frac{N}{v(N)}$	Prognose laut (39) und (42); Z=27000, p_1=0,07: v(N,Z)	Prognose laut (39) und (42); Z=27000, p_1=0,07: $\frac{v}{v(N,Z)}$	Prognose laut (39) und (42); Z=27000, p_1=0,07: $v_1(N,Z)$	Prognose laut (39) und (42); Z=27000, p_1=0,07: $\frac{v_1}{v_1(N,Z)}$
2000	800	600	N_o = 2000	750	1,07	746	1,07	518	1,16
4000	1250	800	N_o = 4000	1300	0,96	1190	1,05	815	0,98
6000	1580	950	" "	1660	0,95	1540	1,03	950	1,00
8000	1900	1070	" "	1950	0,97	1830	1,04	1100	0,98
10000	2120	1160	" "	2200	0,96	2090	1,01	1215	0,95
12000	2320	1240	" "	2420	0,96	2320	1,00	1320	0,94
14000	2505	1320	" "	2610	0,96	2530	0,99	1407	0,94
16000	2710	1390	" "	2800	0,97	2720	1,00	1490	0,93
18000	2775	1450	" "	2960	0,94	2900	0,96	1547	0,96
20000	3040	1500	N_o = 10000	3060	0,99	3070	-	1637	0,92
30000	3650	1710	" "	3730	0,98	3740	0,98	1790	0,96
40000	4200	1880	" "	4270	0,98	4320	0,97	2000	0,94
50000	5000	2100	" "	4700	1,06	4800	1,04	2145	0,98
100000	6000	2350	-	-	-	6240	0,96	2475	0,95
200000	7500	2750	-	-	-	8270	0,91	2850	0,96

lich Z ergab Z ≈ 27000 (d.h. wir fanden diesen Zipfschen Umfang für den Modelltext, der bei einem Stichprobenumfang von N = 20000 den tatsächlich beobachteten Vokabularumfang ergibt; mit anderen Worten: betrachtet man (39) als die Gleichung für die Kurve des Vokabularwachstums, die von dem Parameter Z abhängt, dann setzt man diese Kurve mit einem experimentell beobachteten Punkt gleich). Als p_1 wurde 0.07 angenommen, was für Sprachen mit Artikeln typisch ist (bei einer schwachen Abhängigkeit der benutzten Beziehungen von p_1 ist es möglich, auch "aus der Luft gegriffene" Werte zu benutzen). Aus diesen zwei Zahlen konnte man sowohl den Vokabularumfang als auch die Zahl der hapax legomena in allen Punkten berechnen. Abweichungen von über 5% im Vokabularumfang wurden nur an den äußersten Enden des untersuchten Materials bei zehnfachen Prädiktionen und Retrodiktionen von dem Ausgangspunkt beobachtet (vgl. Spalten 8 und 10, Zeilen für N = 2000 und N = 20000 in Tab. 1).

Der Wert Z, der aus der Analyse dieses Materials folgte, diente als Grundlage für die Berechnung der Häufigkeitsstruktur, wie sie in Abb. 4 dargestellt wird. Sie stellt dadurch die "Restaurierung" der Häufigkeitsstruktur des Materials von Guiraud bei verschiedenen Stichprobenumfängen dar: N_1 = 2000, N_2 = 20000, N_3 = 200000. Starke horizontale Linien am Ende jeder "Treppe", die leicht von der theoretischen Grenze abweichen, stellen empirisch beobachtete "Rechtecke" der hapax legomena dar. Wie aus der Tabelle ersichtlich ist, sollte bei allen dazwischenliegenden Umfängen die Übereinstimmung der theoretischen und experimentellen Daten besser sein als bei diesen extremen Umfängen.

Wie man außerdem in der Tabelle erkennt, liegen die Prognosen für hapax legomena im Durchschnitt etwas höher (ungefähr um 4%); berücksichtigt man diese Verschiebung, dann ist die Unstimmigkeit auch unbedeutend. Kalinins Prognosen erweisen sich trotz des großen Umfangs der verwendeten Ausgangsinformation als weniger genau und als weniger "weitreichend".

Die Hypothese also, daß bei den Stichproben von 27000 Wörtern aus dem Material von Guiraud das Zipf-Mandelbrotsche Gesetz in der "kanonischen" Form (30) gilt (d.h. daß für dieses Mate-

rial der Zipfsche Umfang = 27000 ist), erlaubt es, die Kurve des Vokabularwachstums und die Anzahl der hapax legomena zufriedenstellend zu berechnen. In der vorliegenden Arbeit streben wir nicht nach einer detaillierten Überprüfung des vorgeschlagenen Modells am empirischen Material (das dargelegte Beispiel dient nur zur Illustrierung des Funktionierens des Modells). Weitere Entwicklungen zum Zweck der Beschreibung der Häufigkeitsstruktur des Lexikons und zahlreiche Vergleiche mit empirischen Daten kann man in Orlov (1976) finden. Wie aus den Resultaten dieser Arbeit ersichtlich ist, erlaubt der Wert $\alpha = 1$ die Häufigkeitsstruktur und das Vokabularwachstum in sehr heterogenem lexikalischen Material, dessen Häufigkeitsgraphen manchmal sehr gekrümmt sind, zufriedenstellend zu beschreiben.

Man kann also annehmen, daß α eine Konstante ist, die mit dem Typ der linguistischen Einheiten zusammenhängt (0 für Buchstaben, 1 für Wörter), während Z für einen konkreten Text oder eine Textauswahl charakteristisch ist. Etwas verallgemeinernd kann man sagen, daß α eine Charakteristik einer sprachlichen Ebene (oder, im weiteren Sinne, eines Informationssystems, eines Kodes, der langue), und Z eine Charakteristik der parole (der Nachricht) ist. Der früher nicht entdeckte Zusammenhang dieser Charakteristika mit dem Stichprobenumfang, der zur "Abweichung des Schweifs" führte, diskreditierte die beobachteten Gesetzmäßigkeiten und erlaubte es nicht, von ihnen ausgehend begründete quantitative Schätzungen durchzuführen.

Die Frage, welchen Wert α für andere linguistische und informationstragende Einheiten wie Phonem, Silbe, Digram, Trigram, zwei- und dreigliedrige Wortgruppen usw. annimmt, bleibt vorläufig offen.[3)] Es sind umfangreiche Untersuchungen nötig, aber es ist offensichtlich, daß es nur unter Berücksichtigung der in dieser Arbeit beschriebenen Dynamik der Häufigkeitsstruktur möglich sein wird, sich in der Vielfalt der empirischen Häufigkeitskurven zurechtzufinden. Vorläufige Rechnungen zeigten auch die prinzipielle Adäquatheit des gegebenen Modells für die Beschreibung der Anteile der Farbflächen in malerischen Kunstwerken (vgl. Vološin&Orlov 1972); eine Erscheinung, die der Erfüllung des "kanonischen" Zipf-Mandelbrotschen Gesetzes in abgeschlossenen literarischen Texten vollkommen analog ist, wurde auch in musikalischen Texten beobachtet (vgl. Boroda 1974). Alles dies bezeugt sowohl die außerordentliche Wichtigkeit als auch die außerordentliche Verbreitung des verallgemeinerten Zipf-Mandelbrotschen Gesetzes in dem Bereich der Kommunikationsmittel.

6

In der vorliegenden Arbeit haben wir das Problem der Herkunft und der Begründung des kanonischen Zipf-Mandelbrotschen Gesetzes absichtlich nicht berührt, sondern es als gegeben angenommen. Im Rahmen des dargelegten Modells scheint diese Form lediglich ein Spezialfall zu sein, der durch eine breitere Klasse von Häufigkeitsstrukturen mit gekrümmtem "Schweif" (§4) wesentlich verallgemeinert wird. Die eigenartige Tatsache, daß Häufigkeitsstrukturen größerer literarischer und musikalischer Werke ausgerechnet zu dieser kanonischen Form tendieren (vgl. Orlov 1974, Boroda 1974), zwingt uns, das Problem von einer anderen Seite anzuschauen. Es ist nicht möglich, diese Erscheinung durch irgendwelche statistischen Metatheorien zu erklären. Es muß eingesehen werden, daß die Realisierung der kanonischen Form der Häufigkeitsstruktur gerade in der vollen Länge der Werke sehr wichtig für ihre Perzeption ist und daß der Mensch im Laufe der Schöpfung der Texte imstande ist, ihre Häufigkeitsstruktur so zu organisieren, daß die volle Länge des Textes dem Zipfschen Umfang nahe kommt. Eine detaillierte Besprechung dieses Problems findet man in Arapov & Efimova & Šrejder (1975) und Orlov (1974, 1976).

Wenn man sich schon darüber wundert, daß die Häufigkeitsstruktur eines Textes mit seiner Länge korrespondiert, was ja zielgerichtete Bemühungen des Autors voraussetzt, so muß man sich erst recht darüber wundern, daß der Zipfsche Umfang auch bei lexikalischen Stichproben existiert, die nicht von einem einheitlichen organisierenden Willen erzeugt worden sind (wie z.B. die

durch Kolguškin ausgezählten militärischen Texte oder Kalininas russische Texte über Elektronik, [vgl. Orlov (1976)]. Daraus muß gefolgert werden, daß sich die Worthäufigkeiten "von alleine" zu einer der zahlreichen Formen des Zipfschen Gesetzes formieren - mit anderen Worten, es muß einen latenten, rein statistischen Mechanismus geben, der anscheinend sehr eigenartig ist, - und daß die Übereinstimmung des Zipfschen Umfangs mit der Textlänge in künstlerischen Werken durch eine "parametrische Unterordnung" unter die Länge des geplanten Textes, die mit der "Einstimmung" und "Regulierung" dieses Mechanismus in jedem konkreten Fall verbunden ist, zustande kommt. Die äußeren Charakteristika dieses Mechanismus (wie "input-output") werden durch das dargelegte Modell beschrieben, jedoch muß seine innere Konstruktion noch analysiert werden. Dies ist äußerst wichtig, da man darin die rein statistischen Erscheinungen von den störenden "menschlichen" Faktoren trennen muß.

Zum Schluß fassen wir die Resultate der vorliegenden Arbeit kurz zusammen:

1. Es wird eine neue Art der Beschreibung der Häufigkeitsstruktur vorgeschlagen, die sich von der üblichen geordneten Folge von Häufigkeiten unterscheidet. Der grundlegende Vorzug der vorgeschlagenen Beschreibung, d.h. der Verwendung der Funktion der strukturellen Dichte f(p) oder der äquivalenten Funktionen F(p) bzw. v(p) liegt in der Ersetzung des oft schwierigen Summierens von Zahlenfolgen durch eine Integration.

2. Es wurde die verallgemeinerte Form des Zipf-Mandelbrotschen Gesetzes als eine Funktion der strukturellen Dichte (20) formuliert. Dabei hat sich gezeigt, daß bei $\alpha > 0$ diese Funktion dieselbe Häufigkeitsstruktur wie die Mandelbrotsche Formel (1) beschreibt. Bei $\alpha \leq 0$ erhält man Häufigkeitsfolgen, die bisher unbekannt waren, aber einige von ihnen finden ihre Entsprechungen im Informationsprozeß (Buchstaben, melodische Intervalle). Die gemeinsame formale Genese unterschiedlicher Strukturen zeugt offensichtlich von der Einheitlichkeit der Informationsprozesse auf verschiedenen Ebenen. Außer dieser verallgemeinernden Rolle hat die Darstellung der Zipf-Mandelbrotschen Häufigkeitsstruktur in Form einer Funktion der strukturellen Dichte den Vorteil, daß nur ein unabhängiger Parameter $\alpha = 1/\gamma$ übrig bleibt, die restlichen, K und B kann man als Funktionen von p_1 und p_v ausdrücken.

3. Die Untersuchung kleiner Stichproben, in denen, wie in den lexikalischen, immer hapax legomena vorhanden sind, ermöglichte es, die grundlegenden beobachteten Parameter dieser Stichproben (Umfang, Häufigkeitsfolge, Vokabular, Zahl der einmaligen, zweimaligen usw. Wörter) durch die Hilfe des unabhängigen Parameters α miteinander zu verbinden.

4. Die Untersuchung eines unendlichen homogenen Textes, der beim Umfang Z dem Zipf-Mandelbrotschen Gesetz in seiner "kanonischen Form" folgt, hat gezeigt, daß Z der einzige Wert ist, bei dem diese Form vorkommen kann. Bei allen anderen Stichprobenumfängen folgen unvermeidlich "Abweichungen der Schweife". Die Kenntnis des Z-Wertes erlaubt es, sowohl die Dynamik der Veränderung der Häufigkeitsstruktur, als auch das Anwachsen des Vokabulars bei wachsendem Stichprobenumfang zu beschreiben. Es wurde auch die besondere Rolle von Z geklärt, nämlich seine Beziehung zu dem relativen Vokabularreichtum des Textes [Ungleichung (46)].

5. Der Vergleich mit empirischen Daten, der sowohl hier als auch in anderen Arbeiten durchgeführt wurde, hat gezeigt, daß die Häufigkeitsstruktur und das Vokabularwachstum innerhalb des Textes mit Hilfe des dargelegten Modells bei $\alpha = 1$ adäquat beschrieben werden kann. Das Modell ermöglichte es, eine Übereinstimmung in der Häufigkeitsstruktur des Vokabulars mit vollem Umfang zwischen literarischen und musikalischen Texten zu entdecken. Die Berücksichtigung des Einflusses des Zipfschen Umfangs bei kleinen Stichproben läßt hoffen, daß es auch bei anderen linguistischen Einheiten außer den Wörtern gelingt, die universelle Bedeutung von α für jede solche Einheit zu finden.

In dieser Arbeit wurde also ein Modell der Häufigkeitsstruktur des Zipfschen Typs aufgestellt, das gesetzmäßige Deformationen eben dieser Struktur beschreibt. Wenn man eine klare Vorstellung davon hat, daß die kanonische Häufigkeitsstruktur des Typs (1) nur ein Spezialfall einer breiteren Klasse verwandter

Häufigkeitsstrukturen ist, und wenn man die Dynamik der Häufigkeitsstruktur berücksichtigt, so kann man ohne den überflüssigen Empirismus auskommen, der sich beim Konstruieren aller möglichen spitzfindigen Formeln zur Approximation an empirische Daten einstellt.

Zum Schluß möchte sich der Verfasser bei M.V. Arapov, A.L. Brudno, E.M. Dumanis, N.M. Poliektov-Nikoladze, A.R. Chvoles, R.Ja. Čitašvili und Ju. A. Šrejder, deren Bemerkungen, Ratschläge und kollegiale Unterstützung beim Entstehen dieser Arbeit sehr behilflich gewesen sind, herzlichst bedanken.

ANHANG

Die eingeführte stetige Approximation der Häufigkeitsstruktur in Form einer Funktion der strukturellen Dichte erlaubt es, nicht nur die vorhandenen Parameter miteinander zu verknüpfen und die Werte der Koeffizienten K und B in der Mandelbrotschen Formel zu finden, sondern sie ermöglicht es auch, sowohl an die anderen bekannten Darstellungen der Häufigkeitsstruktur anzuknüpfen als auch neue Darstellungsarten zu finden. Als Beispiel erläutern wir kurz zwei Darstellungsformen des Zipf-Mandelbrotschen Gesetzes, die einen direkten Vergleich mit empirischen Daten erlauben.

Die erste Form ist in den Fällen geeignet, wenn man einen Vergleich mit empirischem Material unternehmen muß, das in Form von Tabellen oder Graphiken der sogenannten "Textauffüllung mit dem Vokabular" gegeben ist. Wir führen die Funktion der kummulativen Häufigkeit ein:

$$\varphi(i) = \sum_{j=1}^{i} p_j, \qquad i = 1,2,\ldots,v \tag{47}$$

wo die Menge der empirischen Häufigkeiten $\{p_j\}$, nach abnehmender Größe geordnet ist. Die Größe $\varphi(i)$ stellt die Wahrscheinlichkeit der zufälligen Wahl eines Wortes aus der gegebenen Stichprobe dar, dessen Häufigkeit in dieser Stichprobe nicht kleiner als p_i ist. Wenn $f(p)$ die Häufigkeitsstruktur der Stichprobe beschreibt, dann erhält man aufgrund von (13)

$$\varphi(i) = P(p > p_{i+1}) = \int_{p_{i+1}}^{p_1} f(p)dp. \tag{48}$$

Wenn $f(p)$ durch (20) mit $\alpha = 1$ gegeben ist, dann ist

$$\varphi(i) = \left(\ln \frac{p_1}{p_{i+1}}\right)\left(\ln \frac{p_1}{p_v}\right)^{-1}, \tag{49}$$

wobei man p_i aus (30) erhält.

Ein Beispiel derartiger Darstellung des Zipf-Mandelbrotschen Gesetzes und die Form der theoretischen und empirischen Kurven wurde in Orlov (1974) gebracht.

Die zweite Darstellungsform der untersuchten Gesetzmäßigkeiten ist bisher weder in der Praxis noch in der Theorie der quantitativen Linguistik bekannt; wir werden jedoch zeigen, daß sie einige Vorteile hat.

Wir führen den Begriff einer "geometrischen Häufigkeitsklasse" ein (im weiteren wird das Wort "geometrisch" einfachheitshalber ausgelassen). Dann zerlegen wir das Intervall der Häufigkeiten $[p_v,p_1]$ in eine Folge von disjunkten Teilintervallen, die einander in Punkten $\varphi_j = p_v C^{j-1}$ berühren, wo $C > 1$ eine beliebige Konstante ist. Die Längen solcher Teilintervalle bilden eine geometrische Folge mit der Basis C.

Bei der Häufigkeitsstruktur des Lexikons werden wir diejenigen Elemente (Wörter), die in das j-te Teilintervall $[p_j,p_{j+1})$ fallen, als die j-te Häufigkeitsklasse, und die Anzahl unterschiedlicher Elemente $v[p_j,p_{j+1})$ in dieser Klasse als den Umfang der j-ten Häufigkeitsklasse bezeichnen. Die Größe, die die Wahrschein-

lichkeit der zufälligen Wahl des Elements der j-ten Häufigkeitsklasse darstellt,

$$P[\rho_j,\rho_{j+1}) = \sum_{\rho_j \le p_i < \rho_{j+1}} p_i, \tag{50}$$

bezeichnen wir als das relative Gewicht der j-ten Häufigkeitsklasse, und $N[\rho_j,\rho_{j+1}) = NP[\rho_j,\rho_{j+1})$, wobei N der Stichprobenumfang ist, als das absolute Gewicht dieser Gruppe.

Aus (13), (17) und (20) erhält man leicht

$$v[\rho_j,\rho_{j+1}) = \frac{A}{\alpha}\left(\rho_{j+1}^{-\alpha} - \rho_j^{-\alpha}\right) \tag{51}$$

$$P[\rho_j,\rho_{j+1}) = \frac{A}{1-\alpha}\left(\rho_{j+1}^{1-\alpha} - \rho_j^{1-\alpha}\right). \tag{52}$$

Das Verhältnis des Umfangs einer Häufigkeitsklasse zu dem Umfang der vorhergehenden Häufigkeitsklasse findet man als

$$\frac{v[\rho_{j+1},\rho_{j+2})}{v[\rho_j,\rho_{j+1})} = \frac{C^{-\alpha j-2\alpha} - C^{-\alpha j-\alpha}}{C^{-\alpha j-\alpha} - C^{-\alpha j}} = C^{-\alpha}. \tag{53}$$

Analog folgt für die Gewichte zweier aufeinander folgenden Klassen

$$\frac{N[\rho_{j+1},\rho_{j+2})}{N[\rho_j,\rho_{j+1})} = \frac{P[\rho_{j+1},\rho_{j+2})}{P[\rho_j,\rho_{j+1})} = C^{1-\alpha}. \tag{54}$$

Daraus ergibt sich: Wenn die Längen der Häufigkeitsteilintervalle eine wachsende geometrische Folge mit der Basis C bilden, dann bilden die Umfänge der betreffenden Häufigkeitsklassen eine abnehmende geometrische Folge mit der Basis $C^{-\alpha}$ und die Gewichte dieser Klassen eine Folge mit der Basis $C^{1-\alpha}$. D.h., wenn man die Funktion $v(\rho_j,\rho_{j+1})$ in ein Koordinatensystem mit doppellogarithmischem Maßstab gegen ρ_j auf der Abszisse einträgt, so erhält man eine Gerade mit dem Steigungskoeffizienten $-\alpha$. Auch für die Gewichte der Häufigkeitsklassen erhält man eine Gerade, jedoch mit dem Steigungskoeffizienten $1-\alpha$.

Wenn also die Rangordnungsdarstellung der Häufigkeitsstruktur laut (1) im allgemeinen im Bereich der großen Häufigkeiten eine durch den Parameter B erfaßte Krümmung hat, so erscheint die Darstellung dieser Struktur mit Hilfe der Umfänge oder der Gewichte der geometrischen Häufigkeitsklassen im doppellogarithmischen Maßstab immer als eine Gerade. Wegen dieses Umstandes ist die Zerlegung der Menge der Häufigkeiten in Häufigkeitsklassen ein effektives Verfahren zum Test, ob das Zipf-Mandelbrotsche Gesetz in seiner kanonischen Form erfüllt ist, sowie zur Schätzung des Parameters α. Im Falle kleiner Stichproben empfiehlt es sich, die Zerlegung in Teilintervalle auf der Achse der absoluten (nicht relativen) Häufigkeiten durchzuführen, indem man mit 1 anfängt und die Konstante C gleich 2 setzt[4]. Da die größte beobachtete Häufigkeit im allgemeinen nicht zweiter Ordnung ist, braucht man bei der Schätzung von α aus dem Diagramm das letzte Teilintervall nicht zu berücksichtigen[5].

Beispiele der Darstellung des empirischen Materials in Form von Häufigkeitsklassen wurden in Orlov (1970) gebracht; alle dort untersuchten Strukturen sind durch $\alpha = 1$ charakterisiert. Auf der Abb. 5 findet man die Diagramme von Umfängen der Häufigkeitsklassen über Verknüpfungen von je zwei (A), je drei (B), je vier (C) und je fünf (D) aufeinanderfolgenden melodischen Intervallen in Nocturnen von F. Chopin [die Berechnung wurde in Orlov (1970) erklärt]. Es wurde insbesondere untersucht, wie die Größe α beim Übergang zu einer Einheit höheren Typs zunimmt[6].

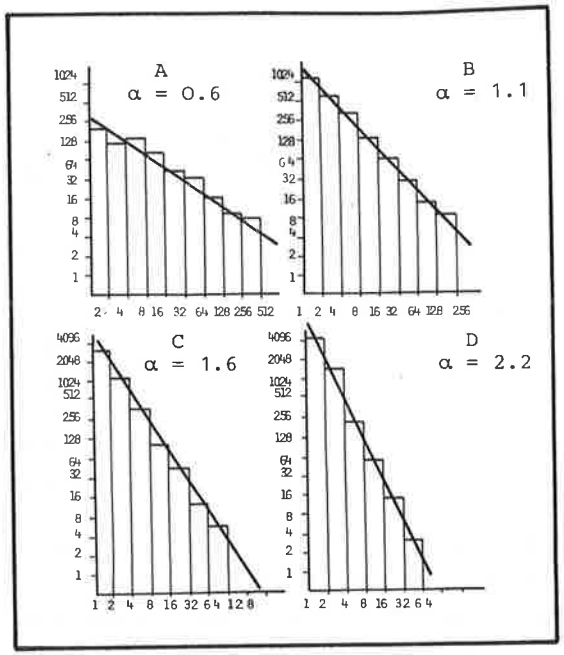


Abb. 5. Häufigkeitsstrukturen der Folgen melodischer Intervalle in Chopins Nokturnen (eigene Zählungen zusammen mit M.G. Boroda)

Anmerkungen

1 Es ist interessant, diesen Ausdruck mit der Häufigkeitsreihe $p_i = pe^{Bi}$, die Mandelbrot (1957) erhalten hat, zu vergleichen. In Mandelbrots Modell stellt diese Reihe die Häufigkeitsstruktur einer degenerierten "Sprache" dar, deren Alphabet aus einem Symbol und dem Zeichen für die Pause besteht.

2 Die Methoden der approximativen Lösung dieser Gleichung wurden in Orlov (1978a, 1980) dargelegt.

3 Zur Zeit der Entstehung dieser Arbeit (1973-1974) analysierte der Verfasser eine große Menge von empirischem Material auf verschiedenen linguistischen Ebenen. Es hat sich gezeigt, daß in den meisten Fällen die Wahl des Parameters Z völlig ausreichte; bisweilen mußte man auch den Wert von p_1 wählen (z.B. in georgischen Texten übersteigt die Häufigkeit des häufigsten Wortes 3-5 mal die Häufigkeit des zweithäufigsten Wortes. In solchen Fällen führt der Vergleich von p_1 mit der beobachteten Häufigkeit des häufigsten Wortes zu einer schlechten Leistung des Modells, und man muß p_1 so wählen, daß die Summe der Häufigkeiten der ersten 20-30 häufigsten Wörter im Modell der entsprechenden Summe im Text gleich ist). Bei der Wahl von α ergaben sich jedoch keine Schwierigkeiten und daher ist es plausibel anzunehmen, daß α = 1 der universelle Wert dieser Konstanten ist, unabhängig von der Art linguistischer Einheiten (außer Buchstaben und anderen Einheiten niedrigerer Ebenen).

4 Solche Häufigkeitsklassen nennt man in der Literatur gewöhnlich Oktaven. Ein Beispiel der Zerlegung in Oktaven findet man in Nadarejšvili & Orlov (1969).

5 Der folgende Fall ist z.B. gut möglich: die Häufigkeit des häufigsten Wortes im Text ist gleich 1050 und sie fällt in die Oktave 1024-2048; daher ist der größte Teil dieser Oktave "leer". Dadurch wird der reguläre Verlauf der "Treppen" (wie in Abb. 5) in dieser letzten Oktave gestört.

6 Die Zählung der Folgen von melodischen Intervallen erfolgt durch die sogenannte "gleitende Erhebung". Numeriert man alle Intervalle im Text laufend, so untersucht man zuerst die Folge mit den Zahlen 1,2,3, danach die mit 2,3,4, dann mit 3,4,5 usw. analog wie man es mit Digrammen, Trigrammen u.a. in linguistischen Massiven tut.

Ein Modell der Häufigkeitsstruktur des Vokabulars

Ju.K. Orlov

In den letzten 50 Jahren sind auf dem Gebiet der Lexikostatistik zahlreiche experimentelle Arbeiten (Häufigkeitswörterbücher, Konkordanzen, usw.) wie auch theoretische Arbeiten (vgl. Zipf 1935, 1949; Mandelbrot 1953, 1957 und viele andere) erschienen, die der Untersuchung und der theoretischen Begründung bzw. Verallgemeinerung empirischer Gesetzmäßigkeiten gewidmet waren. Im Augenblick befindet sich jedoch die Untersuchung der quantitativen Seite des Problems in einer gewissen Krise. Auch wenn man die Häufigkeitswörterbücher für die Kompilation von Minimalwörterbüchern, für maschinelle Übersetzung usw. verwenden kann, so fehlen jegliche verallgemeinernde Arbeiten über die Statistik des Lexikons.

Der Grund liegt in einer gewissen Kluft zwischen der Intention der theoretischen Arbeiten, die üblicherweise das Zipfsche Gesetz irgendwie "begründen" möchten, und den aktuellen Bedürfnissen der linguostatistischen Praxis.

Die allgemeine Kenntnis des Umstandes, daß der Bereich der Wörter mit mittlerer Häufigkeit dem Zipfschen Gesetz folgt (vgl. Frumkina 1961; Finkenstaedt & Wolf 1969), erlaubt es nicht, irgendwelche Rechnungen durchzuführen; ganz unbefriedigend ist das sogenannte Maß der lexikalischen Konzentration, das einen Vergleich des relativen Vokabularreichtums in Texten unterschiedlicher Länge erlauben sollte. Die meistbenutzte Beziehung $R = v/N$ (v = Stichprobenvokabular, N = Stichprobenumfang) ist ganz ungeeignet, da sich dieses Verhältnis innerhalb eines Textes beträchtlich ändert. Beispielsweise in Puškins "Kapitänstochter" (nach Angaben von Josselson 1953) beträgt es in einer Stichprobe von 5000 Wortwendungen 0,313; bei 10000 Wortwendungen 0,243 und im gesamten Text ($N = 29345$) 0,163. Unseres Erachtens sind jegliche Formeln, die diese Beziehung enthalten, ungeeignet (vgl.

Tešitelová 1972). Etwas besser ist der Vorschlag von Guiraud (1954): $R = v/\sqrt{N}$; aber auch diese Größe kann sich im Rahmen eines Textes ändern (für die "Kapitänstochter" ergeben sich die entsprechenden Zahlen: 22,2; 24,3; 27,9).

Wenn man nur das Vokabular bei einem Textumfang kennt, so kann man es für einen anderen Umfang nicht "umrechnen"; sollte diese Möglichkeit gefunden werden, so würde sich das o.a. Problem von alleine lösen. Manchmal werden für empirische Daten die Werte der Koeffizienten K, B und γ der Mandelbrotschen Formel

$$p_i = \frac{K}{(B+i)^{\gamma}}, \qquad (1)$$

(wo p_i die Häufigkeit des i-ten Wortes in einem nach abnehmender Häufigkeit geordneten Wortinventars ist) berechnet, aber die linguistische Bedeutung dieser Konstanten ist unklar, und es ist schwer, irgendwelche Gesetzmäßigkeiten ihrer Veränderung beim Übergang von einem Text zum anderen festzustellen. Die Menge dieser Konstanten, die die Häufigkeitsstruktur einer konkreten Stichprobe beschreibt, ist ebensoweinig mit der analogen Menge einer anderen Stichprobe vergleichbar wie die Häufigkeitsreihen selbst. Deswegen entstand schon längst die Notwendigkeit, ein mathematisches Modell der Häufigkeitsstruktur des Lexikons aufzustellen, das die Lösung der aufgezählten Probleme und darüberhinaus weiterer Probleme ermöglicht.

In der vorliegenden Arbeit wird ein Modell dargestellt, das folgendes leistet:

(1) Die Berechnung des Vokabulars, der Häufigkeitskurve und des lexikalischen Spektrums (Zahl der einmaligen, zweimaligen usw. Wörter) bei beliebigem Testumfang, wenn das Vokabular in einer Stichprobe aus dem gegebenen Text bekannt ist.

(2) Schätzung der möglichen zufälligen Abweichung des tatsächlichen Vokabulars von der theoretischen Voraussage.

(3) Schätzung des Grads der "statistischen Ähnlichkeit"zweier oder mehrerer Texte, die in eine Stichprobe zusammengefaßt wurden.

(4) Charakterisierung des Vokabularwachstums in einem gegebenen Text und des lexikalischen Spektrums in einem beliebigen Abschnitt durch einen den Daten des Texts (oder der Stichprobe) angepaßten Parameter, der direkt als eine Konvention zur Messung des relativen Vokabularreichtums verwendet werden kann.

1.

Bei der Ableitung des Modells gehen wir von einigen Resultaten von Kalinin (1964, 1965) aus. Im folgenden stellen wir kurz diejenigen Ergebnisse dar, die wir in der vorliegenden Arbeit verwenden werden.

Kalinin analysiert den "unbegrenzten Text", in dem jedes Wort eine a priori Vorkommenswahrscheinlichkeit hat, die von der Position im Text und von den früher verwendeten Wörtern unabhängig ist. Obwohl diese Voraussetzung streng genommen in den reellen Texten nicht erfüllt wird, erlaubt sie trotzdem, einige Beziehungen abzuleiten, die mit der Realität annehmbar übereinstimmen.

Die Unabhängigkeitsannahme erlaubt es, die Beziehung zwischen der mathematischen Erwartung des Vokabularumfanges $v(N)$ und der Zahl der m-mal vorkommenden Wörter $v_m(N)$ bei verschiedenen Umfängen N abzuleiten.

Wenn die mathematischen Erwartungen $E[v(N_o)]$ und $E[v_m(N_o)]$ bei dem Umfang N_o bekannt sind, so gilt bei beliebigem Umfang N

$$E[v(N)] = E[v(N_o)] - \sum_{j\geq 1} E[v_j(N_o)]\left(1 - \frac{N}{N_o}\right)^j \qquad (2)$$

und

$$E[v_m(N)] = \frac{1}{m!}\left(\frac{N}{N_o}\right)^m \sum_{j\geq 1} E[v_j(N_o)]j(j-1)\ldots(j-m+1)\cdot$$

$$\cdot \left(1 - \frac{N}{N_o}\right)^{j-m}, \qquad (3)$$

wo E die mathematische Erwartung bedeutet.

Aus diesen Beziehungen sind die uns unbekannten Wahrscheinlichkeiten des Wortvorkommens ausgeschlossen; dabei ist die Umrechnung sowohl "rückwärts" (für $N < N_o$) als auch "vorwärts" (für $N > N_o$) möglich. (Wenn man anstelle der mathematischen Erwartungen $E[v(N_o)]$ und $E[v_m(N_o)]$ nur die beobachteten Werte $\hat{v}(N_o)$ und $\hat{v}_m(n,N_o)$ beim Umfang N_o verwendet, dann ist nur die Umrechnung "rückwärts" möglich.) Für die Varianz der Größen $v(N)$ und $v_m(N)$ gibt Kalinin folgende Beziehungen an:

$$V[v(N)] = E[v(2N)] - E[v(N)] - \frac{\{[E[v_1(N)]\}^2}{N} + \frac{E[v_2(2N)]-E[v_2(N)]}{N} + \varepsilon \qquad (4)$$

$$V[v_m(N)] = E[v_m(N)] - \frac{E[v_{2m}(2N)]}{\sqrt{\Pi m}} - \frac{\{E[v_m(N)]\}^2}{N} + \varepsilon \qquad (5)$$

wo V die Varianz bedeutet.

Wenn man also das erwartete lexikalische Spektrum des Textes, beschrieben durch $E[v_m(N_o)]$, bei einem gegebenen Umfang N_o kennt, dann kann man die Häufigkeitsstrukturen des Textes bei einem beliebigen anderen Umfang berechnen und die Zufälligkeit bzw. Signifikanz der beobachteten Abweichungen beurteilen.

In dem dargelegten Modell spielt eine grundlegende Rolle der Umfang Z, der die Häufigkeitsstrukturen nach dem Zipf-Mandelbrotschen Gesetz gestaltet (vgl. Orlov 1970). Die formale Analyse einer derartigen Struktur, wie in Orlov (1976) durchgeführt, beruht auf folgenden Annahmen:

(a) Das Vorkommen von hapax legomena (einmal vorkommende Wörter) im Text ist obligatorisch (eine Beobachtung von Yule 1944);

(b) die Größe γ in der Mandelbrotschen Formel (1) ist gleich eins.[1)]

Dies führt zu folgenden Beziehungen:

$$p_i = \frac{\frac{1}{\ln (Zp_{max})}}{\frac{1}{p_{max} \ln (Zp_{max})} - 1 + i}; \qquad (6)$$

$$v(Z) = \frac{Z - \frac{1}{p_{max}}}{\ln (Zp_{max})}; \qquad (7)$$

$$v_m(Z) = \frac{v(Z)}{m(m+1)}, \qquad (8)$$

wo p_{max} die größte im Text gefundene relative Worthäufigkeit ist.

Vergleicht man (6) mit (1), so sieht man, daß sie identisch sind, wenn

$$K = \frac{1}{\ln Zp_{max})}; \quad B = \frac{K}{p_{max}} - 1; \quad \gamma = 1. \qquad (9)$$

Die Werte von K und B kann man also mit Hilfe von Z und P_{max} darstellen.

Betrachtet man die durch (8) bestimmten Größen $v_m(Z)$ als die mathematischen Erwartungen der Anzahl unterschiedlicher m-mal vorkommender Wörter beim Textumfang Z[2)], dann kann man durch Einsetzung von (7) und (8) in (2) und (3) den Vokabularumfang v(N,Z) und die Anzahl $v_m(N,Z)$ m-maliger Wörter bei beliebigem Umfang N erhalten (hier lassen wir das Zeichen der mathematischen Erwartung weg):

$$v(N,Z) = v(Z) \sum_{j=0}^{\infty} \frac{(1-X)^j}{j+1}; \qquad (10a)$$

$$v_m(N,Z) = v(Z) \sum_{j=0}^{\infty} \frac{(1-X)^j}{(j+m)(j+m-1)}, \qquad (11a)$$

wo $X = Z/N$[3)]. Wenn X von 1 stark abweicht, dann bekommen diese Größen eine andere Form.[4)]

$$v(N,Z) = v(Z) \frac{\ln X}{X - 1}; \qquad (10b)$$

$$v_1(N,Z) = \frac{v(Z) - Xv(N,Z)}{1 - X}; \qquad (11b)$$

$$v_m(N,Z) = \frac{v_{m-1}(N,Z) - v_{m-1}(Z)}{1 - X}; \quad m = 2,3,... \qquad (11c)$$

(Die Rechnung mit der Rekursionsformel (11c) muß auf mehrere Dezimalstellen durchgeführt werden, z.B. mit einem Mikrorechner; falls $X \approx 1$, so soll man lieber Formel (11a) benutzen.)[5)]

Unser Modell der Häufigkeitsstruktur des Lexikons stellt also einen unbegrenzten, statistisch homogenen Text dar, der folgende Eigenschaften hat:

(1) Er besitzt immer hapax legomena, d.h. das Vokabular wächst unbegrenzt mit dem Anwachsen des Textumfangs.

(2) In Stichproben mit einem festen Umfang Z ist die mathematische Erwartung der Anzahl m-maliger Wörter und die des Vokabularumfangs durch (7) und (8) und die geordnete Häufigkeitsreihe durch (6) gegeben.

(3) In Stichproben mit beliebigem Umfang N ist die mathematische Erwartung der Anzahl m-maliger Wörter durch (11a) und (11b) und die des Vokabularumfangs durch (10a) und (10b) gegeben. Den letzten Ausdruck kann man als die Funktion des Vokabularanwachsens betrachten, die mit der Zunahme des Textumfangs in Zusammenhang steht, wenn Z

die Rolle eines Parameters hat. Der Verlauf der Häufigkeitskurve im Bereich der häufigen Wörter (deren relative Häufigkeiten sich bei der Änderung des Stichprobenumfangs nicht ändern) wird auch in diesem Fall durch (6) beschrieben.

Mit anderen Worten, wenn ein Text mit Umfang Z dem Zipf-Mandelbrotschen Gesetz folgt, dann ist dieser Umfang der einzige, bei dem das Gesetz erfüllt werden kann. Bei Stichproben mit einem beliebigen anderen Umfang sind Abweichungen unvermeidlich. Unten werden wir zeigen, daß die bekannte "Abweichung des Schweifes" im Bereich seltener Wörter eine Folge gerade dieses Umstandes ist. Den Umfang Z werden wir im weiteren als den "Zipfschen Umfang" bezeichnen.

Diese Größe Z ist eine wichtige Charakteristik des Textes. Außer der Möglichkeit, das Vokabular und das Spektrum bei beliebigem Umfang zu berechnen, liefert sie unmittelbar ein geeignetes Maß der lexikalischen Konzentration, ein Maß des relativen Vokabularreichtums. Wenn es also zwei Texte mit unterschiedlichen Werten von Z gibt so, daß $Z_1 > Z_2$, dann ist bei gleichem p_{max}

$$v(N,Z_1) > v(N,Z_2) \qquad (12)$$

für beliebiges N. Das heißt, eine Stichprobe mit Umfang N aus dem ersten Text wird ein reicheres Vokabular haben als eine Stichprobe desselben Umfangs aus dem zweiten Text, dessen Zipfscher Umfang kleiner ist. Mit anderen Worten, wenn man die Zipfschen Umfänge zweier Texte vergleicht, dann wird der Text ein reicherer Vokabular haben, dessen Z größer ist. Dies erlaubt uns, die Größe Z als ein konventionelles Maß des relativen Vokabularreichtums zu nennen.

Damit ist die Aufstellung des Modells beendet. Alles weitere hängt davon ab, wie man dieses Modell verwendet. In Orlov (1970, 1976) wurde festgestellt, daß Z im Extremfall der Länge eines großen literarischen Werkes entspricht.[6] Würden wir nämlich die volle Länge solcher Texte als Z in (6), (7) und (8) einsetzen, so erhielten wir eine Übereinstimmung der beobachteten und der

theoretischen Werte innerhalb der ± 20% Toleranz für alle Werke, deren Länge über 10-20 Tausend Wortverwendungen betrug. Stichproben aus diesen Texten würden mit ungefähr derselben Genauigkeit den Ausdrücken (10) und (11) (vgl. Orlov 1969) folgen. Es hat sich gezeigt, daß man in sehr großen Texten eine bessere Übereinstimmung erreichen kann, wenn man als Z die Umfänge der vom Autor festgelegten Teile dieser Texte (Bände, Bücher u.a.), einsetzt (vgl. Nadarejšvili, Orlov 1971), obwohl in einigen Fällen (z.B. in Joyce "Ulysses") die bessere Übereinstimmung bei vollem Text erfolgte.

Wie interessant diese Erscheinung an sich auch sein mag (der relative Vokabularreichtum wird durch die Textlänge bestimmt: die Wachstumskurve des Vokabulars ist desto steiler, je mehr der Autor zu schreiben vorhat) ihre Überprüfung würde durch die im allgemeinen niedrige Genauigkeit der theoretischen Prognose erschwert. Die Uneindeutigkeit der Prognose für große Texte (der Zipfsche Umfang kann sowohl der vollen Textlänge als auch der Länge eines Teiles entsprechen), die Unmöglichkeit der Prognose für kurze Texte (kürzer als 10000 Wortverwendungen) und die Unklarheit der Grenzen der "Zwischenzone" (10000-?), in der einige Texte mit ihrem Umfang den Gesetzmäßigkeiten (6) - (8) folgen und andere wiederum nicht, entwertet diese Hypothese (Z = volle Textlänge) für praktische Rechnungen.

Es wurde jedoch beobachtet, daß die Häufigkeitsstruktur kurzer Texte der Struktur von Abschnitten aus längeren Texten ähnelt, die in ihrer vollen Länge dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen und dieselbe charakteristische Krümmung der Treppenkurve oberhalb der theoretischen Kurve (6) im Bereich seltener Wörter haben. Diese Beobachtung führte zu der Hypothese: <u>für jeden Text gibt es einen eigenen Zipfschen Umfang</u> unabhängig sowohl von seiner Häufigkeitskurve als auch von seinem reellen Umfang. Mit anderen Worten, entweder kann man aus beliebigem Text Stichproben vom Umfang Z erheben, deren Vokabular und Häufigkeitsstruktur bis auf zufällige Abweichungen durch (6) - (8) genau beschrieben werden, oder man kann den Text (wenn sein Umfang kleiner als Z ist) als einen Abschnitt aus einem Text mit Umfang Z, für den (6) - (8) gelten, betrachten. Wenn

man Z kennt, so kann man in beiden Fällen das Vokabular und das Häufigkeitsspektrum bei beliebigem Umfang aufgrund von (10) und (11) berechnen.

Den letzten Umstand benutzen wir zur Überprüfung der aufgestellten Hypothese. Hat man den Wert von Z für den gegebenen Text geschätzt, dann berechnet man die theoretischen Werte derjenigen Textparameter, die zur Bestimmung von Z nicht direkt verwendet werden. Wenn die aufgestellte Hypothese korrekt ist, dann muß zwischen dem Vokabular und dem lexikalischen Spektrum des Textes bei Stichproben mit beliebigem Umfang der durch (10a), (10b) und (11b) gegebene Zusammenhang bestehen.

Um den Zipfschen Umfang zu schätzen, löst man bezüglich Z die Gleichung

$$v(N,Z) = \hat{v}(N) \qquad (13)$$

wo $\hat{v}(N)$ das beobachtete Vokabular der Stichprobe aus dem Text (eventuell der ganze Text), der den Umfang N hat, darstellt. Mit anderen Worten, wir führen die theoretische Kurve der Vokabularzunahme durch einen einzigen Punkt, nämlich durch den beobachteten Wert des Wortschatzes.

Der Ausdruck (13) ist bezüglich Z leider transzendental und läßt sich mit Hilfe einfacher Funktionen nicht ausdrücken.[7)]

Wenn das Vokabular $\hat{v}^{(')} = v(N_1,Z)$ beim Umfang N_1 gegeben ist, dann kann man offensichtlich nach der Bestimmung von Z für das Vokabular $v(N_2,Z)$ einen Umfang N_2 berechnen. Schätzen wir nun mit Hilfe unseres Modells den Grad möglicher zufälliger Abweichungen des beobachteten Vokabulars von dem prognostizierten beim Umfang N_2.

Wenn man Z genau kennen würde, dann könnte man die Dispersion der Abweichungen des beobachteten Wertes von dem prognostizierten Wert direkt mit Kalinins Formel (4) bestimmen. Setzt man in sie den Ausdruck für den Wortschatz (10b) ein, und behält man nur die ersten zwei Glieder der Entwicklung, so kann man schreiben:

$$V[v(N,Z)] = v(Z)\left(\frac{\ln\frac{Z}{2N}}{\frac{Z}{2N}-1} - \frac{\ln\frac{Z}{N}}{\frac{Z}{N}-1}\right) =$$

$$= v(N,Z)\left(\frac{\ln\frac{X}{2}}{\ln X}\cdot\frac{X-1}{\frac{X}{2}-1} - 1\right) = v(N,Z)\cdot G(X), \qquad (14)$$

wo

$$X = \frac{Z}{N};\ G(X) = \left(1 - \frac{\ln 2}{\ln X}\right)\cdot\frac{X-1}{\frac{X}{2}-1} - 1 \text{ ist.}$$

Für praktische Rechnungen ist es angebracht, die relative Standardabweichung in Prozenten zu benennen. Sie gleicht der Standardabweichung, dividiert durch den Erwartungswert und multipliziert mit 100, d.h.

$$\delta = \frac{\sqrt{V[v(N,Z)]}}{v(N,Z)}\cdot 100\% = \sqrt{\frac{G(x)}{v(N;Z)}}\cdot 100\%.$$

Der genaue Wert Z ist aber unbekannt, und wir schätzen ihn lediglich aus den Daten der Stichprobe. Der Fehler der Prognose setzt sich zusammen aus der zufälligen Streuung in dem prognostizierten Punkt und der Verschiebung des prognostizierten Punktes selbst durch zufällige Abweichungen des Wortschatzes im beobachteten Punkt.

Die relative Varianz der Prognosefehler kann man als die Summe

$$\delta^2 = \delta_1^2 w^2 + \delta_2^2 \qquad (15)$$

darstellen, wo

$$\delta_1 = \sqrt{\frac{G(X_1)}{v(N_1,Z)}} \cdot 100\%; \quad \delta_2 = \sqrt{\frac{G(X_2)}{v(N_2,Z)}} \cdot 100\%;$$

$$X_1 = \frac{Z}{N_1} \; ; \quad X_2 = \frac{Z}{N_2} \; ; \quad v(N,Z) = \hat{v}^{(1)}$$

ist.

Der Koeffizient W berücksichtigt die geometrischen Eigenschaften der Verschiebung des prognostizierten Punktes infolge der Verschiebungen des beobachteten Punktes.[8] Ein exakter analytischer Ausdruck ist aus demselben Grund nicht möglich wie die Lösung der Gleichung (13); eine annehmbare Approximation ist gegeben durch

$$W \approx \left(\frac{\lg v(N_2,Z)}{\lg \hat{v}^{(1)}}\right)^4 . \qquad (16)$$

Im Vergleich mit der exakten Berechnung der Veränderung des Abstandes zwischen den Kurven der Familie (10) bei unterschiedlichen Z ergibt sie einen relativen Fehler, der unter 6-8% liegt.

Auf diese Weise können wir also den Zufälligkeitsgrad der festgestellten Unterschiede zwischen der Voraussage $v(N_2,Z)$ und dem tatsächlich beobachteten Vokabularumfang beim Umfang N_2 abschätzen. Wir werden schließen, daß die festgestellte relative Abweichung rein zufällig ist, wenn sie 3δ nicht überschreitet. Dies entspricht einer Wahrscheinlichkeit größer als 0.003.

Wir bemerken, daß diese Schätzung für solche Stichproben gilt, die unabhängig aus dem Modelltext erhoben wurden. Wenn die Stichprobe mit dem Umfang N_1, für die man eine Prognose aufstellt, ein Teil der Stichprobe mit Umfang N_2 ist (oder umgekehrt) und dieser Umfang nur unbedeutend größer (kleiner) als N_2 ist, so vergrößert sich wesentlich die Genauigkeit der Prognose. Eine derartige Erhöhung der Prognosegenauigkeit wird in den Tabellen im Anhang demonstriert, wenn sich der Stichprobenumfang dem Punkt nähert, für welchen die Prognose aufgestellt

wurde. Falls der Umfang der Unterstichprobe viel kleiner ist als der der großen Stichprobe, so kann man bei diesen Prognosen beide Stichproben als unabhängig betrachten.

2.

Das Material, an dem die Adäquatheit des dargelegten Modells überprüft wurde, ist in den Tabellen 1a, 1b, 2a und 2b (s. Anhang 2) aufgeführt. Die benutzten Werke wurden durchlaufend numeriert, damit wir auf sie mit einer Zahl hinweisen können. Die Hinweiszahl wird in geschweifte Klammern gesetzt. Bevor wir die Tabellen analysieren, demonstrieren wir an einigen Beispielen ausführlich die Übereinstimmung des dargelegten Modells mit konkreten lexikalischen Stichproben.

In der Abb. 1 werden die Häufigkeitskurven des Romans "Kapitänstochter" von Puškin {28}[9] (obere Kurve) und ein Abschnitt aus dem Roman im Umfang von 5000 Wortverwendungen {30} dargestellt. Die Punkte und die dicken Striche stellen die beobachteten Häufigkeiten dar. Die halbdicken ununterbrochenen Kurven sind die theoretischen Häufigkeiten, die aufgrund von (6), in unserem Fall für Z = 45000, berechnet wurden. Die theoretischen Werte der Anzahl seltener Wörter, berechnet nach (11) für den Umfang N = = 29345, sind als "Treppenkurve" mit dünnen Strichen aufgetragen. Man kann leicht sehen, daß in diesem Fall, d.h. wenn N dem Z relativ nah steht, eine gute Übereinstimmung zwischen der Häufigkeitsreihe und dem genannten Verlauf der theoretischen Kurve (6) erreicht wird. Auf der Graphik des Abschnitts sieht man, daß der Treppenabschnitt im Bereich seltener Wörter (d.h. sowohl die empirischen "Rechtecke" als auch die theoretischen "Treppen") überwiegend über der theoretischen Kurve liegt und die typische "Krümmung des Schweifes" aufweist. Die Zahl der hapax legomena beträgt in diesem Fall mehr als die Hälfte des Vokabulars.

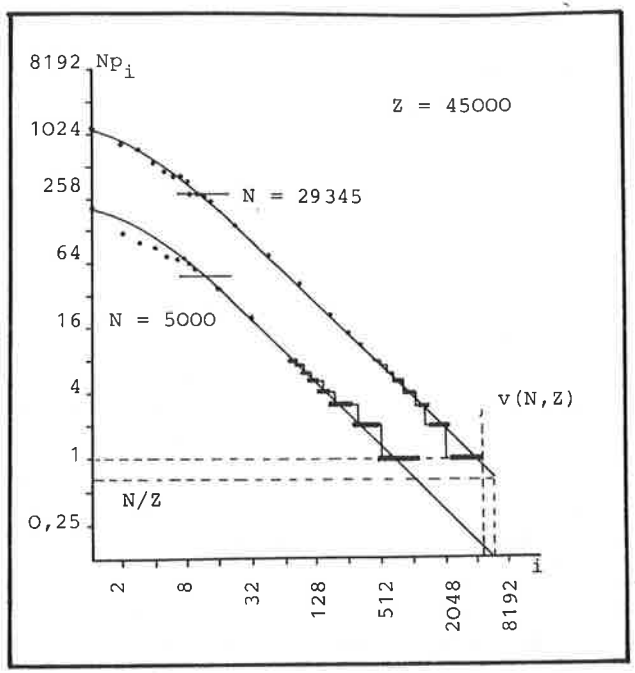


Abb. 1. Die Häufigkeitskurve von Puškins "Kapitänstochter" (obere Kurve) und von einem Abschnitt von 5000 Wortverwendungen {30}. In beiden Fällen ist laut (9) K = 0.134; B = 2.40. Die theoretische Kurve, berechnet nach (6) (die ununterbrochene halbdicke Kurve) läuft bis i = v(Z) = 6020, wo die Summe der relativen Häufigkeiten 1 erreicht, und $P_{v(Z)} = \frac{1}{Z}$.

Die obere Graphik auf der Abb. 2 stellt die Häufigkeitskurve der Gesamtwerke von Puškin {127} dar. Der Stichprobenumfang (N = 544777) überschreitet wesentlich den Zipfschen Umfang und die "Krümmung des Schweifes" erfolgt auf die andere Seite, d.h. unterhalb von (6). Obwohl in diesem Fall die Übereinstimmung der theoretischen "Treppenkurve" und der empirischen "Rechtecke" etwas schlechter ist[10], stimmt der allgemeine Krümmungstrend überein. Die Zahl der hapax legomena ist hier kleiner als die Hälfte des Vokabulars.

Die mittlere Graphik der Abb. 2 stellt die Häufigkeitskurve der Novelle "Pique Dame" {25} dar. Die auffallende Ähnlichkeit dieser Kurve mit der des Abschnitts aus der "Kapitänstochter" (vgl. Abb. 1) ist die Konsequenz der Ähnlichkeit der grundlegenden Textparameter (N = 6861, Z = 35000) mit denen des Abschnitts {30}.

Auf der anderen Seite zeigt "Skazka o rybake i rybke" {5} (die unterste Graphik auf der Abb. 2), die einen erheblich kleineren Zipfschen Umfang hat (N = 948, Z = 2200), am Anfang einen viel langsameren Verlauf der Häufigkeitskurve (relativer Überfluß häufiger Wörter), wodurch der Einfluß von Z auf die lexikalische Konzentration erklärt wird - denn je mehr relativ häufige Wörter im Text vorhanden sind, desto weniger Platz bleibt für die anderen.

Die Abbildungen 1 und 2 zeigen anschaulich die Dynamik der Veränderung der lexikalischen Struktur mit der Veränderung des Stichprobenumfangs. Solange N < Z ist, liegt der treppenförmige Teil der Graphik oberhalb der theoretischen Häufigkeitskurve und nimmt langsamer ab; versucht man den Verlauf der ganzen Kurve mit der Mandelbrotschen Formel (1) zu approximieren, so ergibt sich, daß $\gamma < 1$ ist. Je mehr sich der Stichprobenumfang dem Zipfschen Umfang nähert, desto gerader wird die Graphik: die rechten Ecken der Treppen nähern sich der Häufigkeitskurve (6) und das tatsächliche Vokabular nähert sich der Zahl v(Z); die Zahl von hapax legomena erreicht fast die Hälfte des Vokabulars. Wenn der Stichprobenumfang den Zipfschen Umfang übersteigt, so biegt sich der treppenförmige Teil der Graphik nach unten.

Die Größe p_{max} hat einen geringen Einfluß auf den Verlauf der Häufigkeitskurve im Bereich der Wörter mit mittleren und niedrigeren Häufigkeiten. Es ist eben dieser Umstand, der es erlaubt, einen approximativen Wert von p_{max} zu verwenden, wenn sein genauer Wert für die gegebene Stichprobe unbekannt ist. In den aufgeführten Tabellen wurden solche Orientierungswerte mit dem Zeichen ~ versehen. Sie wurden für die Sprache der Stichprobe als maßgebend betrachtet, wenn die Quelle keine genauen Daten

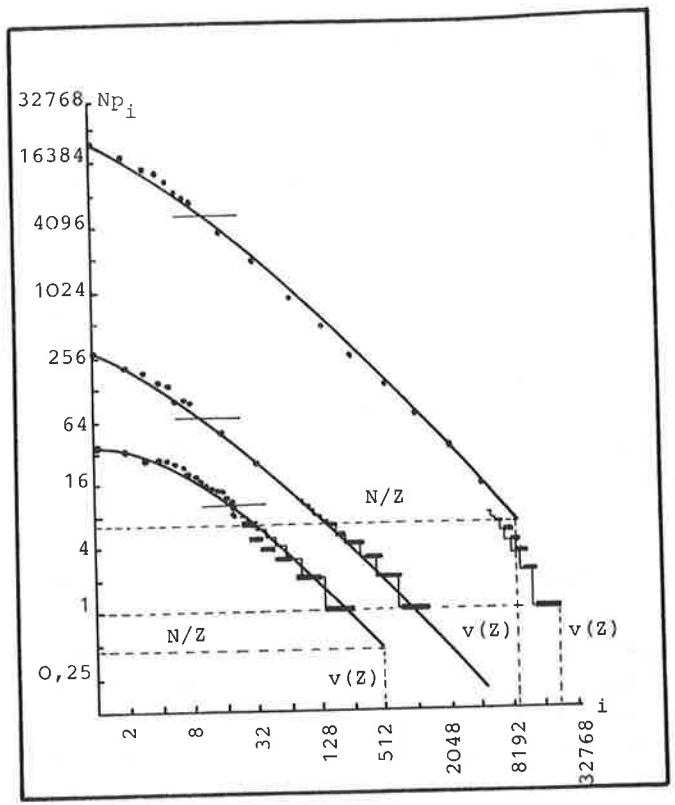


Abb. 2. Häufigkeitskurven der Gesamtwerke von A.S. Puškin {127} mit Z = 75000; K = 0,122; B = 1,54 (oben); "Pique Dame" {25} mit Z = 35000; K = 0,139; B = = 2,66 (in der Mitte) und "Skazka o rybake i rybke" mit Z = 2200; K = 0,228; B = 4,95 (unten). Wie hier und auf Abb. 1 ersichtlich, weisen die Häufigkeiten der häufigsten Wörter nichtreguläre Abweichungen von der Kurve (6) auf. Ähnliche nichtreguläre Abweichungen kann man auch an anderen Daten beobachten. Der Bereich, in dem diese Abweichungen aufhören, liegt bei den Häufigkeiten zwischen 0,01 und 0,005. Die relative Häufigkeit 0,01 ist auf den Abbildungen mit einem dünnen horizontalen Strich gekennzeichnet, der den oberen Teil jeder Graphik schneidet.

lieferte. Wenn uns für einige Sprachen überhaupt keine Angaben über den Wert von p_{max} zur Verfügung standen z.B. für die afghanischen Stichproben von Ludin {169 - 174}, dann haben wir p_{max} = 0,05 gesetzt und diesen Wert mit einem Fragezeichen versehen.

Die Abbildungen 1 und 2 haben lediglich eine illustrative Aufgabe. Alle Angaben über die Häufigkeitsstruktur des Lexikons, die uns zur Verfügung standen, sind in den Tabellen 1 und 2 enthalten. In Tab. 1a und 1b sind Angaben entweder über vollständige Texte einzelner Werke oder über Abschnitte aus ihnen enthalten. In Tab. 2a und 2b sind Daten von Vereinigungen mehrerer Texte zu einer Stichprobe aufgeführt.

Die Tabellen 1a und 2a enthalten Angaben über die Texte und Stichproben, deren Vokabulare sowie lexikalische Spektra bekannt waren. Dies ermöglichte, nach der Berechnung von Z die theoretische Anzahl der ein-, zwei- und dreimaligen Wörter mit den entsprechenden empirischen Werten zu vergleichen. Diese Größen wurden deswegen gewählt, weil sie, nach dem dargelegten Modell, am meisten vom Stichprobenumfang beeinflußt wurden und insgesamt etwa 3/4 des Wortschatzes bilden.

In den Tabellen 1b und 2b sind lediglich Angaben über Umfänge und Vokabulare einiger Texte und Stichproben zusammengestellt; daher mußten wir uns in diesen Tabellen auf die Berechnung des Zipfschen Umfangs beschränken (die hier mit einer geringeren Genauigkeit durchgeführt wurde; vgl. Anhang II).

Die Analyse der Tabellen 1a und 2a zeigt, daß die Übereinstimmung zwischen den theoretischen Prognosen und den beobachteten Werten in praktisch annehmbaren Grenzen liegt. Für alle Texte und Stichproben, deren Vokabular mehr als 1000 Wörter beträgt, wurde eine signifikante, im Durchschnitt dreiprozentige Unterbesetzung von hapax legomena bei dreimaligen Wörtern beobachtet. Bei zweimaligen Wörtern wurde im Schnitt keine signifikante Verschiebung der Prognose beobachtet.[11] Zwecks einer genaueren a priori Berechnung der Zahl von hapax legomena empfiehlt es sich, den Wert von (11) mit einem empirischen Korrekturkoeffizienten von 0,92 zu multiplizieren; für die dreimaligen Wörter mit 1,08. Nach der Durchführung dieser Korrekturen beträgt die relative

Standardabweichung der empirischen Größe von der Prognose für einmalige Wörter ± 10%, für zweimalige Wörter ± 11% und für dreimalige Wörter ± 15%. Eine Überprüfung mit dem χ^2-Kriterium zeigte, daß die Verteilung der Prognosefehler der Hypothese ihrer Normalität nicht widerspricht.

Diese Abweichungen, die üblicherweise größer als zufällig sind, stimmen gut mit der Streuung überein, die man in gleichlangen Stichproben aus einem Text vorfindet. Stichproben von jeweils 1000 Wortverwendungen aus Puškins "Eugen Onegin" (23 Stichproben), "Ruslan i Ljudmila" (11 Stichproben) und "Skazka o care Saltane" (4 Stichproben), die uns freundlicherweise A.Ju. Šajkevič zur Verfügung stellte, weisen folgende Standardabweichungen von ihren Mittelwerten (in %) auf:

für das Vokabular	± 6,43
für einmalige Wörter	± 9,85
für zweimalige Wörter	± 11,0
für dreimalige Wörter	± 20,8

Alle diese Abweichungen übersteigen die möglichen zufällig. Speziell die Standardabweichung des Vokabulars bei N = 1000 und Z = 240000 laut (14) ist gleich ± 3,6%. Der Chi-Quadrat-Test zeigte, daß die beobachteten Abweichungen eindeutig signifikant sind. Also gibt es in Texten nichtzufällige Schwankungen im Prozeß der Generierung neuer Wörter; ähnliche Erscheinungen kann man in einigen Daten in den Tabellen 3 und 4 beobachten.

In der Tabelle 3 sind Angaben über Texte enthalten, die aus mehr als einer Stichprobe bestehen (der volle Text und Abschnitte davon; verschiedene Teile eines Textes). Die Prognose für das Vokabular einer jeden Stichprobe erfolgt aus den anderen Stichproben desselben Textes. Mit N_1, v_1 und z_1 werden die Parameter der Stichprobe 1, für die man die Prognose macht, gekennzeichnet. Stichprobe 2 ist diejenige, aufgrund derer man die theoretische Prognose aufstellt.

Die Analyse der Abweichungen der empirischen Wortschatzumfänge von der Prognose zeigt keine wesentlichen Verschiebungen: die algebraische Summe aller mittleren Prognosefehler beträgt ungefähr 0,5%. Die relative Standardabweichung ist gleich ± 6,04%,

was mit der oben angegebenen Streuung des Wortschatzes in gleichlangen Abschnitten aus den Werken Puškins (± 6,43%) gut übereinstimmt. Es ist bemerkenswert, daß für die Texte, die nicht in Teile aufgeteilt sind, die Prognose genauer ist, als wenn man aufgrund eines Textteils die Prognose für einen anderen Textteil aufstellt. Während im ersten Fall die Abweichung die theoretische ± 3σ Grenze nicht oder nur geringfügig übersteigt, so findet man im zweiten Fall eine wesentlich größere Abweichung.

Solche signifikanten Abweichungen findet man auch bei Prognosen für die Texte der altgeorgischen Evangelien {6 - 10} (Prognosen aufgrund eines Evangeliums für ein anderes oder für den vollständigen Text), die in der Tabelle 3 nicht aufgeführt wurden. Diese Tatsache bezeugt eine gewisse "statistische Selbstständigkeit" der Teile eines großen Textes. Große Abweichungen gibt es auch bei Vereinigungen vieler Texte über Automobilbau {161 - 162} zu einer Stichprobe.

In den Tabellen 4a und 4b wird das Anwachsen des Vokabulars vom Anfang einiger Texte an im Detail untersucht. Als Ausgangsgröße wurde der Wortschatzumfang der ganzen untersuchten Stichprobe genommen. Es ist interessant, daß bei einigen Texten ("Slovo o polku Igoreve"; "Vojna i mir" und "Voskresenie" von Tolstoj) eine sehr genaue Übereinstimmung mit theoretischen Prognosen besteht (in Extremfällen keine Abweichung über 2-4%), während bei anderen ("Ilja Muromec i Kalin-car'"; "Krejcerova sonata" und "Kazaki" von Tolstoj) eine signifikante Herabsetzung (gekennzeichnet mit Sternchen) des tatsächlichen Wortschatzes beobachtbar ist, die im Vergleich mit dem theoretischen Wortschatz um 12 - 14% kleiner ist.[12)]

Fassen wir die Resultate des Vergleichs des dargelegten Modells mit empirischen Daten zusammen.

1. In dem untersuchten Material gab es keinen Text, dessen Häufigkeitsstruktur von dem Modell abweichen würde; mit anderen Worten, es besteht ein Zusammenhang zwischen dem Vokabularumfang in einem und der Häufigkeitsstruktur in einem beliebigen anderen Punkt des Textes. Die Häufigkeitsstruktur ist dynamisch und hängt von dem Textumfang ab. Der Parameter Z, berechnet aus den Angaben der Stichprobe, reicht sowohl für die Beschreibung des

Vokabulars innerhalb des Textes als auch für die Beschreibung der Häufigkeitsstruktur in einem beliebigen Textabschnitt.

2. Der Zipfsche Umfang Z scheint ein annehmbares Maß des relativen Vokabularreichtums zu sein. Obwohl die Streuung der Schätzungen von Z für unterschiedliche Stichproben aus einem Text auf den ersten Blick ziemlich groß ist (z.B. schwanken die Schätzungen für Puškins "Kapitänstochter" zwischen 33000 und 47000), ist die Streuung der "Kreuzprognosen" aufgrund dieser Schätzungen (Tab. 3) nicht signifikant groß.[13] Als Maß des relativen Vokabularreichtums kann man auch die Größe v(Z) benutzen; in dem Falle wird auch der Einfluß des Wertes von p_{max} berücksichtigt.

3. Die Analyse der Abweichungen empirischer Größen von ihren theoretischen Prognosen führt zu dem Schluß, daß bei zufriedenstellender Übereinstimmung, die im Durchschnitt besteht, einige Abweichungen als nicht zufällig zu betrachten sind. Dies kann entweder durch die Wirkung von Faktoren, die im Modell nicht berücksichtigt wurden und die den in seinem Wesen zufälligen Prozeß überlagern, oder durch die Wirkung irgendwelcher Kontroll- und Regelmechanismen, die den Wortzuwachs im Text mit nicht absoluter Genauigkeit steuern, erklärt werden. Die Argumente für die Existenz eines solchen Mechanismus' werden im letzten Teil dieser Arbeit erörtert.

4. Das dargelegte Modell beruht auf der Hypothese der Texthomogenität (die in reellen Texten nicht erfüllt ist). Aber gerade diese Annahme macht das Modell auch bei der Analyse nichthomogener Stichproben nützlich. Vereinigt man mehrere unähnliche Stichproben mit gleichem Z, so wird in der vereinigten Stichprobe der Wert von Z größer (vgl. die Angaben von Ludin über afghanische Texte {169 - 173}; auch für das "Häufigkeitswörterbuch des Tschechischen", in dem alle Texte vereinigt sind, wird Z größer; vgl. {134} mit {58 - 123}, {135 - 142}). Dies ist nicht der Fall, wenn die vereinigten Texte verhältnismäßig homogen sind. Man kann sich jedoch auch die Situation vorstellen (nicht vorhanden in unserem Material), daß Z in der vereinigten Stichprobe kleiner wird als in den sie konstituierenden Stichproben. Eine solche "übermäßige" Homogenität könnte in dem Fall zustandekommen, wenn die konstituierenden Stichproben nur Varianten eines Textes wären. Diese Eigenschaft macht aus Z einen gewissen Indikator der statistischen Homogenität. Aufgrund ähnlicher Überlegungen kann man z.B. schließen, daß die Unterschiede der Z-Werte in Stichproben, die aus sehr heterogenem Material zusammengestellt wurden (vgl. {156 und 159}), den unterschiedlichen Methoden der Stichprobenerhebung und der Auszählung des Lexikons, die die Autoren dieser Stichproben benutzt haben, zuzuschreiben sind.

3.

Man könnte bei diesen Thesen über die formal beschreibende und "instrumentale" Rolle des dargelegten Modells stehenbleiben, wenn nicht durch die angeführten Zahlen einige neue, teilweise ziemlich unerwartete Gesetzmäßigkeiten hindurchschimmerten.

An erster Stelle fällt die außerordentlich große Streuung der Werte von Z auf: von einigen hundert (in Russischen Bylinen) bis zu 11 Millionen in Joyces "Finegans Wake" {50}. Sogar in Werken eines Autors kann diese Größe beträchtlich schwanken: es reicht, das Z in Puškins "bylinenähnlichen" Texten {1,5} und in seinen Gedichten {47,48} zu vergleichen. Obwohl der Wert von Z = 75000, der in seinem Gesamtwerk festgestellt wurde, in großen Zügen die Häufigkeitsstruktur des Gesamtwerks beschreibt (vgl. die obere Graphik auf der Abb. 2), ist er als eine universelle Charakteristik eines beliebigen Textes Puškins praktisch ungeeignet, da der größte Teil seiner Werke ein von 75000 stark abweichendes Z hat.[14] Auch bei L.N. Tolstoj und J. Joyce ist Z variabel. Daher ist Z weder für die Sprache als ganze, noch für den Autor, sondern nur für den einzelnen Text charakteristisch.

Auf der anderen Seite ist das Material in der Tabelle 1a nach wachsendem Z geordnet.[15] Es ist leicht zu sehen, daß diese Anordnung gleichzeitig, zumindest in groben Zügen, einer chronologischen Anordnung entspricht. Wenn man die gereimte Poesie ausschließt und wenn man nicht das Entstehungsdatum eines Textes,

sondern die "Geschichte der Form", in der der Text geschrieben wurde, in Betracht zieht, dann ist die chronologische Anordnung ziemlich exakt. Das Material in der Tabelle 1a beginnt mit den archaischsten Formen des Typs der russischen Bylinen und endet mit den Werken Šolochovs und Joyces.

Wenn Z als ein konventionelles Maß des relativen Vokabularreichtums betrachtet wird, so kann man schließen, daß sich im Laufe großer Zeiträume eine evidente, wachsende Tendenz zur lexikalischen Konzentration der Texte durchsetzt. Die gereimte Poesie hat die Tendenz die gleichzeitige Prosa zu "überholen". So erscheinen Rustavelis "Der Held im Tigerfell" {27} im Bereich der Prosa von Puškin und Tolstoj und Puškins Verse im Bereich der Prosa von Šolochov und Joyce. Ähnliche Beobachtungen sind schon früher gemacht worden (vgl. z.B. die Arbeiten von Kuraškevič), jedoch nur an beschränktem Material; den diesbezüglichen Verallgemeinerungen fehlte nämlich ein Maß des relativen Vokabularreichtums. Es besteht die Hoffnung, daß mit Zunahme empirischer Daten die erwähnten Beobachtungen zu weiterem Fortschritt in dieser Frage führen werden.

Leider muß man bei diesem Plan auf das Material des Häufigkeitswörterbuchs des Tschechischen (HC) {73 - 82} verzichten, da es nicht bekannt ist, was mit Reim und was in verse libre geschrieben wurde.

Es gibt noch einen Aspekt im Verhalten von Z, der uns höchst interessant erscheint. Da diese Größe dieselbe Dimension wie die Textlänge hat, kann man sie miteinander vergleichen. Auffällig ist, daß in der Mehrzahl der Fälle beide Größen von derselben Größenordnung sind. Im allgemeinen erreichen kurze Texte nicht ihr eigenes Z; lange Texte, deren voller Textumfang Z wesentlich überschreitet, werden von Autoren immer in Teile aufgeteilt (Bände, Bücher u.a.), deren Umfang wiederum von der Größenordnung von Z ist. So beträgt der Umfang von "Krieg und Frieden" etwa 20 Z (N = 472000, Z = 24000), aber der Verfasser hat ihn in 17 Teile aufgeteilt, beide Epiloge eingeschlossen. Ein analoges Bild findet man in der "Auferstehung" (3 Teile), in "Podnjataja celina" (2 Bücher), in "Privalovskie milliony" (5 Teile), in den Romanen von Ju. Smolič (6 Teile), G. Tjutjunnik

(2 Bücher), in "Znamenosci" von A. Gončar (3 Bücher) und im Roman "Chleb i sol'" von Stel'mach (3 Teile).

Dieser offensichtliche Zusammenhang zwischen der Textlänge und dem Zipfschen Umfang brachte uns auf den Gedanken, die Korrelation zwischen x = lg N und y = lg Z zu untersuchen. Die Resultate der Analyse unterschiedlicher Gruppierungen des vorhandenen Materials sind in der Tabelle 5 angegeben.

Die Zeilen I-VI in dieser Tabelle beziehen sich auf die Texte, d.h. als N wurde entweder die volle Textlänge (Zeile I),

Tabelle 5.

Gruppierung des Materials				Korrelationskoeffizient		Lineare Regression $y = ax+b$		Standardabweichung	
		Text- oder Stichprobengruppe	Zahl der Texte oder Stichproben	ρ_{xy}	σ_z	a	b	σ_x	σ_y
Texte	I	$N_o > 10^4$ ohne Daten des HC	27	0,660	0,109	0,833	0,754	0,498	0,630
	II	Idem, mit Berücksichtigung der Teile	28	0,810	0,065	1,255	1,085	0,404	0,622
	III	Künstlerische Prosa des HC (Gruppen A,C,D)	31	0,760	0,076	1,362	-1,310	0,205	0,368
	IV	Vereinigung der Gruppen II und III	59	0,772	0,053	1,160	-0,533	0,337	0,507
	V	Nichtkünstlerische Texte des HC (E,G,H)	14	0,526	0,194	0,378	2,814	0,294	0,211
	VI	$N_o < 10^4$	15	0,653	0,148	0,937	0,940	0,657	1,005
Stichproben	VII	Beliebige Stichproben ohne HC	40	-0,103	0,142	-0,065	5,044	0,663	0,438
	VIII	Beliebige Stichproben aus dem HC	43	0,260	0,142	0,168	3,990	0,617	0,397
	IX	Vereinigung der Stichproben VII und VIII	83	0,031	0,111	0,017	4,641	0,754	0,424

oder die mittlere Länge des Textteils, falls vom Verfasser unterteilt (Zeile II), genommen. Die Zeilen VII-IX beziehen sich auf beliebige Stichproben und als N wurde der tatsächliche Umfang der untersuchten Stichprobe genommen, d.h. entweder der Umfang des Abschnitts oder der Umfang der Vereinigung mehrerer Texte zu einer Stichprobe. Das Material des Häufigkeitswörterbuchs des Tschechischen (HC) (vgl. Jelínek, Bečka, Tešitelová 1961) dient dabei als Kontrollmaterial zu den restlichen Korpora.

Während zwischen der Textlänge und dem Zipfschen Textumfang in allen Fällen offensichtlich eine deutliche Korrelation besteht, gibt es praktisch keine Korrelation zwischen dem Stichprobenumfang und dem zu der Stichprobe gehörenden Z. Am größten ist der Korrelationskoeffizient in den Fällen, wo man die Unterteilung großer Texte vom Verfasser selbst berücksichtigt (Zeile II). Die Texte, die für das "Häufigkeitswörterbuch des Tschechischen" verarbeitet wurden, weisen auch einen hohen Korrelationskoeffizienten auf (Zeile III); der Gesamtkorrelationskoeffizient für alle großen Texte beträgt 0,772 (Zeile IV). Die lineare Regression $y = 1{,}16x - 0{,}533$ liegt in der untersuchten Zone $4 \leq x \leq 5$ relativ nah zur Winkelhalbierenden des Koordinatensystems, was die im Durchschnitt recht große Nähe von Z zum Textumfang bezeugt.[16)]

Bedeutend kleiner ist die Korrelation für nichtkünstlerische Texte des HC (populäre, wissenschaftliche und politische Literatur) und für Texte, die weniger als 10000 Wortverwendungen enthalten (Zeilen V und VI), jedoch übersteigt der Korrelationskoeffizient auch in diesen Fällen den Wert 0,5. Für die nichtkünstlerischen Texte ist der Regressionskoeffizient der linearen Regression wesentlich kleiner als 1, jedoch erlaubt die geringe Anzahl dieser Texte (insgesamt 14) nicht, den Wert von $\underline{a} = 0{,}378$ als endgültig zu betrachten.

Diese Resultate erklären eine früher gemachte Beobachtung, daß große Texte in ihrem vollen Umfang dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen (vgl. Nadarejšvili, Orlov 1971; Orlov 1969, 1970a,b, 1976). Es entsteht der Eindruck, als ob der Zipfsche Umfang eine obere Grenze für die Textlänge wäre, d.h. der Text kann sich bis zu diesem Umfang erstrecken, aber er kann

ihn nicht wesentlich überschreiten. Wenn es nötig ist, diese Grenze zu überschreiten, dann muß der Text in Teile derselben Größenordnung wie Z unterteilt werden.

Was kann denn einen Autor "aufhalten", wenn er den Zipfschen Umfang erreicht hat? Die Suche nach einer Größe, die beim Zipfschen Umfang einen universellen, für alle Texte gemeinsamen Wert annimmt und gleichzeitig linguistisch sinnvoll ist, führte zu dem Begriff der differentiellen Geschwindigkeit des Vokabularwachstums.

Seien vom Anfang des Textes N Wörter (tokens) geschrieben (oder beobachtet), unter denen es v unterschiedliche Wörter (types) gibt. Bei den nächsten ΔN tokens soll das Vokabular um Δv types anwachsen. Die Größe $\frac{\Delta N}{N}$ bezeichnen wir als den relativen Textzuwachs und $\frac{\Delta v}{v}$ als den relativen Vokabularzuwachs. Das Verhältnis $\frac{\Delta v}{v} / \frac{\Delta N}{N}$ bei kleinem ΔN werden wir die differentielle Geschwindigkeit des Vokabularwachstums nennen (DGVW).

Da wir über die stetige theoretische Vokabularwachstumskurve (10b) verfügen, ist es zweckmäßig, die DGVW durch Differenzierung abzuleiten. Es ist

$$L(X) = \lim \frac{N}{v} \cdot \frac{\Delta v}{\Delta N} = \frac{N}{v} \frac{dv(N,Z)}{dN} = \frac{1}{X-1} \cdot \frac{1}{\ln X} \qquad (17)$$

wo $X = Z/N$.

Es ist interessant, daß diese Größe offensichtlich weder von N, noch von Z, noch von p_{max}, sondern ausschließlich von $X = Z/N$ abhängt. Daraus folgt: Wenn man die Textlänge in Einheiten von der Größe Z ausdrückt, dann erweist sich die Kurve der DGVW als eine universelle, für alle Texte gemeinsame Größe. Wenn sich der Textumfang dem Zipfschen Umfang nähert, d.h. wenn $X \to 1$, dann $L(X) \to 0{,}5$. Der allgemeine Verlauf dieser Funktion ist auf der Abb. 3 dargestellt (vgl. die obere Graphik).

Direkt unter dieser Kurve sind in demselben horizontalen Maßstab die Histogramme der Verteilung der Beziehung $X_o = Z/N_o$ für die untersuchten Texte aufgetragen. Hier ist N_o entweder die volle Textlänge oder die mittlere Textlänge, wenn ein großer Text vom Autor unterteilt wurde. Zwecks Vergleichs wurden darunter die Histogramme der Beziehung $X = Z/N$ für beliebige Stichproben (Ab-

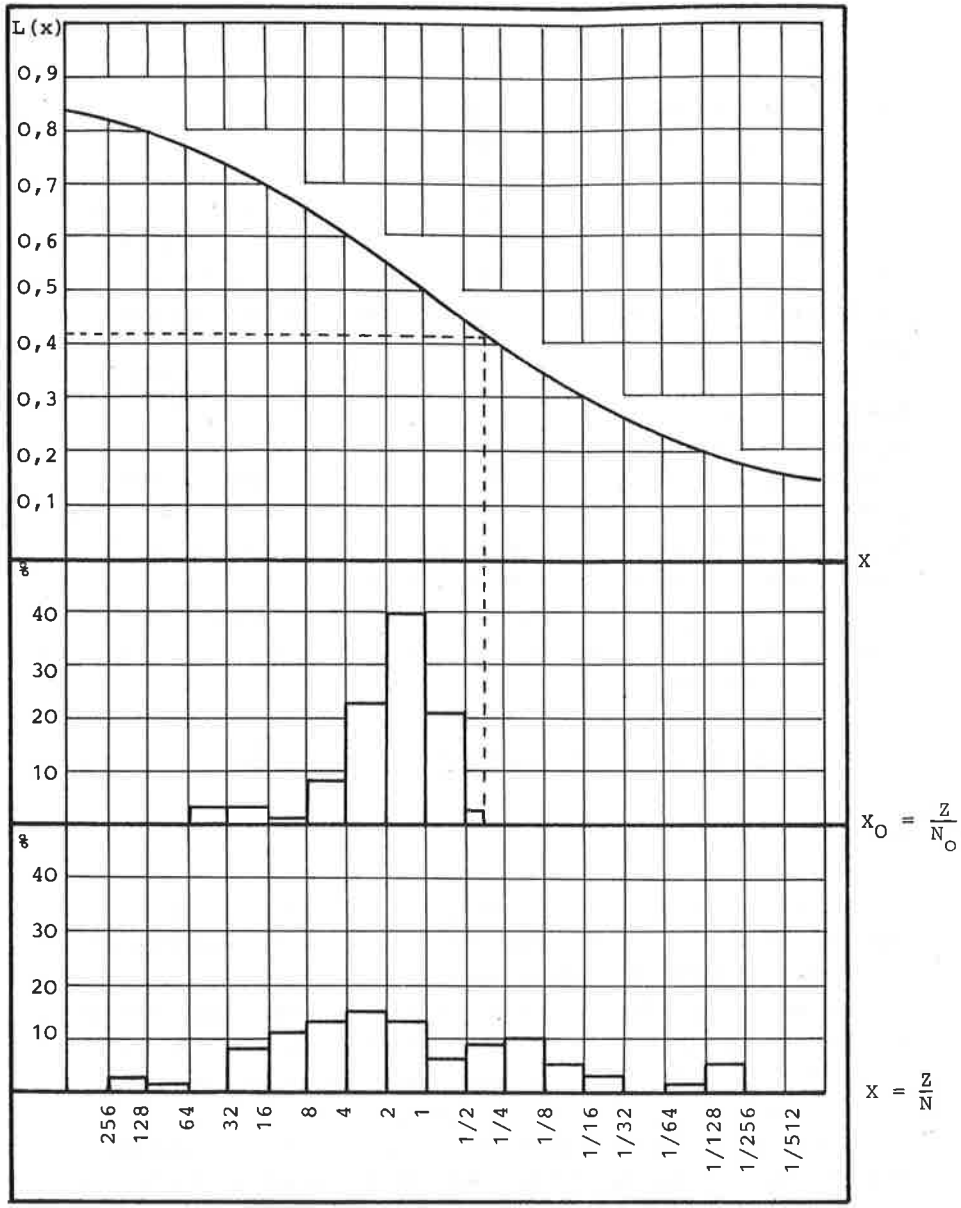


Abb. 3. Graphik der Funktion der DGVW L(X) und die Verteilungen der Größen $X_o = Z/N_o$, wo N_o die volle Textlänge (oder die Länge eines vom Autor bestimmten Teils eines großen Textes) ist, und $X = Z/N$, wo N die Länge einer beliebigen Stichprobe (Abschnitt eines Textes oder Vereinigung mehrerer Texte in einer Stichprobe) ist. Die unterbrochene Linie bezeichnet den kleinsten beobachteten Wert von X_o und $L(X_o)$ für einzelne Texte.

schnitte oder Vereinigungen mehrerer Texte) gezeichnet, und zwar in solchen Fällen, wo der Stichprobenumfang durch Zufall oder durch die Willkür der Forscher festgelegt wurde (z.B. Gesamtwerke von Puškin).

Man kann leicht sehen, während die Verteilung von X_o ein deutliches Maximum in der Nähe von X = 1 (d.h. bei N = Z) hat, besitzt die Verteilung von X kein so deutliches Maximum und ist außerdem beträchtlich auseinandergezogen. 40% aller untersuchten Texte haben eine Länge von 0,5Z bis Z, 62% der Texte von 0,5Z bis 2Z und 84% der Texte von 0,25Z bis 2Z; die Grenzen von 0,25Z bis 2Z entsprechen einer DGVW am Ende des Textes zwischen 0,62 und 0,44. Der niedrigste beobachtete Wert L = 0,414 befindet sich im altgeorgischen Text des "Johannes-Evangeliums". Die DGVW scheint also ein recht begrenzter Parameter zu sein.

Wir werden L(1) = 0,5 als den "nominellen" Wert der DGVW, der die Notwendigkeit der "Textbeendung" signalisiert, betrachten, und wir analysieren die Texte, deren Umfang Z überschreitet. Der mittlere Wert X_o für diese Texte (in unserem Material gibt es 17 dieser Art) ist ungefähr 0,71. Dies entspricht der DGVW L(0,71) = 0,46. Der "nominelle" Wert 0,5 wird also im Schnitt um 8% reduziert, und die größte beobachtete Reduktion ist 17%, wenn eine Reduktion überhaupt stattfindet. Diese Werte liegen in den Grenzen der gewöhnlichen menschlichen Empfindlichkeit für Veränderungen irgendwelcher physikalischer Größen (das Weber-Fechnersche Gesetz).[17)]

Daher bietet sich die folgende Hypothese an: bei der Erzeugung oder der Wahrnehmung eines Textes in dieser oder jener Form verwirklicht sich (offenbar unbewußt) eine Kontrolle der DGVW, und die Stelle, wo die Textlänge anfängt mehr als doppelt so schnell zu wachsen wie das Vokabular, wird als Vollendung, Abgeschlossenheit des Textes erlebt.

Ein weiteres Zurückbleiben des Vokabularwachstums hinter dem Textwachstum wird als "Langatmigkeit" des Textes empfunden usw. Außer durch die dargelegten statistischen Erläuterungen kann diese Hypothese auch durch die Beobachtung unterstützt werden, daß bei großen Schriftstellern die volle Textlänge nah an Z liegt, z.B. bei Š. Rustaveli {27}, L.N. Tolstoj {19, 33}, K. Čapek {70},

in der Byline "Vol'ga i Mikula" {3}. Besonders interessant ist die praktisch exakte Übereinstimmung der Textlänge mit dem Zipfschen Umfang in Joyces "Stephen Hero" {53}.

Wie bekannt (vgl. Žantieva 1967) hat dieser frühe Roman von Joyce ursprünglich einen größeren Umfang gehabt, aber der Verfasser hat einen Teil verbrannt. Offensichtlich, wenn der Rest mit Z übereinstimmen sollte, dann müßte der ursprüngliche Umfang Z überstiegen haben. War es vielleicht eben dieser Umstand, der als "Langatmigkeit" empfunden wurde und Joyce beunruhigte? Man kann nicht ausschließen, daß die "Kollisionen mit dem Zipfschen Umfang" für die Autoren eine dramatische Rolle spielen können.

Interessant ist beispielsweise K.G. Paustovskijs Aussage (erwähnt von A. Ionov in "Ogonek" Nr. 22, 1972), daß er nicht imstande ist, ein Werk länger als 11 Druckbogen zu schreiben: "Es ist direkt eine verhexte Zahl: elf, was du auch immer willst! Wenn ich auf den zwölften übergehe, kann ich keine einzige Zeile mehr schreiben". Elf Druckbögen enthalten etwa 50000 Wortverwendungen und in der Tat haben die meisten großen Werke Paustovskijs ungefähr diesen Umfang (das größte Werk "Dalekie gody" hat 14,3 Druckbögen. Nach den Angaben von L. Sudovičene (Učennye zapiski vyss. uč. zav. Lit. SSR. Jazykoznanie 22/2, Vilnjus 1971) hat der Roman "Dym otečestva" den Umfang von N_o = 57000 Wortverwendungen, den Wortschatz von v = 7829 Wörtern und p_1 = 0,0337 (12,35 Druckbögen). Der Wert Z, berechnet aus diesen Angaben, beträgt 65400, d.h. er liegt sehr nah bei der tatsächlichen Romanlänge. Es wäre interessant, die lexikalische Konzentration in anderen Texten von Paustovskij zu untersuchen.

Der parallele Zuwachs des relativen Vokabularreichtums und der maximal "zugelassenen" Umfänge ruft offensichtlich eine Krisensituation hervor. Auf jeden Fall kann man die in der Literaturwissenschaft umstrittene Lebensfähigkeit der Romanform in Begriffen der vorliegenden Arbeit vollständig interpretieren. Der Mensch kann offenbar lexikalisch gesättigte Texte großen Umfangs schwer bewältigen und dieses Gefühl der Erschöpfung der Aufnahmefähigkeit der Leser beunruhigt die Schriftsteller. Anscheinend führte die Entwicklung des Romangenres zur Entstehung

der "literarischen Dinosaurier" ... Künftige Untersuchungen werden zeigen, ob diese Behauptung stimmt.

Zusammenfassend kann man bemerken, daß das dargelegte Modell einige unerwartete Erscheinungen zu beschreiben erlaubt:

(a) das Anwachsen des relativen Vokabularreichtums im Verlauf großer Zeiträume;
(b) das vorauseilende Anwachsen des relativen Vokabularreichtums in der gereimten Poesie;
(c) die Korrelation zwischen der Länge und dem Zipfschen Umfang des Texts;
(d) die Universalität der Funktion der DGVW, wenn die Textlänge in relativierten Einheiten von Z ausgedrückt ist;
(e) die Unzulässigkeit einer wesentlichen Unterschreitung der 0,5-Grenze durch die DGVW (der kleinste beobachtete Wert in zusammenhängendem Text war 0,414).

Offensichtlich hängen (c) und (e) eng zusammen und das eine bedingt das andere. Es ist schwer zu sagen, welche von ihnen primär ist, die Hypothese, daß (e) primär ist, ist wahrscheinlicher. In jedem Fall führt uns der Zusammenhang zwischen N und Z zu literaturwissenschaftlichen Problemen, zu Problemen der Psychologie künstlerischer Kreativität (und zur Psychologie der künstlerischen Rezeption, da man sich schwer vorstellen kann, daß die Autoren diesen Zusammenhang nur zum "eigenen Vergnügen" verwirklichen) und, letzten Endes, zu tieferen Problemen der Organisation und Rezeption von Information durch den Menschen. Es ist offensichtlich, daß dieser Problemkomplex eine weitere vertiefte Untersuchung durch viele Spezialisten, darunter Mathematiker, Linguisten und Psychologen, benötigt.

An dieser Stelle möchte ich meine tiefe Dankbarkeit an A.N. Kolmogorov ausdrücken, dessen Interesse diese Arbeit stark angeregt hatte. Ich bedanke mich herzlich bei V.M. Andrjušenko, N.P. Darčuk, I.Š. Nadarejšvili, T.I. Pataraja und A.Ja. Šajkevič, die mir eine Menge linguistischer Zählungen zur Verfügung stellten.

TABELLE 1A.

	Text	Quelle	Volle Textlänge N_0	Stichprobenumfang N	p_{max}	Wortschatz V	$R=\frac{V}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
1	Puškin, Iz byta povolžskych razbojnikov /Aus dem Leben der Wolga-Räuber/	Zählung von I.S. Nadarejšvili	278	278	0,040	112	6,72	400
2	Byline "Brat'ja-razbojniki i sestra" /Räuberbrüder und die Schwester/	- " -	293	293	0,044	116	6,77	486
3	Byline "Vol'ga i Mikula" /Wolga und Mikula/	- " -	930	930	0,040	254	8,32	960
4	Byline "Il'ja Muromec i Kalincar" /Ilja Muromec und Zar Kalin/	- " -	3380	3380	0,059	550	9,46	2130
5	Puškin, Skazka o rybake i rybke /Das Märchen vom Fischer und Fischlein/	Mitteilung von A.Ja. Šajkevič	948	948	0,038	318	10,3	2200
6	Altgeorgisches Evangelium (voller Text)	Imnajšvili (1948)	53043	53043	0,120	2672	11,6	8500
7	Altgeorgisches Johannes Evangelium	- " -	11908	11908	0,084	1146	10,5	4156
8	Altgeorgisches Matthäus Evangelium	- " -	14317	14317	0,115	1708	14,3	10000
9	Altgeorgisches Markus Evangelium	- " -	9528	9528	0,142	1400	14,3	11300

Theoretischer Wortschatz $v(N,Z)$	$x_0=\frac{Z}{N_0}$	Einmalige Wörter beobachtet ν_1	Einmalige Wörter theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter beobachtet $\hat{v}_2$	Zweimalige Wörter theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter beobachtet $\hat{v}_3$	Dreimalige Wörter theoretisch $v_3(N,Z)$	$\frac{\nu_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
113	1,44	39	60	0,65	36	19	2,00	9	10	0,90
115	1,66	52	61	0,85	27	19	1,40	8	9	0,89
253	1,03	116	127	0,91	56	42	1,33	22	21	1,05
547	0,63	245	258	0,95	82	85	0,97	42	40	1,05
316	2,32	192	178	1,08	45	52	0,87	21	23	0,91
2675	0,16	769	947	0,81	416	400	1,04	242	234	1,03
1147	0,35	432	475	0,91	184	184	1,00	98	99	0,99
1690	0,70	655	789	0,83	277	273	1,01	145	145	1,00
1403	1,22	587	722	0,80	244	234	1,04	133	115	1,16

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R=\frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
10	Altgeorgisches Lukas Evangelium	Imnajšvili (1948)	17290	17290	0,136	1932	14,7	12000
11	"Čtenie o Borise i Glebe"	Vjalkina & Lukina	7380	7380	0,077	1249	14,5	9000
12	"Žitie Feodosija Pečerskogo"	- " -	18754	18754	0,081	2164	16,4	12000
13	"Skazanie o Borise i Glebe"	- " -	8590	8590	0,107	1611	17,4	18000
14	"Slovo o Polku Igoreve"	Zählung von A.J. Pataraja	2772	2772	0,032	893	16,9	12500
15	Puškin, Skazka o care Saltane /Das Märchen von Zar Salton/	Mitteilung von A.Ja. Šajkevič	~4000	1000	0,041	446	14,1	13500
16	Kumsiašvili, Die Portweinproduktion in Georgien (in Georgisch)	Zählung von I.Š. Nadarejšvili	9998	9998	0,051	1906	19,1	17000
17	- " -	- " -	9998	6000	0,056	1484	19,2	20000
18	- " -	- " -	9998	3000	0,068	943	17,2	22400
19	Tolstoj, Krejcerova sonata /Kreutzersonate/	- " -	~25200	10000	0,047	2083	20,8	22000
20	Tolstoj, Vojna i mir /Krieg und Frieden/	- " -	~472000	10000	0,045	2146	21,5	24000
21	Unsuri, Divan (in Arabisch)	Osmanov (1970)	46472	46472	0,048	4824	22,4	26400
22	Puškin, Skazka o zolotom petuške /Das Märchen vom goldenen Hahn/	Mitteilung von A.Ja. Šajkevič	902	902	0,037	457	15,2	26000

Theoretischer Wortschatz $V(N,Z)$	$x_o=\frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
1930	0,69	813	908	0,90	301	321	0,92	177	166	0,92
1247	1,22	607	643	0,94	199	201	0,99	114	102	1,12
2165	0,64	908	992	0,92	296	352	0,84	184	183	1,00
1605	2,09	855	845	1,01	257	310	0,83	123	78	1,42
890	4,50	516	555	0,93	138	137	1,01	66	59	1,12
446	3,37	293	297	0,99	67	62	1,09	31	24	1,31
1903	1,70	817	1035	0,79	295	318	0,93	160	122	1,31
1488	-	706	890	0,79	239	236	1,01	135	138	1,25
948	-	518	623	0,83	142	140	1,01	98	57	1,72
2080	0,87	1169	1173	0,99	349	340	1,03	157	153	1,03
2145	0,051	1181	1230	0,96	393	350	1,12	150	155	0,97
4830	0,57	2268	2200	1,03	653	798	0,82	397	421	0,94
457	28,8	335	337	0,99	51	56	0,91	24	21	1,14

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichproben-Umfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
23	Saint-Trond, Gesta abbatum trudoniensium I-VII	Tombeur (1945)	26275	26275	0,038	3957	24,4	30000
24	Puškin, Skazka o pope i rabotnike ego Balde /Das Märchen vom Popen und seinem Arbeiter Bald/	Mitteilung von A.Ja. Šajkevič	910	910	0,034	477	15,8	35000
25	Puškin, Pikovaja dama /Pique Dame/	Zählung von I.S. Nadarejšvili	6861	6861	0,038	1928	23,3	35000
26	"Merilo pravednoe"	Vjalkina & Lukina (1964)	31262	31262	0,062	4311	24,4	36000
27	Rustaveli, Der Held im Tigerfell (in Georgisch)	Osmanov (1980)	42120	42120	0,021	5965	29,0	38000
28	Puškin, Kapitanskaja dočka /Die Kapitänstochter/	Josselson (1953)	29345	29345	0,039	4783	27,9	45000
29	- " -	- " -	29345	10000	0,036	2432	34,3	33400
30	- " -	- " -	29345	5000	0,040	1671	23,6	47000
31	- " -	- " -	29345	5000	0,044	1567	22,2	37300
32	Mamin-Sibirjak, Privalovskie milliony /Die Priwalowschen Millionen/	Genkel' (1966)	127927	127927	0,023	10063	28,2	40000
33	Tolstoj, Kazaki /Die Kosaken/	Zählung von I.S. Nadarejšvili	~48100	10000	0,056	2582	25,8	53000

Theoretischer Wortschatz $v(N,Z)$	$x_o = \frac{Z}{N_o}$	Einmalige Wörter			Zweimalige Wörter			Dreimalige Wörter		
		beobachtet $\hat{v}_1$	theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	beobachtet $\hat{v}_2$	theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	beobachtet $\hat{v}_3$	theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
3980	1,14	1775	2040	0,87	705	666	1,06	367	328	1,12
481	38,4	372	336	1,11	42	57	0,74	20	16	1,24
1930	5,08	1146	1218	0,94	335	296	1,12	129	124	1,04
4330	1,15	2147	2220	0,97	706	723	0,98	334	357	0,94
5960	0,90	2995	2940	1,02	937	996	0,94	392	504	0,78
4780	1,54	2384	2575	0,93	847	802	1,06	433	385	1,13
2430	-	1477	1460	1,01	404	382	1,06	167	172	0,97
1662	-	1133	1120	1,01	236	238	0,99	104	96	1,08
1575	-	927	1033	0,90	281	232	1,21	127	95	1,34
10030	0,31	5093	4020	1,27	-	-	-	-	-	-
2570	1,10	1487	1030	0,91	419	391	1,07	190	165	1,15

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R=\frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
34	Tolstoj, Voskresenie /Die Auferstehung/	Zählung von I.S. Nadarejšvili	~145000	10000	0,057	2587	25,9	53000
35	Richards, Arifmetičeskie operacii na vyčislitelnych masinach /Arithmetische Operationen auf dem Computer/	Borodin & Mater (1967)	60803	29358	0,045	5075	29,7	57000
36	Appolinaire, Calligrammes		9865	9865	0,026	2941	29,8	60000
37	Dovženko, Poema o more (in Ukrainisch) /Gedicht über das Meer/	Mitteilung von N.L. Darčuk	25837	20000	0,037	4187	29,6	57000
38	Smolić, Mir chižinam, vojna dvorcam (in Ukrainisch) /Friede den Hütten, Krieg den Palästen/	- " -	154682	20000	0,033	4906	34,6	94000
39	Tjutjunnik, Vir (in Ukrainisch)	- " -	171860	20000	0,033	4735	33,5	85000
40	Gončar, Mikita Bratus' (in Ukrainisch)	- " -	15078	15078	0,027	3777	30,8	60000
41	Gončar, Znamenosci (in Ukrainisch)	- " -	121055	20000	0,031	4960	35,0	100000
42	Stel'mach, Chleb i sol' (in Ukrainisch) /Brot und Salz/	- " -	194850	20000	0,039	5116	36,1	120000

Theoretischer Wortschatz $v(N,Z)$	$x_o=\frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
2570	0,34	1934	1630	0,94	408	391	1,04	175	165	1,08
5110	0,94	2500	2830	0,88	861	846	1,02	301	377	1,04
2900	6,1	1676	1820	0,92	476	446	1,07	221	179	1,23
4175	2,20	2420	2455	0,99	683	692	0,98	314	306	1,03
4890	0,61	3054	3060	1,00	773	754	1,03	332	324	1,02
4760	0,49	2953	2930	1,01	727	746	0,97	344	318	1,08
3760	4,00	2483	2300	1,08	608	588	1,03	238	253	0,94
4980	0,82	3097	3160	0,98	737	760	0,97	360	362	0,99
5100	0,62	3221	3260	0,99	818	766	1,07	320	320	1,00

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R=\frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
43	Gončar, Krov' ljudskaja - ne vodica (in Ukrainisch) /Menschliches Blut ist kein Wasser/	Mitteilung von N.L. Darčuk	72272	20000	0,038	4791	33,9	95000
44	Šolochov, Podnjataja celina Bd. 1 /Neubruch/	Ljatina (1968)	87883	87883	0,032	12778	43,1	130000
45	Šolochov, Podnjataja celina. Bd. 2 /Neubruch/	- " -	106167	106167	0,038	12762	39,2	106167
46	Šolochov, Podnjataja celina. Gesamttext/Neubruch/	- " -	194035	194035	0,035	18508	41,9	127000
47	Puškin, Ruslan i Ljudmila	Mitteilung von A.Ja. Šajkevič	~11000	1000	0,045	579	18,3	178000
48	Puškin, Evgenij Onegin	- " -	~23000	1000	0,046	590	18,6	240000
49	Joyce, Ulysses	Hart (1963)	260430	260430	~0,07	29899	58,7	360000
50.	Joyce, Finnegans Wake	- " -	218077	218077	~0,07	63924	137	$1{,}1\cdot10^7$

Theoretischer Wortschatz $v(N,Z)$	$x_o=\frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
4810	1,23	2890	3000	0,96	777	747	1,04	339	316	1,07
12760	1,48	6326	6820	0,93	2212	2040	1,07	1138	947	1,20
12780	1,00	6010	6390	0,94	2390	2130	1,12	1137	1065	1,07
18550	0,65	8048	8610	0,94	3066	3030	1,01	1688	1400	1,20
580	16,2	438	472	0,93	81	53	1,53	23	18	1,28
592	10,4	470	508	0,93	63	52	1,21	22	16	1,37
30100	1,38	16432	16100	1,02	-	-	-	-	-	-
64000	50,5	51922	49200	1,05	-	-	-	-	-	-

Zusatz zu Tabelle 1A

	Text	Quelle	volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$
1	2	3	4	5	6	7	8
10a	Beaumarchais,"Figaros Hochzeit"	Musso (1972)	23829	23829	0,060	2367	15,33
15a	Dante,"Das neue Leben"	Alinei (1971)	14392	14392	0,045	2275	18,96
15b	Cervantes, "Don Quijote" (2 Teile, 99 Kapitel)	Gomez (1962)	357255	357255	0,061	7872	18,17
18a	Shakespeare, "König Lear"	Kvaracchelija (1966)	25471	25471	0,039	3391	21,25
19a	Seneca, "De Clementia"	Seneque (1968a)	8218	8218	0,038	1952	21,53
22a	Povest' vremennych let /Die Erzählung von Obergangsjahren/	Tvorogov (1967)	47371	47371	0,081	4684	21,52
23a	Seneca, "De Brevitate vitae"	Seneque (1968b)	6115	6115	0,032	1811	23,16
32a	Pisarev, "Realisty /Die Realisten/	Bulachov(1969)	48354	48354	0,054	6348	28,87
34a	Griboedov, "Gore ot uma" /Verstand schafft Leiden/	Kunickij (1896)	13246	13246	0,032	3343	29,05
37a	Paustovskij, "Dym otečestva /Rauch des Vaterlandes/	Sudavičene (1971)	57000	57000	0,034	7829	32,79
37b	Mickiewicz, "Herr Tadeusz, I-XII	Sambor (1969)	64510	64510	0,032	9250	36,42
37c	-"-, Buch I	-"-	64510	6587	0,037	2257	27,81
37d	-"-, Bücher I-VI	-"-	64510	34280	0,032	6823	36,85
40a	Gladkov, "Povest' o detstve /Erzählung von der Kindheit/	Sinenko (1973)	127917	127917	0,060	12821	35,85

Zipfscher Umfang Z	Theoretischer Wortschatz v(N,Z)	$\chi_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
9	10	11	12	13	14	15	16	17	18	19	20
10340	2356	0,43	1095	1039	1,05	381	413	0,94	214	256	0,84
15230	2267	1,06	1278	1111	1,15	361	370	0,98	157	185	0,84
17300	7905	0,05	2730	2209	1,24	1099	1016	1,08	634	632	1,00
20170	3392	0,79	1933	1630	1,19	483	565	0,85	226	312	0,72
23000	1945	2,80	995	1137	0,88	371	313	1,19	167	140	1,19
28375	4685	0,60	2187	2143	1,02	730	774	0,94	397	407	0,98
32510	1808	5,32	1052	1145	0,92	303	276	1,10	155	116	1,34
52500	6328	1,09	3092	3077	1,00	1058	1012	1,04	546	502	1,09
56146	3342	4,24	2079	2060	1,01	526	521	1,01	210	225	0,93
63800	7845	1,12	3798	3997	0,95	1277	1307	0,98	705	646	1,09
82530	9241	1,28	4360	4810	0,91	1595	1537	1,04	797	855	0,93
84050	2259	-	1460	1564	0,93	331	311	1,06	137	121	1,13
95000	6815	-	3648	3977	0,92	1143	1098	1,04	537	495	1,08
98000	12862	0,77	5636	6146	0,91	2087	2139	0,98	1211	1099	1,10

Zusatz zu Tabelle 1a (Fortsetzung)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$
1	2	3	4	5	6	7	8
43a	Dante, "Die göttliche Komödie"	Alinei (1971)	101554	101554	0,039	13004	40,81
43b	Dante, "Die göttliche Komödie, Hölle"	- " -	101554	34126	0,041	6579	35,61
43c	Dante, "Die göttliche Komödie, Fegefeuer"	- " -	101554	34042	0,039	6450	34,96
43d	Dante, "Die göttliche Komödie, Paradies"	- " -	101554	33386	0,039	6273	34,33
46a	Byron, "Don Juan"	Byron (1967)	130745	130745	0,046	14411	39,84

Anmerkung: Der Zusatz zur Tabelle 1a wurde nach der Fertigstellung der deutschen Übersetzung erstellt und konnte nicht in die Tabelle 1 eingeordnet werden. Die Zahlen in der ersten Spalte zeigen, an welche Stelle in der Tabelle 1a die Daten einzuordnen sind.

Zipfscher Umfang Z	Theoretischer Wortschatz $v(N,Z)$	$\chi_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
9	10	11	12	13	14	15	16	17	18	19	20
118000	12966	1,16	7487	6645	1,13	1966	2159	0,91	891	1063	0,84
94000	6582	-	4100	3838	1,07	995	1061	0,94	439	478	0,92
86850	6447	-	4009	3720	1,08	937	1044	0,90	399	474	0,84
82200	6269	-	3939	3599	1,09	899	1017	0,88	407	464	0,88
120200	14434	0,92	7250	7512	0,97	2405	2504	0,96	1194	1252	0,95

TABELLE 1B

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
51	Byline "Avdot'ja-ženka rjazanočka"	Zählung von I.S. Nadarejšvili	524	524	0,034	163	7,13	415	0,79
52	La Rochefoucaud, Maximes	Mitteilung von A.Ja. Šajkevič	13500	13500	0,066	1893	16,3	12000	0,96
53	Joyce, Stephen Hero (???)	Hart (1963)	74459	74459	~0,07	8740	31,9	~74500	~1
54	Richards, Areifmeticeskie operacii na vyčislitel'nych mašinach /Arithmetische Operationen auf dem Computer/	Borodin & Mater (1967)	60803	60803	0,045	8978	36,4	~90000	~1,48
55	Macaulay, "Essay on Bacon"	Frumkina (1964)	?	2046	~0,07	964	21,4	~150000	-
56	- " -	- " -	?	4049	~0,07	1432	22,5	~ 65000	-
57	- " -	- " -	?	6045	~0,07	2048	26,4	~100000	-
58	V. Vančura, "Konec starých časů" 7-186	HC, Gruppe A künstlerische Prosa	?	30281	~0,047	5469	31,4	~ 60000	-

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
59	K. Horký, "Piskání v lese" (2)	HC, Gruppe A, künstlerische Prosa	28028	28028	~0,047	6111	36,5	~ 90000	~ 3,22
60	J. Fučík, "Reportáz psana na oprátce" (3)	" - "	21640	21640	-"-	5006	34,0	~ 90000	~ 4,15
61	K. Čapek, "Zivot a dílo skladatele Foltýna (4)	" - "	21963	21963	-"-	4145	27,9	~ 45000	~ 2,06
62	J. Morávek, "Srdce na zâmek" 5-171 (5)	" - "	?	35187	-"-	6927	36,9	~100000	-
63	I. Olbracht, "Bratr Žak" (6)	" - "	29803	29803	-"-	5559	32,2	~ 67000	~ 2,25
64	K. Nový, Rytíři a lapkově" 5-185 (7)	" - "	?	33774	-"-	6265	34,1	~ 80000	-
65	J. Spáčil, "Náš svět zemře s námi" (8)	" - "	23436	23436	-"-	4582	29,9	~ 57000	~ 2,43
66	J. Marek, "Vesnice pod zemí" 9-160 (9)	" - "	?	31195	-"-	4188	23,7	~ 48000	-
67	E. Bass, "Lidé z maringotek" (11)	" - "	47542	47542	-"-	8763	39,8	~120000	~ 2,53

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
68	J. Hora, "Hladový rok" (12)	HC, Gruppe A, künstlerische Prosa	35273	35273	~0,047	6939	36,9	~100000	~ 2,88
69	M. Pujmanová, "Predtucha" (13)	- " -	24353	24353	-"-	4858	31,1	~ 63000	~ 2,59
70	K. Čapek, "Obyčejný zivot" (14)	- " -	39360	39360	-"-	5539	27,9	~ 40000	~ 1,02
71	J. Marânek, "Barbar Vok" (15)	- " -	30145	30145	-"-	6498	37,4	~120000	~ 3,98
72	V. Řezáč, "Černě svetlo" (16)	- " -	55164	55164	-"-	8675	36,9	~ 95000	~ 1,73
73	Vl. Holan, "Havranîm brkem"(31)	HC, Gruppe B, Poesie	?	12153	-"-	3855	35,0	~170000	-
74	J. Seifert, "Jaro sbohem" (32)	- " -	?	6171	-"-	2095	25,5	~ 55000	-
75	S.K. Neumann, "Zamořená leta" (33)	- " -	?	6658	-"-	2435	29,8	~130000	-
76	J. Hora, "Knîha domova" (34)	- " -	?	10900	-"-	2961	28,4	~ 60000	-
77	Er. Halas, "Ladenî" (35)	- " -	?	5249	-"-	2078	28,7	~150000	-
78	J. Hořec, "Květen Č. I" (36)	- " -	?	3374	-"-	1204	20,7	~ 35000	-

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
79	V. Nezval, "Historický obraz" I-III (37)	HC, Gruppe B, Poesie	?	5340	~0,047	1825	25,0	~ 60000	-
80	P. Křička, "Svetlý oblak" (38)	- " -	?	4782	-"-	1689	24,4	~ 60000	-
81	J. Hiršal, "Student nebe" (39)	- " -	?	3614	-"-	1066	17,7	~ 18000	-
82	Fr. Hrubîn, "Krásno po chudobě" (40)	- " -	?	2240	-"-	914	19,3	~ 36000	-
83	M. Majerová, "Robinsonka" (41)	HC, Gruppe C, Jugendliteratur	37509	37509	~0,043	7420	38,3	~120000	~ 3,20
84	J. Prchal, "Bilý jestrâb" (42)	- " -	13268	-"-	-"-	2761	24,0	~ 30000	~ 2.44
85	J. Pilař, "Slunečnî vůz" (43)		18400	-"-	-"-	4164	30,7	~ 60000	~ 3,26
86	J. John, "Narodil se.." (44)	- " -	24779	-"-	-"-	6286	39,3	~150000	~ 6,07
87	Fr. Langer, "Deti a dýka" (45)	- " -	34192	34192	-"-	5599	30,3	~ 55000	~ 1,61
88	J.V. Pleva, "Budîk" (46)	- " -	24249	24249	-"-	3120	20,0	~ 17000	~ 0,70

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
89	B. Říha, "Na útěku" (47)	HC, Gruppe C, Jugendliteratur	30072	30072	~0,043	4562	26,4	~ 37000	~ 1,24
90	E. Štorch, "Meč proti meči" (48)	- " -	52031	52031	-"-	7350	32,2	~ 70000	~ 1,34
91	V. Deyl, "Vyzvědačí"(49)	- " -	31079	31079	-"-	5680	32,2	~ 67000	~ 2,16
92	J. Kopta, "Přivor pod ořechy" (50)	- " -	32284	32384	-"-	6011	33,4	~ 75000	~ 2,39
93	M. Kratochvil, "České jaro" (51)	HC, Gruppe D,	12384	12384	~0,031	2156	19,4	~ 15000	~ 1,21
94	Er. Tetauer, "Krivdu napravovati" (52)	- " -	16250	16250	-"-	2324	18,1	~ 13000	~ 0,79
95	J. a M. Tomanovi,"Vinice" (53)	- " -	12444	12444	-"-	1999	17,9	~ 12000	~ 1,09
96	O. Scheinpflugová, "Guayana" (54)	- " -	14694	14694	-"-	2417	19,9	~ 16000	~ 1,09
97	Er. Gotz, "Soupeři" (55)	- " -	15400	15400	-"-	2817	22,6	~ 24000	~ 1,56
98	St. Lom, "Karel IV" (56)	- " -	18683	18683	-"-	3840	28,1	~ 50000	~ 2,68
99	L. Suchý, "Hrstka věrných" (57)	- " -	8910	8910	-"-	1790	18,9	~ 15000	~ 1,68

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
100	J. Klíma, "Na dosah ruky" (58)	HC, Gruppe D,	12852	12852	~0,031	1480	13,0	~ 5000	~ 0,39
101	J. Drda, "Hrátky s čertem" (59)	- " -	13908	13908	-"-	2899	24,6	~ 30000	~ 2,16
102	Fr. Langer, "Jiskra v popelu" (60)	- " -	14418	14418	-"-	2661	22,1	~ 22000	~ 1,52
103	V. Příhoda, "Idem na školu II stupne" (61)	HC, Gruppe E,	24658	24658	~0,038	4308	27,5	~ 47000	~ 1,91
104	B. Václavek, "Lidová slovesnost" (62)	- " -	9290	9290	-"-	2360	24,5	~ 36000	~ 3,88
105	A. Piša, "Poesie nové doby" (63)	- " -	23802	13802	-"-	3381	21,9	~ 22000	~ 0,92
106	"Ústava Československé republiky" (65)	- " -	8374	8374	-"-	1337	14,6	~ 6000	~ 0,72
107	K. Chochola, "Spalovací motory" (66)	- " -	31655	31655	-"-	3388	19,0	~ 15000	~ 0,47
108	V. Vojtíšek, "Česká města" (67)	- " -	8729	8729	-"-	2116	22,6	~ 28000	~ 3,21
109	B. Gloss, "Metoda plánování práce" (68)	- " -	18085	18085	-"-	2372	17,6	~ 13000	~ 0,72

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
110	P. Reiman, "Století vědeckého socialismu" (69)	HC, Gruppe E,	10103	10103	~0,038	2831	28,1	~ 60000	~ 5,94
111	Zd. Nejedlý, "Dějiny národa českého" (81)	HC, Gruppe C, Jugendliteratur	31250	31250	~0,035	4366	24,7	~ 30000	~ 0,96
112	J. Ulrich, "Zaklady marxistické ekonomie" (82)	- " -	?	26908	-"-	3577	21,8	~ 22000	-
113	V. Úlehla, "Zamyšleni nad životem II" (83)	- " -	?	20340	-"-	3870	27,1	~ 40000	-
114	O. Chlup, Pedagogika (84)	- " -	?	29813	-"-	4516	26,1	~ 35000	-
115	J. Janko, "Základy statistické indukce" (85)	- " -	?	17249	-"-	1835	13,9	~ 6000	-
116	K. Přerovský. Therapie uhlečitá (86)	- " -	18?	18448	-"-	3088	22,7	~ 24000	-
117	R. Souček, "Psychoanalisa" (87)	- " -	20603	20603	-"-	3502	24,4	~ 28000	1,36
118	K. Honzík, "Tvorba životního slohu" (88)	- " -	?	33700	-"-	5916	32,2	~ 65000	-
119	V. Láska, "Úvod do geofysiky" (89)	- " -	14714	14714	-"-	2790	23,0	~ 26000	1,76

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
120	J. Popelová, "Tri studie z filosofie dejin" (90)	HC, Gruppe C, Jugendliteratur	?	23237	~0,035	3970	26,0	~ 25000	-
121	Zd. Nejedlý, "O kulturu národní a lidovou" (92)	HC, Gruppe H,	17579	17579	~0,050	3137	23,6	~ 27000	~ 1,53
122	E. Burian, "Voláno roz-hlasem" (93)	- " -	22420	22420	-"-	3886	26,0	~ 36000	~ 1,60
123	A. Zápotocký, "Po staru se žít nedá" (94)	- " -	30797	30797	-"-	4849	27,6	~ 40000	~ 1,30

TABELLE 2A

	Text	Quelle	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
124	Častotnyj slovar' voennych tekstov /Häufigkeitswörterbuch der Texte aus dem Militärwesen/	Kolguškin (1970)	689214	0,029	3000	3,62	2150
125	Russische Texte über Elektronik	Kalinina (1968)	200388	0,033	6826	15,3	15200
126	Častotnyj slovar' jazyka Abaja /Häufigkeitswörterbuch der Sprache von Abaj/	Bektaev (1966)	46847	~0,05?	6017	27,8	46847
127	Puškin, Gesamtwerk	Materialy (1963)	544777	0,046	21197	30,0	75000
128	Abschnitte aus dem Werk von V. Pšavela (in Georgisch)	Abuladze (1966)	26467	0,046	5656	34,8	100000
129	- " -	- " -	18187	0,044	4593	34,1	105000
130	- " -	- " -	8280	0,050	2681	29,5	100000
131	Häufigkeitswörterbuch des Russischen	Zasorina (1966)	120474	0,044	14208	40,8	130000
132	Lenins Werke, Bd.1, 1-130	Borodin & Mater (1967)	27110	0,039	6250	38,0	130000
133	Häufigkeitswörterbuch der modernen ukrainischen Prosa	(1969)	100000	~0,035	13945	44,2	142000
134	Häufigkeitswörterbuch des Čechischen	Jelínek & Bečka & Tešitelová (1961)	1623527	0,0413	54486	42,7	210000

$x = \frac{Z}{N}$	Theorethischer Wortschatz v(N,Z)	Einmalige Wörter $\hat{v}_1$	Einmalige Wörter v(1,N,Z)	Einmalige Wörter $\frac{\hat{v}_1}{v(1,N,Z)}$	Zweimalige Wörter $\hat{v}_2$	Zweimalige Wörter v(2,N,Z)	Zweimalige Wörter $\frac{\hat{v}_2}{v(2,N,Z)}$	Dreimalige Wörter $\hat{v}_3$	Dreimalige Wörter v(3,N,Z)	Dreimalige Wörter $\frac{\hat{v}_3}{v(3,N,Z)}$
0,0031	3010	712	512	0,72	284	253	1,12	150	166	0,90
0,076	6800	2009	2080	0,96	863	932	0,93	536	570	0,94
~1	6050	2975	3025	0,98	902	1010	0,89	491	504	0,97
0,138	21200	6388	7300	0,88	2910	3140	0,93	1803	1845	0,96
3,38	5670	-	3450	-	-	890	-	-	390	-
5,77	4570	2714	2920	0,93	733	675	0,92	321	291	1,10
12,1	2665	1675	1840	0,91	418	370	1,13	154	145	1,06
1,08	14200	6752	7200	0,96	2337	2380	0,98	1171	1180	0,99
4,8	6270	3607	3920	0,92	1013	985	1,03	500	410	0,82
1,42	13950	6679	7350	0,91	2220	2300	0,97	1147	1135	1,01
0,129	54500	20476	18530	1,10	7762	7950	0,98	-	-	-

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
135	Häufigkeitswörterbuch des Čechischen, biologische Texte	Jelínek & Bečka & Tešitelová (1961)	32972	~0,038	4750	26,2	~ 36000	1,09
136	-"-, Zeitschrift Tvorba 27,35	- " -	26725	~0,040	5477	33,4	~ 75000	2,78
137	-"-, Zeitung Práce 30.3.1950	- " -	24903	-"-	5315	33,6	~ 80000	3,21
138	-"-, Zeitung Právo lidu 4.9.1946	- " -	13049	-"-	3718	32,6	~ 90000	6,67
139	-"-, Zeitschrift Svět práce 5,36	- " -	32079	-"-	6782	37,8	~120000	3,74
140	-"-, Zeitung Svobodné slovo 20.9.1946	- " -	12574	-"-	3742	33,4	~120000	9,55
141	-"-, Zeitung Rudé právo 5.9.1946	- " -	14502	-"-	4166	34,6	~130000	8,95
142	-"-, Lidová demokracie 11.9.1946	- " -	13369	-"-	4160	36,1	~140000	10,47
143	Čechisch (nach Vey)	Novak (1962)	1523627	~0,05	55795	45,1	~250000	0,164
144	Russische Umgangs-sprache	Ovsienko (1966)	100000	~0,05	6300	19,9	~ 25000	0,25

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
145	Häufigkeitswörterbuch des Russischen	Štejnfel'd (1963)	400000	0,046	24224	38,3	122000	0,30
146	Russische wissenschaftliche Texte	Ludin (1971)	50000	~0,05	9464	42,3	~170000	3,40
147	Russische Umgangssprache	- " -	50000	~0,05	10304	46,2	~230000	4,60
148	Französisch	Novak (1962)	400000	~0,07	7628	12,1	~ 17000	0,042
149	Französische Umgangs-sprache	Novak (1962)	312135	~0,07	7995	14,3	~ 20000	0,064
150	Französische Umgangs-sprache	Ovsienko (1966)	312200	~0,07	8000	14,3	~ 20000	0,064
151	Rumänisch, Belletristik	Novak (1962)	50000	~0,035	4547	20,3	~ 20000	0,40
152	Rumänisch, Publizistik	- " -	70000	~0,035	5710	21,6	~ 25000	0,36
153	Häufigkeitswörterbuch des Rumänischen und des Moldauischen	- " -	300416	0,035	14250	25,8	~ 55000	0,183
154	Englische Telephongespräche (French)	Alekseev (1971)	79300	~0,07	2240	7,95	~ 4250	0,058
155	- " -	- " -	80000	~0,07	2400	8,50	~ 5000	0,063

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
156	Englische Umgangssprache	Alekseev (1971)	512647	~0,07	4539	6,33	~ 6300	0,012
157	Englische Korrespondenz	- " -	23629	~0,07	2001	13,0	~ 8000	0,340
158	Englische Korrespondenz (Cook, O'Shea)	- " -	200000	~0,07	5200	11,6	~ 12000	0,060
159	Englische Umgangssprache	- " -	288000	~0,07	6800	12,5	~ 15000	0,052
160	Englische Elektronik	Alekseev (1969)	200000	0,10	7160	16,0	21000	0,105
161	Englische Texte über Automobilbau	Budman (1971)	100000	0,105	5200	16,4	20000	0,20
162	- " -	- " -	200000	0,107	7355	16,5	24000	0,12
163	Engl. wissensch. Texte	Ludin (1971)	50000	~0,07	5399	24,1	~ 38000	0,76
164	Geschriebenes Englisch der Schüler	Alekseev (1971)	6012359	~0,07	25632	10,45	~ 40000	0,007
165	Englische Zeitungssprache	Ludin (1971	44000	~0,07	6002	28,7	~ 60000	1,37
166	Englische Literatursprache (Dewey)	Alekseev (1971)	100000	~0,07	10161	32,1	~ 80000	0,80
167	Häufigkeitswörterbuch medizinischer Lehrbücher	Levitskij (1966)	82155	0,035	6698	23,4	28300	0,34

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
168	Wörterbuch der organischen Chemie (Zarechnak)	Zasorina (1966)	100000	~0,05	8000	25,3	~ 40000	0,40
169	Afghanisch, Umgangssprache	Ludin (1971)	40000	~0,05?	6411	32,0	~ 70000	1,75
170	-"-, wissenschaftliche Texte	- " -	40000	-"-	6536	32,6	~ 75000	1,87
171	-"-, Zeitungstexte	- " -	40000	-"-	6755	33,8	~ 80000	2,00
172	-"-, künstlerische Prosa	- " -	40000	-"-	7000	35,0	~ 95000	2,38
173	-"-, Poesie	- " -	40000	-"-	7280	36,3	~120000	3,00
174	Gemischte Stichprobe aller afghanischen Texte	- " -	200000	-"-	21268	47,7	~190000	0,95

TABELLE 3

	Text	Nr. der Stich-probe 1	N_1	v_1	Z_1	Nr. der Stich-probe 2 $v(N_1,Z_i)$	Z_2 $\frac{v_1}{v(N_1,Z_2)}$ 100%	Theoretisches Vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
1	Puškin, Kapitanskaja dočka /Kapitänstochter/	28	29345	4783	45000	29	33400	4420	+8,21*	± 6,38
2	- " -	28	-"-	-"-	-"-	30	47000	4860	-1,59	±11,73
3	- " -	28	-"-	-"-	-"-	31	37300	4510	+6,05	± 9,81
4	- " -	29	10000	2432	33400	28	45000	2565	-5,19*	± 4,92
5	- " -	29	-"-	-"-	-"-	30	47000	2606	-6,68	± 8,07
6	- " -	29	-"-	-"-	-"-	31	37300	2440	-0,33	± 8,35
7	- " -	30	5000	1671	47000	28	45000	1640	+1,89	± 5,82
8	- " -	30	-"-	-"-	-"-	29	33400	1570	+6,43*	± 6,40
9	- " -	30	-"-	-"-	-"-	31	37300	1575	+6,10	± 7,94
10	- " -	31	5000	1567	37300	28	45000	1640	-4,45	± 5,80
11	- " -	31	-"-	-"-	-"-	29	33400	1570	-0,19	± 5,92
12	- " -	31	-"-	-"-	-"-	30	47000	1662	-5,72	± 9,80
13	Šolochov, Podnjataja celina/Neubruch/	44	87883	12778	130000	45	106167	11600	+10,15*	± 2,37

TABELLE 3 (FORTSETZUNG)

	Text	Nr. der Stich-probe 1	N_1	v_1	Z_1	Nr. der Stich-probe 2 $v(N_1,Z_i)$	Z_2 $\frac{v_1}{v(N_1,Z_2)}$ 100%	Theoretisches Vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
14	Šolochov, Podnjataja celina /Neubruch/	44	87883	12778	130000	46	127000	12550	+ 1,82	± 2,06
15	- " -	45	106167	12762	106167	44	130000	14130	- 9,68*	± 2,40
16	- " -	45	-"-	-"-	-"-	46	127000	13800	- 7,52*	± 1,92
17	- " -	46	194035	18508	127000	44	130000	18800	- 1,55	± 2,41
18	- " -	46	-"-	-"-	-"-	45	106167	17000	- 8,87*	± 2,30
19	Englischer Automobilbau	162	100000	5200	20000	163	24000	5750	- 9,57*	± 2,60
20	- " -	163	200000	7355	24000	162	20000	6670	+10,27*	± 3,24
21	Važa Pšavela, Auswahl von Texten	129	26467	5656	100000	129	105000	5860	- 3,52	± 4,40
22	- " -	128	-"-	-"-	-"-	130	100000	5670	- 0,25	± 6,86
23	- " -	128	18187	4593	105000	128	100000	4500	+ 2,06	± 4,13
24	- " -	129	-"-	-"-	-"-	130	100000	4500	+ 2,06	± 6,46
25	- " -	130	8280	2681	100000	128	100000	2665	+ 0,60	± 4,67

TABELLE 3 (FORTSETZUNG)

	Text	Nr. der Stichprobe 1	N_1	v_1	Z_1	Nr. der Stichprobe 2 $v(N_1,Z_i)$	Z_2 $(1-\frac{v_1}{v(N_1,Z_2)})100\%$	Theoretisches vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
26	Važa Pšavela, Auswahl von Texten	130	8280	2681	100000	129	105000	2740	- 2,16	± 5,05
27	Kumsiašvili, Portweinproduktion...	16	9998	1906	17000	17	20000	1990	- 4,22	± 7,75
28	- " -	16	-"-	-"-	-"-	18	22400	1980	- 3,74	± 11,71
29	- " -	17	6000	1484	20000	16	17000	1427	+ 3,99	± 6,66
30	- " -	17	-"-	-"-	-"-	18	22400	1470	+ 0,95	± 10,83
31	- " -	18	3000	943	22400	16	17000	933	+ 1,07	± 7,86
32	- " -	18	-"-	-"-	-"-	17	20000	962	- 1,97	± 8,31
33	Macaulay, "Essay on Bacon"	55	2046	964	150000	56	65000	870	+10,80*	± 9,45
34	- " -	55	-"-	-"-	-"-	57	100000	926	+ 4,11	± 8,79
35	- " -	56	4049	1432	65000	55	150000	1625	-11,88	± 11,90
36	- " -	56	-"-	-"-	-"-	57	100000	1540	- 7,01	± 7,56
37	- " -	57	6045	2048	100000	55	150000	2185	- 6,27	± 12,40
38	- " -	57	-"-	-"-	-"-	56	65000	1880	+ 8,93*	± 8,90

TABELLE 4A

	Pikovaja Dama Z = 35000				Il'ja Muromec i Kalin-car Z = 2130				Slovo o polku Igoreve Z = 12500			
N	v	v(N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3δ	v	v (N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3δ	v	v (N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3δ
250	165	169	-2,37	-	107	124	-13,70	±20,42	159	166	-4,22	±19,40
500	281	293	-4,10	±14,70	175	194	- 9,79	-	277	279	-0,72	-
1000	462	497	-7,05	-	257	291	-11,69	±16,75	461	457	+0,88	-
1500	619	685	-9,63*	± 9,60	369	364	+ 1,37	-	613	602	+1,83	±10,78
2000	787	845	-6,87	± 8,13	419	423	- 0,94	-	737	725	+1,65	-
2500	934	985	-5,08	-	461	444	- 3,83	±11,04	842	840	+0,24	±10,35
3000	1067	1120	-4,73	± 7,77	512	518	- 1,16					
4000	1348	1360	-0,87									
5000	1541	1575	-2,16									
6000	1752	1770	-1,02									

TABELLE 4B

	Krejcerova Sonata Z = 22000			Kazaki Z = 53000			Vojna i mir Z = 24000			Voskresenie Z = 53000		
N	v	$\nu(Z,N)$	$(1-\frac{\nu(Z,N)}{v})100\%$	3σ	v	$\nu(Z,N)$	$(1-\frac{\nu(Z,N)}{v})100\%$	3σ	v	$\nu(Z,N)$	$(1-\frac{\nu(Z,N)}{v})100\%$	3σ	v	$\nu(Z,N)$	$(1-\frac{\nu(Z,N)}{v})100\%$	3σ
1000	431	445	- 3,15	-	438	506	-13,42*	±11,15	454	474	-4,22	±11,13	489	506	-3,36	±11,13
2000	685	760	- 9,87*	±8,64	732	850	-13,90*	± 8,53	759	774	-1,94	-	824	850	-3,06	-
3000	886	997	-11,12*	±7,64	1002	1140	-12,12*	± 7,50	1025	1020	+0,49	-	1132	1140	-0,70	-
4000	1092	1200	- 9,00*	±7,14	1233	1395	-11,61*	± 6,94	1218	1230	-0,98	-	1409	1395	+1,00	-
5000	1288	1380	- 6,66	±6,77	1540	1630	- 5,51	± 6,55	1883	1415	-2,26	± 6,70	1640	1630	+0,61	± 6,70
6000	1505	1542	- 2,39	-	1814	1840	- 1,41	-	1553	1585	-2,02	-	1863	1840	+1,25	-
7000	1682	1695	- 0,77	-	2048	2040	+ 0,34	-	1691	1740	-2,82	-	2049	2040	+0,44	-
8000	1819	1830	- 0,60	-	2253	2225	+ 1,26	-	1830	1881	-2,70	± 6,23	2237	2225	+0,54	± 6,23
9000	1958	1960	- 0,10	-	2438	2400	+ 1,58	-	2005	2020	-0,74	-	2411	2400	+0,46	-
10000	2048	2080	-	-	2582	-	-	-	2146	2145	-	-	2587	2570	-	-

ANHANG I

EINE APPROXIMATIVE METHODE ZUR BERECHNUNG DES ZIPFSCHEN UMFANGS

Gleichungen wie (13) löst man üblicherweise durch iterative Approximation. Damit das Verfahren beginnen kann, muß man die erste Approximation, den groben Orientierungswert der Gleichungswurzel kennen.

Eine gute erste Approximation kann man erhalten, wenn man Gleichung (13) graphisch löst. Die beigefügten Nomogramme erlauben, eine Übereinstimmung der linken und rechten Seite von (13) mit höchstens 3-5% Abweichung zu erreichen.

Jede der Kurven auf Abb. 4-8 stellt den Quotienten von Guiraud für einen homogenen Text mit festgelegten Werten Z und p_{max} dar:

$$R(N) = \frac{v(N,Z)}{\sqrt{N}}.$$

Es ist leicht zu sehen, daß diese Beziehung am Anfang des Textes wächst und dann anfängt abzunehmen. Gerade wegen dieses Umstandes ist diese Beziehung als Maß des relativen Vokabularreichtums ungeeignet.

Die stark schiefe Funktion

$$R_2(N) = \frac{v(N)}{\sqrt{N}}$$

stellt eine analoge Beziehung für Texte dar, die bei ihrem Umfang dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen, mit anderen Worten, jeder Punkt dieser Kurve entspricht einem Text mit N = Z, und der Zähler dieser Beziehung wird laut (7) berechnet. Daraus ist evident, daß die Abzissen der Schnittpunkte von R(N) und $R_2(N)$ gleich den Werten von Z für die Kurve R(N) sind. (Aus diesem Grund wurden die Werte von Z, die jede Kurve charakterisieren, auf die Graphiken nicht aufgetragen.) Es ist interessant, daß die Kurve $R_2(N)$ durch die Maxima der Kurven

R(N) läuft: das bedeutet, daß Guirauds Beziehung gerade beim Zipfschen Umfang den für den gegebenen Text höchsten erreichbaren Wert annimmt.

Sei gegeben eine lexikalische Stichprobe mit bekannten N, $\hat{v}$ und p_{max}. Für die Bestimmung von Z berechnen wir $\hat{R} = \frac{\hat{v}}{\sqrt{N}}$ und tragen es als Punkt auf das Nomogramm auf, bei dem der Parameter p_{max} der Häufigkeit des häufigsten Wortes am nächsten steht. Es kann eines von drei Ereignissen zustande kommen:

1. Der Punkt $(N,\hat{R})$, unserer Stichprobe liegt auf der stark schiefen Kurve $R_2(N)$. In dem Falle nehmen wir an, daß $Z = N$ ist.
2. Der Punkt $(N,\hat{R})$ liegt auf einer der konkaven Kurven R(Z). In diesem Falle ist Z als die Abzisse des Schnittpunkts dieser Kurve mit der Kurve $R_2(N)$ gegeben.
3. Am wahrscheinlichsten ist aber, daß der Punkt $(N,\hat{R})$ irgendwo zwischen den Kurven liegt. In dem Fall ziehen wir von diesem Punkt "frei" eine Kurve parallel zu den Kurven der Familie R(N) bis zum Schnittpunkt mit $R_2(N)$. Die Abzisse des Schnittpunkts ergibt dann den Wert von Z.

Den auf diese Weise bestimmten Wert von Z muß man zusammen mit N und p_{max} der gegebenen Stichprobe in die Formel (10a) oder (10b) einsetzen, den Wortschatz der Stichprobe beim Umfang N und den gefundenen Wert Z berechnen, und mit dem tatsächlichen Wortschatz $\hat{v}$ vergleichen. Alles weitere hängt sowohl von der erhaltenen Übereinstimmung als auch von den verfolgten Zielen ab.

Wenn die Divergenz zwischen $\hat{v}$ und v(N,Z) 5% nicht übersteigt, so kann man annehmen, daß die graphische Berechnung korrekt ist (man soll aber nicht vergessen, daß die Divergenz einfach durch arithmetische Fehler bei der Berechnung von v(N,Z) entstehen kann). Und wenn man nicht das Ziel einer möglichst exakten Prognose "vorwärts" oder "rückwärts" oder eines Vergleichs zweier Texte mit ähnlichen Z verfolgt, so reicht die Berechnung aus. Die Z-Werte, die wir in der vorliegenden Arbeit mit dieser "graphischen Genauigkeit" berechnet haben, sind in den Tabellen mit ~ gekennzeichnet. Dieser Fall entspricht einer 10-15%-Genauigkeit der Berechnung von Z.

Wenn man Z exakter berechnen möchte, muß man die iterative Approximation verwenden. In dieser Arbeit wurde eine vereinfachte Chordenmethode benutzt, die im Vergleich mit anderen, auf Kosten der Einfachheit einzelner Schritte schneller konvergierenden Prozeduren, eine beträchtliche Rechenökonomie bedeutet.

Wenn der aus dem Nomogramm berechnete Zipfsche Umfang als die erste Approximation betrachtet und als Z_1 bezeichnet wird, dann kann man die zweite Approximation aus der Formel

$$Z_2 = \left[\frac{\hat{v}}{v(N,Z_1)}\right]^2 \cdot Z_1$$

finden. Die dritte und die weiteren Approximationen findet man dann als

$$Z_i = Z_{i-1} + \frac{\hat{v} - v(N,Z_{i-2})}{v(N,Z_{i-1}) - v(N,Z_{i-2})} \cdot (Z_{i-1} - Z_{i-2}).$$

Nach der Berechnung eines Z_i überprüft man die Übereinstimmung zwischen der rechten und der linken Seite von (13). In der vorliegenden Arbeit wurde der Prozeß solange fortgeführt, bis die Genauigkeit für den Wortschatz nicht schlechter als 1% war. Und obwohl dies ungefähr einer 5% Genauigkeit bei der Bestimmung von Z selbst entspricht, hat eine größere Genauigkeit offensichtlich keinen Sinn, da die beobachtete zufällige Streuung der Größe Z bedeutend größer ist.

Wenn wir Z_1 graphisch berechnen, so erreichen wir diese Genauigkeit üblicherweise spätestens im dritten Schritt. Bei der Analyse einiger "exotischer " Stichproben (entweder zu großer wie z.B. {164}, oder zu kleiner {1,2,3,22,24} oder schließlich solcher mit zu großem Z {50}) lag der Punkt $(N,\hat{R})$ freilich außerhalb des Nomogramms. In dem Fall wurde der Wert Z_1 "aus der Luft gegriffen" und der Prozeß der iterativen Approximation endete beim 6-7-ten Schritt.

Als Beispiel führen wir die Berechnung von Z für einen Abschnitt aus Puškins "Kapitänstochter" {30} an. Hier ist

$$N = 5000$$

$$\hat{v} = 1671$$

$$p_{max} = 0{,}04$$

$$\hat{R} = \frac{1671}{\sqrt{5000}} = 23{,}6.$$

Der an p_{max} nächstliegende Wert auf unseren Nomogrammen ist $p_{max} = 0{,}035$. Auf dieses Nomogramm (Abb. 5) tragen wir den Punkt (5000; 23,6) auf. Wir führen "optisch" diesen Punkt entlang der konkaven Kurven bis zu der steilen Kurve $R_2(N)$, und die Stelle, wo sie geschnitten wird, projizieren wir auf die Achse N. Die erhaltene Abszisse hat ungefähr den Wert 47000. Diese Zahl betrachten wir als den Wert von Z.

Überprüfung:

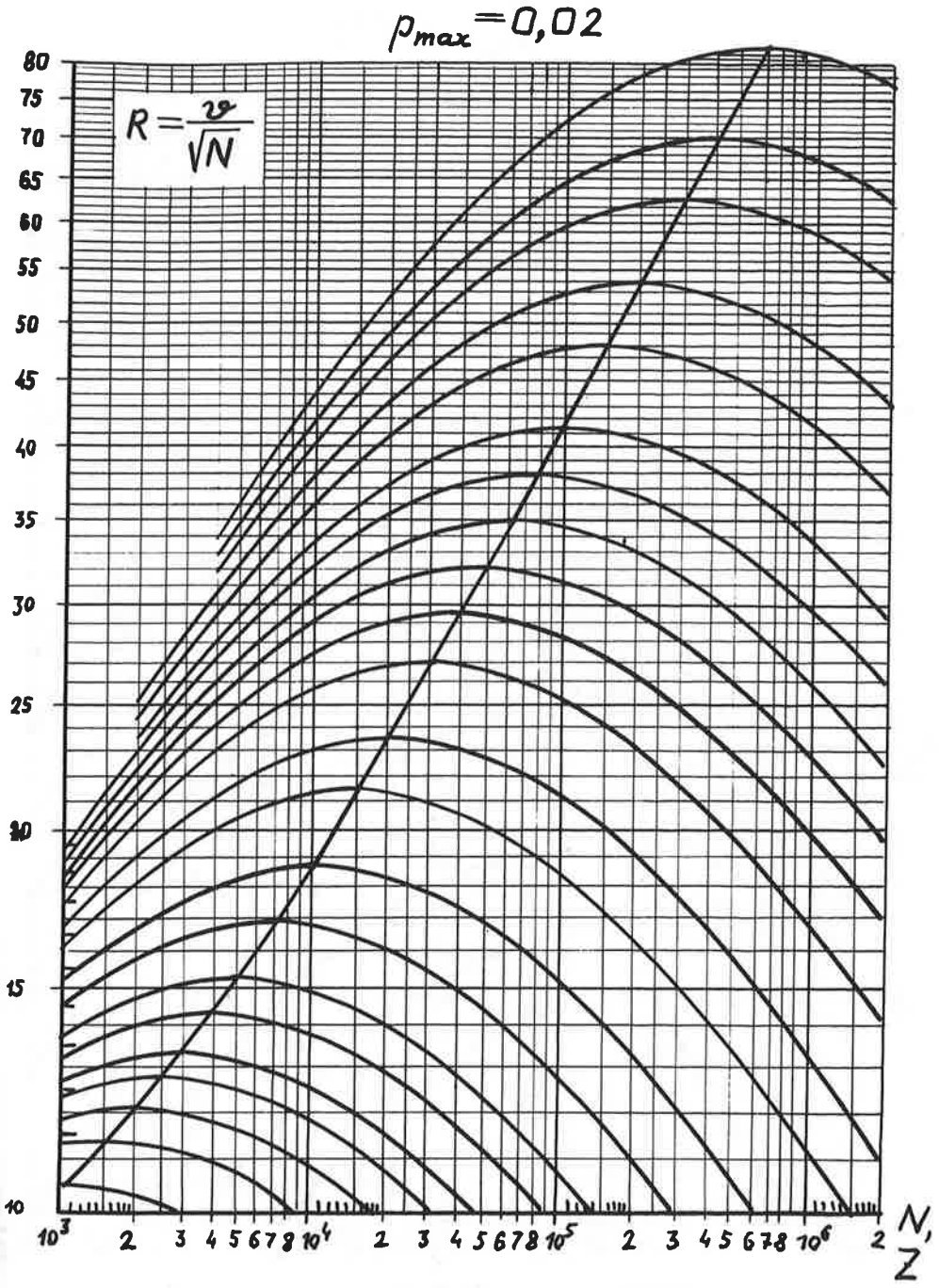
$$X = \frac{47000}{5000} = 9{,}4$$

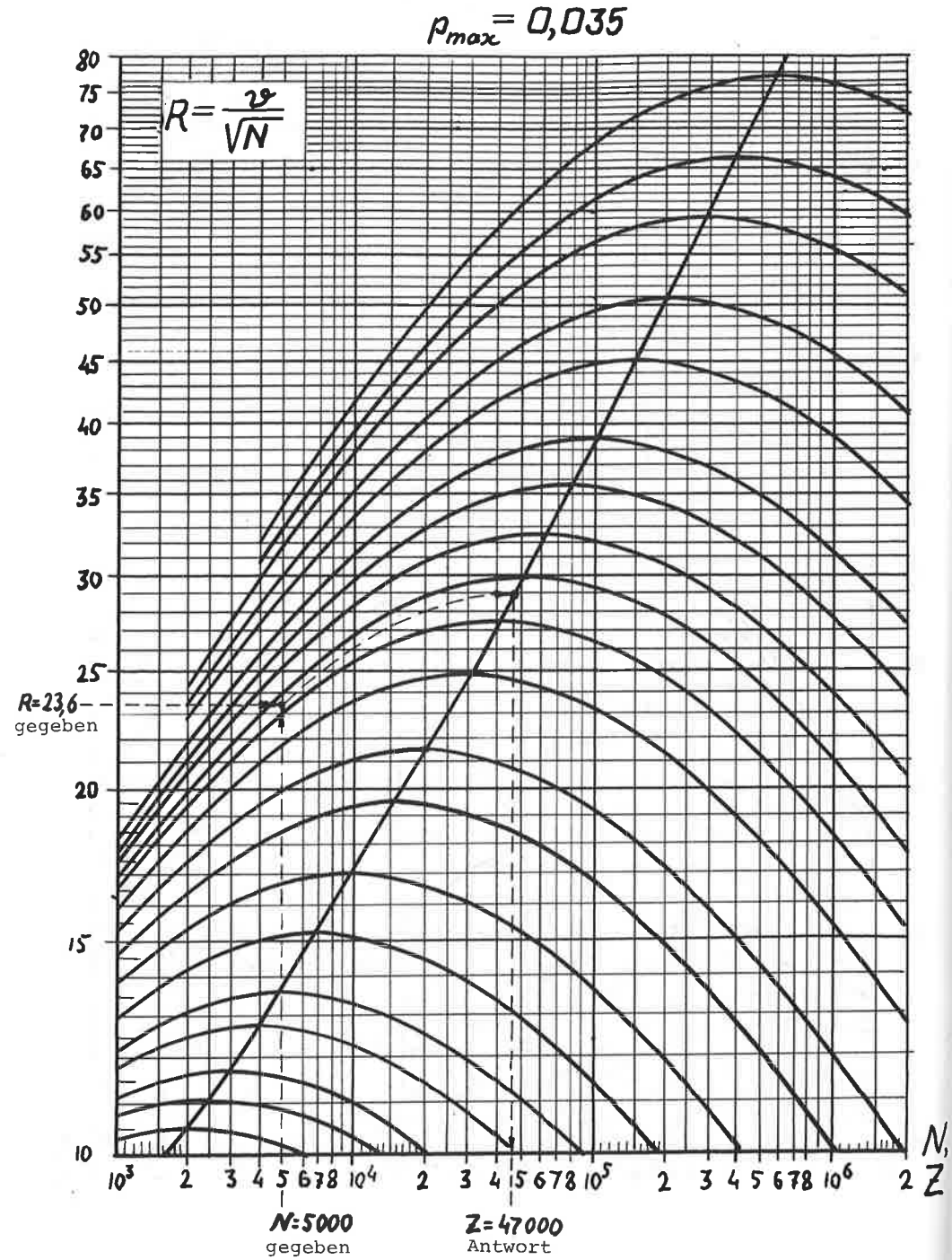
$$v(Z) = \frac{47000}{\ln\,[47000(0{,}04)]} = \frac{47000}{7{,}54} = 6234$$

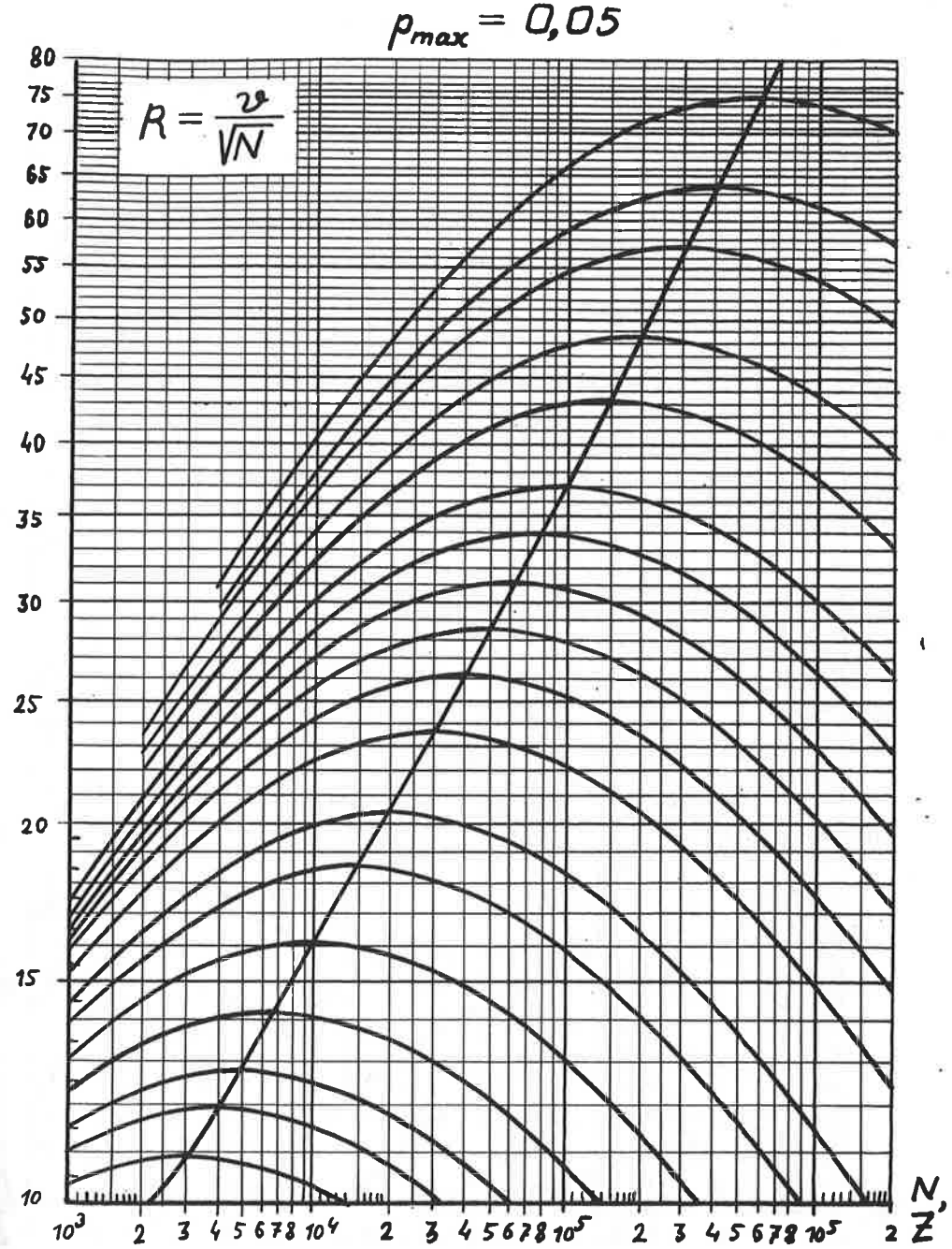
$$v(N,Z) = \frac{6234\ \ln\ 9{,}4}{8{,}4} = 1663.$$

Da sich diese Größe von dem Wortschatz der Stichprobe (1671) um weniger als 1% unterscheidet, ist die Berechnung beendet.

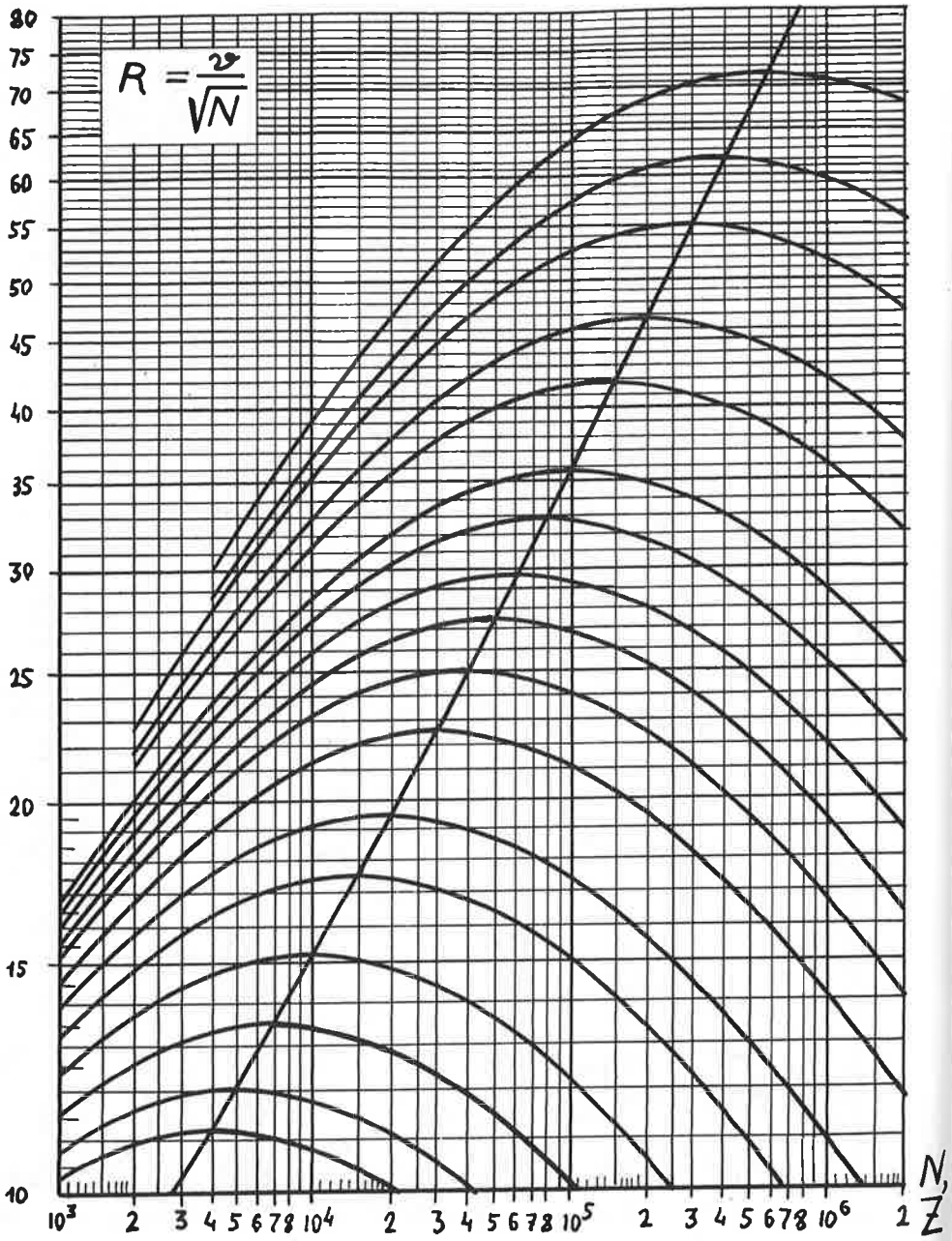
Zum Schluß möchten wir bemerken, daß es mit Hilfe angeführter Nomogramme möglich ist, den Wortschatz des Textes bei verschiedenen Umfängen vereinfacht zu berechnen. Speziell im besprochenen Beispiel kann man für N = 29345 (voller Umfang der "Kapitänstochter") $R \approx 28{,}5$ ablesen. Multipliziert man diese Größe mit der Wurzel aus der Textlänge der "Kapitänstochter", so erhält man für den gegebenen Abschnitt $28{,}5\sqrt{29345} = 4882$, was sich "nach allen Regeln" nur unbedeutend von der theoretischen Prognose 4860 unterscheidet (vgl. Zeile 2 der Tab. 3).



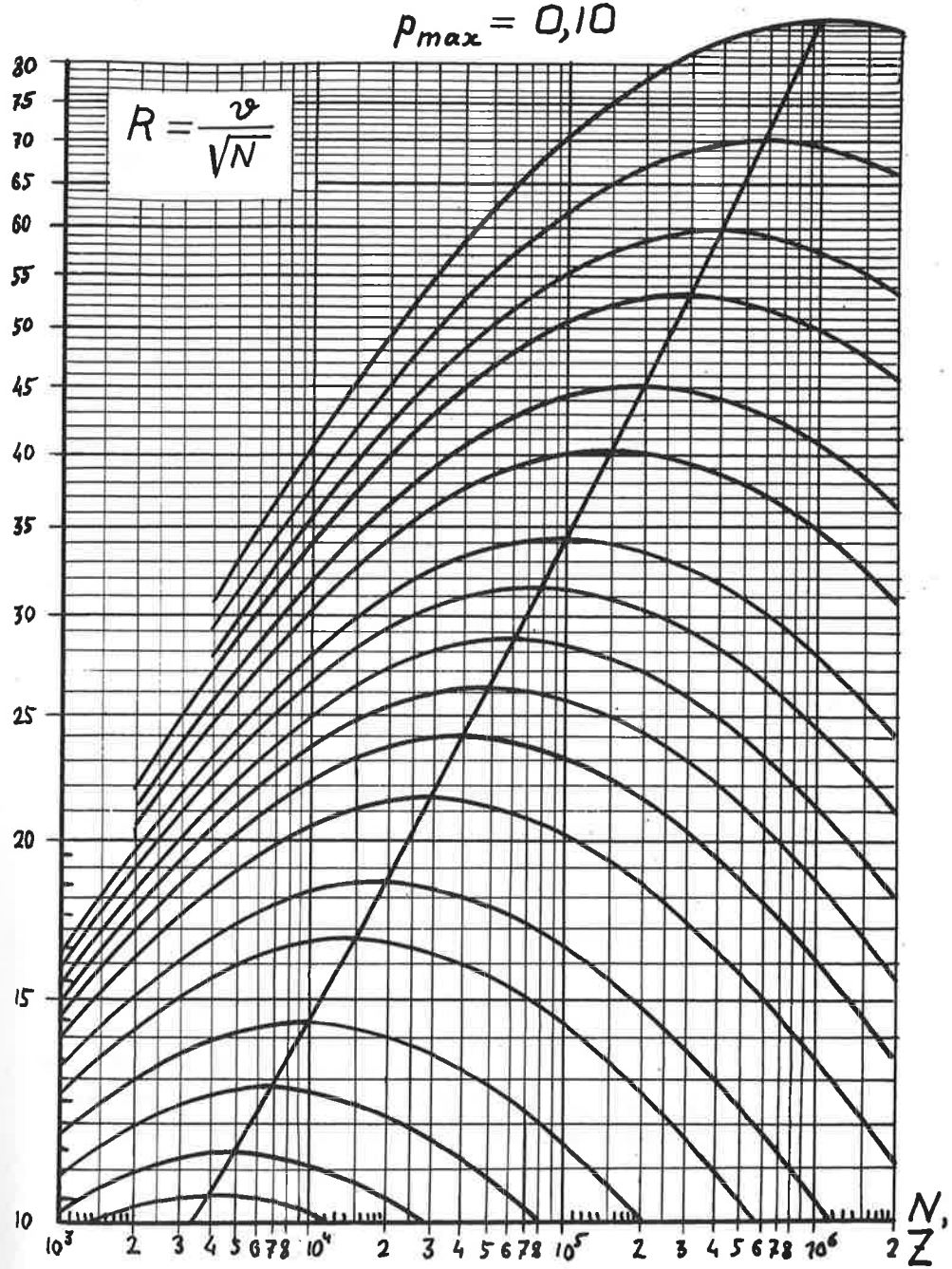

$p_{max} = 0{,}035$
$R = \frac{v}{\sqrt{N}}$
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
R=23,6
gegeben
10^3 2 3 4 5 6 7 8 10^4 2 3 4 5 6 7 8 10^5 2 3 4 5 6 7 8 10^6 2
N, Z
N=5000
gegeben
Z=47000
Antwort


$p_{max} = 0{,}05$
$R = \frac{v}{\sqrt{N}}$
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
10^3 2 3 4 5 6 7 8 10^4 2 3 4 5 6 7 8 10^5 2 3 4 5 6 7 8 10^5 2
N, Z

$p_{max} = 0{,}07$



$p_{max} = 0{,}10$



ANMERKUNGEN

1 Auch wenn die Meinung verbreitet ist, daß γ nicht einmal im Rahmen eines Textes konstant ist (vgl. Frumkina 1961), werden wir unten zeigen, daß dieser Effekt durch den nichtberücksichtigten Stichprobenumfang entsteht, wenn sich dieser von Z unterscheidet.

2 Genauer, wir postulieren eine solche Wahrscheinlichkeitsverteilung der Wörter in einer allgemeinen lexikalischen Gesamtheit (zufällige Stichprobe, aus der der Text generiert wird), daß bei einem Umfang N_o die mathematische Erwartung der Anzahl m-maliger Wörter durch (8) gegeben wird. Diese etwas ungewöhnliche Form der statistischen Hypothese erlaubt es, das "Raten" der unbekannten Menge der Wortwahrscheinlichkeiten zu vermeiden. Man kann eine solche Hypothese formal nicht begründen, aber die erhaltenen Resultate rechtfertigen ihre Annahme.

3 Um eine Genauigkeit von mindestens 1% zu erreichen (bei linguistischen Berechnungen besteht kein Anlass zu einer größeren Genauigkeit), reicht es, in (10a) die ersten drei Glieder zu nehmen, wenn $-0{,}36 \leq X - 1 \leq 0{,}36$, und fünf Glieder, wenn $-0{,}5 \leq X - 1 \leq 0{,}5$.

4 Vgl. auch die Formel (45b) in Orlov (1976).

5 Es ist zu betonen, daß Formel (11a) einerseits und Formeln (11b), (11c) andererseits wie auch die Formeln (40-44) in Orlov (1976) keine unterschiedlichen mathematischen Abhängigkeiten , sondern nur unterschiedliche Formen einer und derselben Abhängigkeit darstellen, ähnlich wie $(a+b)^2$ und $a^2+2ab+b^2$.

6 In den erwähnten Arbeiten ist die entsprechende Größe mit N_o bezeichnet: s. auch die Arbeit von Ju.K. Orlov. Why, How and When does the Zipf-Mandelbrot Law Fail? SMIL Quarterly, Stockholm, Skriptor, 4, 1977, 5-27.

7 Die etwas langwierigeren Methoden der approximativen Lösung der Gleichung (13) werden im Anhang 1 erläutert.

8 Wenn wir den exakten Wert der mathematischen Erwartung des Vokabularumfangs beim Stichprobenumfang N_1 kennen würden, so könnten wir durch Lösung von (13) den exakten Wert von Z für den gegebenen Text berechnen. In dem Falle würde sich die relative Standardabweichung für die Prognose ausschließlich aus δ_2 ergeben. Jedoch ist das tatsächliche Vokabular $\hat{V}^{(1)}$ im Beobachtungspunkt N_1 eine Zufallsvariable, die im Rahmen eines Konfidenzintervalls (a_1,b_1) (vgl. Abl.) schwanken kann. Dementsprechend kann auch die Prognose für den Punkt N_2 schwanken und zwar im Intervall (a_2,b_2), dessen Größe sowohl vom Intervall (a_1,b_1) als auch vom Unterschied zwischen N_1 und N_2 und von den geometrischen Eigenschaften der Kurven (10) abhängt. Wenn wir also die relative mittlere Standardabweichung S_1 im Punkt N_1 berechnet haben, müssen wir sie auf den Umfang N_2 "projizieren". Diese Aufgabe erfüllt eben der Koeffizient

$$W = \frac{b_2 - a_2}{b_1 - a_1} .$$

Der erste Summand in (15) stellt also die Dispersion der Prognose dar, die aus zufälligen Schwankungen des Vokabulars in Punkt N_1 folgen, und der zweite Summand die Dispersion des Vokabulars im prognostizierten Punkt N_2.

9 Die Zahlen in geschweiften Klammern beziehen sich auf die Numerierung der Texte und Stichproben in den Tabellen 1 - 4 (vgl. Anhang 2).

10 Eine mögliche Ursache wird unten besprochen.

11 In dieser Analyse wurden die beobachteten relativen Abweichungen (in %) algebraisch summiert.

12 Die für einige Texte charakteristische Unterbelegung des vorhandenen Vokabulars bei der Prognose "rückwärts" veranlaßte uns, die Analyse der Abweichungen in Text 3 genauer zu betrachten. Die Abweichungen in Tab. 3 sind gewissermaßen symmetrisch (während die Prognose für eine kleinere Stichprobe aus den Daten einer größeren Stichprobe höher ausfällt, so ist die umgekehrte Prognose für eine größere Stichprobe aus den Daten einer kleineren Stichprobe niedriger; bei der algebraischen Summierung der relativen Abweichungen kompensieren sich diese Fehler gegenseitig) und das Fehlen einer systematischen Verschiebung kann eben durch diese Symmetrie erklärt werden. Die Unterscheidung von "vorwärts"- und "rückwärts"-Prognosen zeigte (Tab. 3), daß in jeder Klasse der Prognosen signifikante Abweichungen vorhanden sind.

13 Wir bemerken, daß bei der Hypothese R = const., der Fehler der Prognose über den Vokabularumfang viel größer ist; für die "Kapitänstochter" (beim vollen Umfang ist R = 27,9) ergibt sich das folgende Bild:

für N = 10000, $v = 27{,}9\sqrt{10000} = 2790$
(der tatsächliche Wert ist 2432; Abweichung -12,8%);

für N = 5000, $v = 27{,}9\sqrt{5000} = 1970$
(die tatsächlichen Werte sind 1671 und 1567; Abweichungen -15,2% bzw. -20,5%).

Man vergleiche die Zeilen 4,7 und 10 in Tab. 3.

14 Offensichtlich werden die auf Abb. 2 ersichtlichen Abweichungen der experimentellen Kurve von der theoretischen durch die besonders hohe Inhomogenität der Gesamtwerke von Puškin hervorgerufen.

15 In einigen Fällen wird diese Anordnung geringfügig gestört, damit die Reihenfolge der Zeilen, die sich auf einen Text oder auf ähnliche Texte beziehen, intakt bleibt.

16 Wir bemerken, daß es bei der Untersuchung der Entsprechung zwischen Z und der Textlänge keine Möglichkeit gibt, die Unterschiede der lexikalischen Zählungen zwischen einzelnen Forschern auf einen gemeinsamen Nenner zu bringen. Bisher gibt es keine allgemein angenommene Bestimmung des Wortes und die unterschiedlichen Arten der Identifizierung der Wortform machen sich im Wortschatz bemerkbar. So schreibt z.B. Osmanov (1970): "Wir bringen absichtlich keine zusammengesetzten Verben, die aus Substantiven und Verben gebildet werden, da es im Persischen oft ziemlich schwierig ist festzustellen, ob ein Substantiv und ein Verb ein zusammengesetztes Verb oder eine Wortgruppe bilden. Wenn wir die zusammengesetzten Verben erfassen würden, so würde Unsuris Vokabular beträchtlich anwachsen". In diesem Falle würde offensichtlich auch das Z des Textes größer werden und die Übereinstimmung zwischen der Textlänge von Unsuris "Divan" {21} und seinem Zipfschen Umfang würde sich verbessern (nach Osmanovs Angaben ist X = 26400/46472 = 0,57). Infolgedessen besteht Grund zur Annahme, daß bei einer standardisierten Zählung und Benutzung der Materialien in einer Sprache die beschriebene Korrelation stärker sein müßte. (Die letztgenannten Bedingungen werden zwar in dem "Häufigkeitswörterbuch des Tschechischen" erfüllt, aber trotzdem tritt die erwartete Vergrößerung der Korrelation nicht ein, da die Angaben über die volle Länge der Texte, aus denen die Stichproben erhoben wurden (z.B. {58, 62, 64, 66} u.a.), und die Angaben über die eventuellen Gliederungen der Texte fehlen.)

17 Der Mensch bemerkt die Veränderung einer Größe, wenn sie 10-30% des ursprünglichen Wertes übersteigt (Konstanz der sogenannten differentialen Wahrnehmungsschwelle, das Weber-Fachnersche Gesetz). Kleinere Veränderungen werden von dem Menschen nicht wahrgenommen. Alle charakteristischen Abweichungen zwischen theoretischen Prognosen und tatsächlichen Beobachtungen in dieser Arbeit (nicht nur die Abweichungen des L(X) von 0.5) liegen in dem Intervall der relativen Abweichungen von ± 10-30%. Früher haben wir bemerkt (Orlov 1969a,b, 1970a, 1974, 1978b), daß bei der Prognostizierung des Vokabulars einzelner voller literarischer Texte nach Formel (7), die Benutzung der Textlänge N_o statt Z (in der überwältigenden Mehrheit der Fälle) zu einem Prognosefehler von ± 20% führt. D.h. nimmt man die Zipf-Mandelbrotsche Häufigkeitsstruktur (beschrieben durch (6), (7), (8)) als "ideales Modell" der quantitativen Organisation des literarischen Werks, dann befinden sich die beobachteten Abweichungen von diesem "Ideal" genau an der Grenze, wo der Mensch gerade anfängt eine Abweichung zu beobachten. Die Annahme der differentiellen Geschwindigkeit des Vokabularwachstums (17) als "Ideal" hat den Vorteil, daß diese Kurve universell ist und ihr kritischer Wert 0.5 für Texte beliebiger Länge konstant ist, wenn sie dem Zipf-Mandelbrotschen Gesetz folgen (d.h. wenn L(X) sich 0.5 nähert, dann kann man den Text mit (6), (7) und (8) gut beschreiben).

Vergleichende statistische Wortschatzanalyse als Methode zur Untersuchung des Werkes eines Schriftstellers (am Beispiel der Prosa von K. Gamsachurdija)

I.Š. Nadarejšvili

In den letzten Jahrzehnten haben sich die Linguisten intensiv mit der Zusammenstellung von Häufigkeitswörterbüchern befaßt. Das Fernziel dieser Arbeiten war eine statistische Beschreibung der Sprache, in der jedem Wort seine Verwendungshäufigkeit zugeordnet wäre. Die gewonnenen Zahlenwerte gerieten aber immer mehr miteinander in Widerspruch. Jetzt setzt sich in der Sprachstatistik die Ansicht durch, daß es unmöglich ist, ein Häufigkeitswörterbuch der Gesamtsprache zu verfassen (vgl. Arapov, Efimova, Šrejder 1975). Selbst zwischen den Werken eines einzelnen Schriftstellers schwanken die Worthäufigkeiten stark, und der Wortschatz nimmt unterschiedlich schnell mit dem Textumfang zu (d.h. bei gleicher Textlänge findet man unterschiedlich viele verschiedene Wörter) (vgl. Nadarejšvili, Orlov 1969; Nadarejšvili 1978). Auch andere Befunde legten die Annahme nahe, daß es quantitative Gesetzmäßigkeiten gibt, die an den einzelnen Text als ein geschlossenes Ganzes gebunden sind (vgl. Nadarejšvili, Orlov 1969; Nadarejšvili 1970). Hierdurch hat die statistische Analyse zwar als Instrument zur Untersuchung gesamtsprachlicher Gesetzmäßigkeiten an Interesse verloren, gleichzeitig aber als Instrument zur Textuntersuchung an Interesse gewonnen. Sie erweist sich nämlich als ein sehr wirksames Instrument zur Lösung textwissenschaftlicher Probleme, von der Ermittlung der Autorenschaft bis hin zu sehr

generellen literaturwissenschaftlichen Fragen und der Psychologie der Kreativität.

In der vorliegenden Arbeit wird mit statistischen Methoden der Wortschatz in K. Gamsachurdijas Romanen "David, der Baumeister" (DS) und "Die Hand des großen Meisters" (DVM) sowie seiner Erzählung "Der Fotograf" (F) untersucht. Die beiden Romane beschreiben das Georgien des 11. - 12. Jahrhunderts, der Epoche Zar Georgs I. und seines Urenkels David II. Sie sind ähnlich im Stil, jedoch verschieden im Aufbau, in der Handlung und im Thema. Obwohl "David, der Baumeister" ein halbes Jahrhundert nach "Die Hand des großen Meisters" spielt, setzt seine Handlung nicht einfach die von "Die Hand" fort. In der Erzählung "Der Fotograf" spielt die Handlung zu Beginn unseres Jahrhunderts. Die Hauptperson der Erzählung, ein Fotograf, reist aus Paris in seine Heimat Georgien; die Handlung spielt an vielen Orten - in Paris, im Zug, auf dem Schiff, im Heimatdorf.

Eine vorläufige Analyse zeigte, daß der Wortschatz mit zunehmendem Textumfang in beiden Romanen praktisch gleich schnell ansteigt (vgl. Nadarejšvili, Orlov 1974) (siehe auch Tabelle 1). Es drängt sich die Hypothese auf, daß beide Texte gleichsam aus einem Gesamtlexikon (wie aus einer Urne mit Kugeln) entnommen sind, wobei die Wörter mit konstanten, nicht vom Werk abhängigen Wahrscheinlichkeiten selegiert worden sind.

Um diese Hypothese zu überprüfen, wurden aus den beiden Romanen zwei Stichproben genommen, und zwar jeweils die ersten 10000 Wortvorkommen, vom Textanfang an gerechnet. Die Stichproben umfassen etwa 1/7 des Text von DVM und 1/34 des Text von DS. Das in diesen Stichproben ermittelte verbale Repertoire enthält, wie noch gezeigt wird, annähernd 1/3 bis 1/4 des Gesamtwortschatzes von DVM und 1/7 bis 1/8 des Gesamtwortschatzes von DVM und 1/7 bis 1/8 des Gesamtwortschatzes von DS. Man kann daher die beiden Stichproben als genügend repräsentativ ansehen, und ihre identische Plazierung am Textanfang rechtfertigt eine vergleichende Analyse.

Tabelle 1.

Stichprobenumfang	DVM		DS	
	empirischer Wortschatzumfang v	theoretischer Wortschatzumfang v(N,Z)	empirischer Wortschatzumfang n	theoretischer Wortschatzumfang v(N,Z)
1000	625	572	622	588
2000	1066	995	1022	1020
3000	1437	1365	1383	1397
4000	1735	1700	1733	1740
5000	2035	2010	2026	2050
6000	2315	2310	2306	2340
7000	2551	2580	2592	2630
8000	2817	2850	2858	2890
9000	3102	3090	3119	3140
10000	3321	3330*	3378	3380*
11000	3607	3560	3693	3610
12000	3818	3780	3910	3860
13000	4041	4000	4118	4080
14000	4262	4200	4300	4300
15000	4466	4420	4501	4500
16000	4653	4620	4687	4710
17000	4856	4800	4849	4900
18000	5044	5010	5046	5090
19000	5238	5180	5276	5270
20000	5443	5370	5443	5450

* Mit einem Sternchen sind die theoretischen Werte des verbalen Repertores markiert, die den entsprechenden faktischen durch Berechnung von Z nach Formel (2) "angepaßt" sind.

Die Wörter jeder Stichprobe wurden nacheinander auf einzelne durchnumerierte Kärtchen geschrieben; so wurde jedem Wortvorkommen die Nummer seiner Position im Text zugeordnet. Danach wurden die Kärtchen alphabetisch geordnet und die Auftretenshäufigkeit jedes Wortes gezählt. Für die Wortschatzkartei wurde dann von jedem Wort die Karte mit der kleinsten Nummer genommen, die die Stelle des ersten

Auftretens des Wortes angab. Die Wortschatzkartei wurde dann nach diesen Positionsnummern geordnet, wobei jede Karte eine fortlaufende zweite Nummer erhielt. Die zweite Nummer eines Wortes gibt den mit seinem ersten Auftreten erreichten Umfang des Wortschatzes an, die erste Nummer den an der Stelle erreichten Textumfang. Mit Hilfe dieses Verfahrens kann man verfolgen, wie der Umfang des Wortschatzes mit wachsendem Textumfang anwächst, und zwar mit der Genauigkeit von einem Wort. Tabelle 1 (ihre obere Hälfte bis einschließlich N = 10000) die das Anwachsen des Wortschatzes in beiden Stichproben zeigt, ist nur eine übersichtliche Zusammenfassung der erhaltenen Inventare von jeweils mehr als 3000 Wörterbucheintragungen. Man sieht leicht, daß der Wortschatz in beiden Stichproben praktisch identisch anwächst.

Zur Überprüfung der Hypothese, daß die aus verschiedenen Romanen stammenden beiden Stichproben zu einer gemeinsamen lexikalischen Gesamtheit gehören, wurde der gemeinsame Wortbestand ermittelt. Er umfaßte 1256 Wörter. Die Auftretenshäufigkeiten dieser Wörter in den beiden Texten, F_1 und F_2, wurden nach dem Studentschen Kriterium auf einem Konfidenzniveau von 0.95 verglichen. D.h. der Unterschied zwischen den beiden Häufigkeiten eines Wortes wurde als statistisch bedeutsam angesehen, wenn $|F_1 - F_2|/\sqrt{F_1 + F_2} > 1.96$ war[1].

Insgesamt fanden sich 71 Wörter, d.h. 71/1296 x 100 = 5.65% des gemeinsamen Wortschatzes, deren Häufigkeiten sich signifikant unterschieden. Dieser Bruchteil liegt nur ganz unbedeutend (um 0.65%) über dem gewählten Signifikanzniveau (5%). Mit anderen Worten, man kann annehmen, daß die Mehrzahl der beobachteten Abweichungen letztlich zufällig sind.

Allerdings enthält der gemeinsame Wortschatz nur etwas mehr als ein Drittel des Wortbestandes von jedem Text. Die Frage, wie groß der nichtübereinstimmende Teil des Wortbestands in Stichproben aus einer einheitlichen lexikali-

schen Gesamtheit sein kann, läßt sich mit einem Modell der Worthäufigkeitsstruktur (vgl. Orlov 1978) lösen, mit dessen Hilfe man die Zunahme des Wortschatzes in einem statistisch homogenen Text mit wachsendem Stichprobenumfang beschreiben kann und das auch die Berechnung eines Konfidenzintervalls zur Prognose des Wortbestands nach empirischen Ausgangsdaten ermöglicht. Nach diesem Modell, das an einer großen Anzahl von Texten und Stichproben Bestätigung gefunden hat, ist der Wortbestand in einer Stichprobe aus N Wortvorkommen gleich

$$v(N,Z) = v(Z)\frac{\ln X}{X-1}, \text{ wo } X = \frac{Z}{N} \cdot \qquad (1)$$

Hierin ist $v(Z) = Z/\ln(p_1)$; p_1 ist die relative Häufigkeit des häufigsten Wortes, und Z ist ein Parameter. Wenn die Größe des Wortschatzes v bei einem bestimmten Stichprobenumfang N_1 bekannt ist, so bestimmt sich der Parameter Z aus der Gleichung

$$v(N_1, Z) = v. \qquad (2)$$

Tabelle 2 gibt numerische Lösungen dieser Gleichung (2) nach Einsetzung der Daten aus den Stichproben von DS und DVM. Es werden dort auch die Ausgangsdaten p_1, N, v und der nach Formel (1) berechnete Wortschatzumfang für den Gesamtumfang N_o des jeweiligen Texts angegeben (N_o wurde annähernd bestimmt über die Schätzung der mittleren Anzahl von Wortvorkommen je Romanseite). Wie aus der Tabelle ersichtlich, sind die Z-Werte beider Texte sehr ähnlich (150000 und 170000). Dies spiegelt den ähnlichen Verlauf des Wortschatzwachstums in beiden Texten wider. Eine gewisse Differenz ist durch die unterschiedlichen p_1 in beiden Texten entstanden.

Tabelle 2

Text	p_1	N	v	Z	N_o	$v(N_o,Z)$
DS	0.0338	10000	3378	150000	337200	25600
DVM	0.0498	10000	3321	170000	69600	11620
F	0.0348	4381	2002	250000	4381	2002

Die Werte für v(N,Z) in Tabelle 1 sind durch Einsetzung der Z-Werte aus Tabelle 2 in Formel (1) berechnet worden. Man sieht eine vorzügliche Übereinstimmung zwischen den theoretischen und den empirischen Werten. Man kann daher die Vorhersagen für die Gesamttexte in der letzten Spalte von Tabelle 2 als sicher ansehen. Obgleich der Umfang von DS fast fünfmal so groß ist wie der von DVM, ist der Wortschatz von DS nur etwas mehr als zweimal so groß wie der von DVM, und die lexikalische Konzentration (der relative Vokabularreichtum), d.h. die Anzahl unterschiedlicher Wörter bei verschiedenen Textumfängen, ist in beiden Romanen gleich.

Da Formel (1) unter der Voraussetzung einer Zufallsstichprobe aus homogener Grundgesamtheit gewonnen wurde, kann sie zur Überprüfung der Hypothese verwendet werden, daß die untersuchten Stichproben aus verschiedenen Romanen einer einzigen lexikalischen Grundgesamtheit angehören. Dazu ist der Umfang des nichtübereinstimmenden Teils des Wortschatzes abzuschätzen.

Wenn man die Wortrepertoires beider Stichproben vereinigt, dann nimmt der Wortschatz gegenüber den beiden einzelnen Stichproben nur um den nicht gemeinsamen Anteil der Wortrepertoires zu. Wären die Wortrepertoires beider Stichproben identisch, so würde sich der Wortschatz bei ihrer Vereinigung nicht vergrößern. Die Zunahme des Wortschatzes bei der Vereinigung von Stichproben erfolgt nur auf Grund dessen, daß die Wortrepertoires der Stichproben unterschiedlich sind. Berechnet man den Zuwachs nach Formel (1), kann man ihn unter der Voraussetzung der zu prüfenden[2] Nullhypothese schätzen. Der untere Teil von Tabelle 1 (von N = 11000 an abwärts) stellt sowohl die fortlaufenden theoretischen Vorhersagen nach Formel (1) als auch die Ergebnisse der tatsächlichen Vereinigung beider Stichproben dar. Spalte DVM enthält die Wortschatzwerte, die man erhält, wenn man zu Stücken des Texts DVM Stücke des Texts DS hinzufügt, während in Spalte

DS fortlaufend Stücke des Texts DS mit gleichgroßen Stücken des Texts DVM vereinigt wurden. (Es ist anzumerken, daß dies mit der in unserer Arbeit verwendeten Technik der Kartennumerierung relativ einfach ging, ohne daß man den Wortschatz der wachsenden Teilstichproben tatsächlich aussondern und kombinieren mußte). Der Grad der Übereinstimmung zwischen theoretischen und empirischen Werten ist im unteren Teil von Tabelle 1 im ganzen praktisch genau so gut wie im oberen Teil.

In Orlov (1978) ist angegeben, wie man ein Konfidenzintervall für die theoretische Vorhersage des Wortschatzes aus empirischen Daten unter der Annahme, daß die Stichproben aus einer homogenen Grundgesamtheit stammen, berechnet. Eine Berechnung nach den Daten aus DVM ergibt, daß die Diskrepanz zwischen theoretischem und empirischem Wortschatz beim Umfang N = 20000 nicht größer als 197 Wörter sein dürfte (Konfidenzniveau 0.95). Die beiden beobachteten Differenzen (5443 - 5370 = 73 und 5450 - 5443 = 7) liegen beträchtlich darunter. Die Berechnung auf der Basis von DS ergibt entsprechende Werte.

Der Wortschatzzuwachs (gleich der Anzahl nichtübereinstimmender Wörter in den beiden Stichproben) liegt also im Bereich der Zufallsstreuung des Zuwachses, der zu erwarten ist, wenn beide Stichproben aus einer homogenen Grundgesamtheit stammen.

Die Resultate sprechen dafür, daß die Stichproben aus den beiden verschiedenen Romanen sowohl hinsichtlich ihres gemeinsamen lexikalischen Teils als auch hinsichtlich des Umfangs des nichtgemeinsamen Teils als Stichproben aus einer einzigen lexikalischen Gesamtheit angesehen werden können.

Eine derartige Ähnlichkeit von zwei verschiedenen Texten ist anscheinend eine Ausnahme, selbst wenn man nur Texte von K. Gamsachurdija vergleicht. Die Erzählung "Der Fotograf" z.B. zeigte ein steileres Anwachsen des Wortschatzes (siehe Tabelle 3).

Tabelle 3

N	v	v(N/Z)
1000	637	612
2000	1141	1073
3000	1535	1482
4000	1899	1860
4381	2002	1990*

* Mit einem Sternchen wird der theoretische Wortschatzwert gekennzeichnet, der an den entsprechenden empirischen Wert "angepaßt" wurde, wobei Z nach Formel (2) berechnet wurde.

Außerdem ergab sich bei der Vereinigung des ganzen Texts "Der Fotograf" mit einer gleichgroßen Stichprobe aus DVM (beide vom Umfang N_o = 4381) in der Gesamtmenge von 8762 Wortverwendungen ein Wortschatz von 3434 Wörtern, was in Tabelle 4 mit den Repertoires anderer Texte von entsprechendem Umfang verglichen wird.

Tabelle 4

N	F	DVM	DS	F+DVM	DVM+F
4381	2002	1842	1842	2002	1842
8762	3360*	3042	3055	3434	3434
Zuwachs	1358*	1200	1213	1432	1592

* Mit einem Sternchen sind die nach Formel (1) berechneten theoretischen Werte des Wortschatzzuwachses gekennzeichnet.

Wie aus der Tabelle ersichtlich, ist der Wortschatzzuwachs (letzte Zeile der Tabelle) in den Fällen der Kombination der Stichproben merklich größer als innerhalb jedes einzelnen untersuchten Textes. Man kann daher die Erzählung "Der Fotograf" schon nach diesen Daten nicht als eine Stichprobe aus derselben lexikalischen Gesamtheit ansehen, aus

der die Romane DVM und DS entnommen sind. Dieser Schluß wird anschaulich gestützt durch die Daten in Tabelle 5.

Tabelle 5

Wort	Übersetzung	DVM P %	DS P %	F P %	P max / P min	P max / P min
ar	nicht	0.69	0.85	0.55	1.55	1.23
iqo	sein (Verb)	0.96	0.78	0.52	1.84	1.23
am	diesen	0.51	0.48	0.32	1.59	1.06
da	und	4.98	3.38	2.46	2.02	1.48
misi	sein (Pron.)	0.70	0.32	0.57	2.19	2.18
thavisi	sein (refl.)	0.26	0.32	0.57	2.19	1.23
ese	dieser	0.61	0.58	0.25	2.44	1.05
thvali	Auge	0.52	0.16	0.66	4.12	3.25
magram	aber, sondern	0.43	0.51	0.11	4.64	1.19
igi	er	0.09	0.51	0.20	5.67	5.67
khali	Frau	0.07	0.13	0.89	12.70	1.86
mephe	Zar	1.04	0.71	-	∞	1.50
giorgi	Georgij	0.91	0.45	-	∞	2.02
photographi	Fotograf	-	-	3.48	∞	-
mamamze	Mamamze	1.59	-	∞	∞	∞
davithi	David	-	0.74	-	∞	∞

In der Tabelle sind die Häufigkeiten der häufigsten Wörter (mit relativer Häufigkeit von 0.5% an aufwärts) aufgeführt. In diesen Bereich fielen 7 bis 10 Wörter aus jedem Text, wovon aber nur drei allen drei Texten gemeinsam waren. Insgesamt gab es 16 Wörter, die in wenigstens einem der drei Texte mit einer Häufigkeit von mindestens 0.5% vorkamen. Zu jedem dieser Wörter ist die Auftretenshäufigkeit in jedem Text angegeben. Ferner wurde der Quotient aus der maximalen und der minimalen dieser Häufigkeiten berechnet. Dann wurden die Wörter nach der Größe dieser Quotienten angeordnet, d.h. die Liste zeigt die Wörter in der Reihenfolge abnehmender Häufigkeitsstabilität (und nicht abnehmender Häufigkeit).

Nur bei den drei ersten Wörtern schwanken die Häufigkeiten innerhalb des 95%-Konfidenzintervalls. Bei den übrigen Wörtern sind die Häufigkeitsschwankungen zwischen den drei Texten erheblich stärker. In der letzten Spalte wurden die

Häufigkeitsschwankungen nur zwischen den Stichproben aus DVM und DS berechnet. Bei diesem Vergleich gingen nur von sieben Wörtern die Häufigkeitsschwankungen über das Konfidenzintervall hinaus (Wörter durch Sternchen gekennzeichnet). Drei dieser Wörter sind Eigennamen. Die Häufigkeitsstabilität der häufigsten Wörter ist also beim Wechsel zwischen DVM und DS wesentlich größer.

Die statistische Ähnlichkeit der lexikalischen Stichproben aus den beiden Romanen legt den Gedanken nahe, daß der Autor beim Schreiben gleichsam von ein- und demselben Realitätsmodell ausgegangen ist. Damit ist nicht einfach die offensichtliche Tatsache gemeint, daß er in beiden Romanen dasselbe Altgeorgien beschrieben hat. Im gemeinsamen Wortschatz der Stichproben befinden sich sehr viele Wörter mit engem Anwendungsbereich, z.B.: amala - Gefolge (10 mal in der Stichprobe aus DS und 13 mal in der Stichprobe aus DVM), brdzola - Kampf (17, 11), qodoli - Turm (8, 8), darbazi - Saal (21, 21), daphdaphi - Trommel (3, 2), zari - Glockenton (4, 5), thathbizi - Beratung (6, 4), keisari - Caesar (20, 11), muzaradi - Helm (6, 11), mchedari - Reiter (23, 30), usw. Jedes dieser Wörter kann nur zur Beschreibung eines beschränketen Bereichs von Situationen verwendet werden. Es ist bezeichnend, daß von allen eben angeführten Wörtern nur zwei in der Erzählung "Der Fotograf" vorkamen, nämlich 'Saal' und 'Glockenton' (je einmal). Daß sich die Verwendungshäufigkeiten solcher Wörter in den Romanen nicht erheblich unterscheiden, spricht dafür, daß auch die entsprechenden Situationen in annähernd gleichen Verhältnissen in ihnen dargestellt werden. Das ist ganz plausibel, denn beide Romane beschreiben das Leben der Herrscher Georgiens und ihrer engsten Umgebung. Man könnte sich aus derselben Epoche auch einen Roman vorstellen, dessen Held, sagen wir, ein armer Bauer wäre (im Geiste der naturalistischen Romane von Zola; allerdings gäbe es praktisch kein Material zu einem solchen Roman, wie K. Gamsachurdija im Nachwort zum ersten Band des "David, der Baumeister"

anmerkt). Von einem solchen Helden bis zu einem Zaren wäre es ein weiter Weg, und das Wort 'Zar' käme im ganzen Roman vielleicht dreimal vor. Unser angenommener Roman bezöge sich zwar auch auf dieselbe Realität wie die hier untersuchten Werke von Gamsachurdija, aber es läge ihm ein anderes Realitätsmodell zugrunde. Man kann ein Realitätsmodell als eine minimal strukturierte Vorstellung über das zukünftige Werk ansehen, aus der sich allmählich solche Kategorien wie Sujet, Fabel, Komposition herauskristallisieren.

Die Existenz von Realitätsmodellen ließ sich bisher nur aus den Erinnerungen von Autoren über die allmähliche Verfertigung und Entwicklung des Plans zu ihrem Werk belegen. Ein Realitätsmodell erschien da als der unfertige Keim des zukünftigen Werks, der mit dem Heranreifen des Planes von selbst verschwindet, sich in die Konkretisierungen der Bilder, des Sujets, der Komposition, usw. auflöst.

Der außergewöhnliche Fall der Romane Gamsachurdijas, die eine Art "statistischer Zwillinge" bilden, zeigt aber, daß ein Realitätsmodell auch als ein werkunabhängiges Gebilde existieren kann. Als ein vom konkreten Thema, den Protagonisten, der Handlung unabhängiges Gebilde kann das Realitätsmodell gleichsam aus einem Werk in ein anderes übergehen, d.h. den Prozeß der Erzeugung des neuen Textes steuern, den Autor zur Schaffung eines neuen Themas, neuer Protagonisten und eines neuen Handlungsaufbaus inspirieren.

Die vergleichende statistische Analyse des Wortschatzes von literarischen Texten eines Schriftstellers gewährt also einen Einblick in die Geheimkammern seines schöpferischen Laboratoriums.

Anmerkungen

[1]Das übliche Studentsche Kriterium (vgl. ein beliebiges Lehrbuch der mathematischen Statistik) besteht darin, daß der Unterschied zwischen den beobachteten Zufallsvariablen x und y, wenn sie aus derselben lexikalischen Gesamtheit stammen (und annähernd normal verteilt sind), $t_\beta \sigma_{x\pm y}$ nicht überschreiten soll; dies ist identisch mit der Aussage, daß der Unterschied zwischen den beobachteten Größen signifikant (nicht zufällig) ist, wenn

$$\frac{x - y}{\sigma_{x\pm y}} > t_\beta \ , \qquad (A)$$

wo $\sigma_{x\pm y}$ die Standardabweichung der Summe oder des Unterschiedes der Zufallsvariablen und t_β die tabellierte, vom Konfidenzkoeffizienten β abhängige Größe ist (speziell, wenn in unserem Fall $\beta = 0.95$, dann ist $t_\beta = 1.96$). In unserem Fall, wenn x und y absolute Häufigkeiten aus Stichproben mit demselben Umfang sind, d.h. $x = F_1$, $y = F_2$, kann man annehmen, daß die Varianzen dieser Größen mit den Größen selbst identisch sind und die Varianz ihres Unterschieds der Summe der Varianzen gleich ist, d.h. $F_1 + F_2$, woraus $\sigma_{x\pm y} = F_1 + F_2$. Setzt man diesen Ausdruck in (A) ein, so erhält man

$$\frac{|F_1 - F_2|}{\sqrt{F_1 + F_2}} > t_\beta = 1.96.$$

[2]Die Nullhypothese besagt, daß die Stichproben aus den zwei Romanen als Stichproben aus ein- und derselben lexikalischen Grundgesamtheit zu betrachten sind.

DIE MELODISCHE ELEMENTAREINHEIT

M.G. Boroda

1. Die melodische Elementareinheit, das "formale Motiv" (F-Motiv), wurde in Boroda (1973) beschrieben. Die Untersuchungen in Boroda (1974) zeigten die Fruchtbarkeit dieser Einheit bei der Analyse der statistischen Struktur musikalischer Texte mit unterschiedlichen Stilen. Wie sich in Boroda (1973) zeigte, kann das F-Motiv auch für die Untersuchung weiterer Organisationsebenen musikalischer Texte und möglicherweise ebenso für die stilistische Analyse angewendet werden. Deshalb ist es zweckmäßig, den "physischen Sinn" des F-Motivs und seine Wechselbeziehungen zu den in der Musikwissenschaft üblichen Einheiten genau zu erörtern. Dies zu tun, ist das Ziel der vorliegenden Untersuchung.[1)]

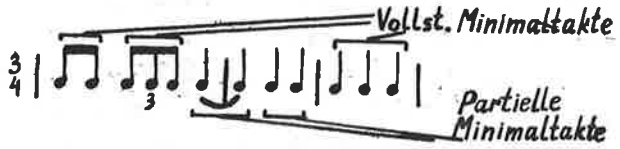
2. Das F-Motiv wird in Boroda (1973) durch die Beschreibung metrorhythmischer Gruppen, die unter den Bedingungen einer Taktordnung entstehen, definiert. Es gibt vier solcher Gruppen.

2.1. Gruppen gleichlanger Töne:

2.1.1. Vollständiger Minimaltakt: Eine Folge zweier gleichlanger Töne (in einer dreiteiligen Gruppe sind es drei Töne), wobei der erste metrisch stärker als die übrigen ist.[2)]

2.1.2. Partieller Minimaltakt: Ein Ton (bzw. zwei gleichlange Töne in einer dreiteiligen Gruppe), der mit dem folgenden Ton keinen vollständigen Minimaltakt bildet (weil dieser eine andere Länge hat oder metrisch stärker ist als der erste).

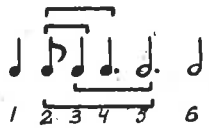
Kurz gesagt, der vollständige Minimaltakt ist eine nach dem Schema der Duole oder Triole gebildete metrische Gruppe; der partielle Minimaltakt ist ein Teil dieser Gruppe. Beispielsweise:



Wie aus der Definition und dem Beispiel ersichtlich, ist der vollständige Minimaltakt eine Verallgemeinerung des einfachen Taktes, während der partielle Minimaltakt eine Verallgemeinerung des (einfachen) Auftaktes ist. Im Folgenden wird der Minimaltakt (vollständig oder partiell) mit M bezeichnet.

2.2. Gruppen unterschiedlich langer Töne:

2.2.1. Anwachsende Sequenz: Eine Folge von Tönen, in der jeder folgende Ton länger ist, als der vorangegangene, z.B.:

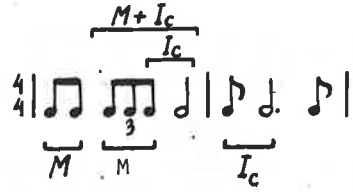


Wie aus dem Beispiel ersichtlich, kann die anwachsende Sequenz vollständig sein, d.h. sie ist in keiner weiteren anwachsenden Sequenz enthalten, wie etwa 2-3-4-5 als Ganzes im obigen Beispiel, und unvollständig, d.h. sie ist in irgend einer weiteren anwachsenden Sequenz enthalten: so ist z.B. 3-4 enthalten in 2-3-4, 3-4-5 und 2-3-4-5. Im weiteren werden wir nur noch auf die vollständigen anwachsenden Sequenzen eingehen (die mit I_c bezeichnet werden).

Wir bemerken, daß das Prinzip des Aufbaus anwachsender Sequenzen eine Verallgemeinerung des in der Musiktheorie bekannten Prinzips der Tonstrebung zum nächsten Ton mit größerer Länge darstellt (vgl. Mazel'/Cukkerman 1967).

2.3. Mischgruppen

2.3.1. Minimale metrorhythmische Gruppe (im Folgenden als $M+I_c$ bezeichnet): Verbindung des Minimaltaktes M mit der vollständigen anwachsenden Tonsequenz I_c, die beim letzten Ton dieses Minimaltaktes anfängt. (Es ist klar, das $M+I_c$ metrorhythmisch untrennbar sind, sonst würden M oder I_c zerstört.) Beispielsweise:



Wie aus diesem Beispiel ersichtlich, können der Minimaltakt M und die vollständige anwachsende Sequenz I_c einen Teil von $M+I_c$ darstellen, sie können aber auch selbständig, d.h. abgeschlossen, dastehen, falls hinter dem letzten Ton von M kein Ton mit größerer Länge folgt. Im weiteren werden wir nur noch abgeschlossene M und I_c betrachten (bezeichnet als M* und I_c*). Es bleibt zu beachten, daß $M+I_c$ eine Verallgemeinerung des in der Musik als elementare metrorhythmische Gruppe bekannten "Gesamtrhythmus" ist (vgl. Mazel'/Cukkerman 1967).

3.1. Wie aus den Definitionen in § 2 ersichtlich, sind M*, I_c* und $M+I_c$ untereinander alternativ und erschöpfen alle Varianten elementarer metrorhythmischer Tongruppierungen im Rahmen einer Taktordnung und unter dem Prinzip der Strebung eines kürzeren Tones zum nachfolgenden längeren Ton.

Das F-Motiv wird mittels M*, I_c*, und $M+I_c$ folgendermaßen definiert: Das Formalmotiv (F-Motiv) ist der im Rahmen von M*, I_c* oder $M+I_c$ stehende Abschnitt der Melodie.

3.2. Aus 3.1 geht hervor, daß das F-Motiv eine formal bestimmte Einheit ist. Gleichzeitig werden in das F-Motiv mittels M*, I_c* und M+I_c alle elementaren (unter bestimmten Umständen) Tongruppierungen eingeschlossen. Deshalb kann die melodische Linie eines musikalischen Textes mit einer Taktordnung von Anfang bis Ende lückenlos in F-Motive segmentiert werden.[3)]

Das Schema eines Zerlegungsalgorithmus', der die Melodielinie sukzessiv von links nach rechts abtastet, sieht folgendermaßen aus:

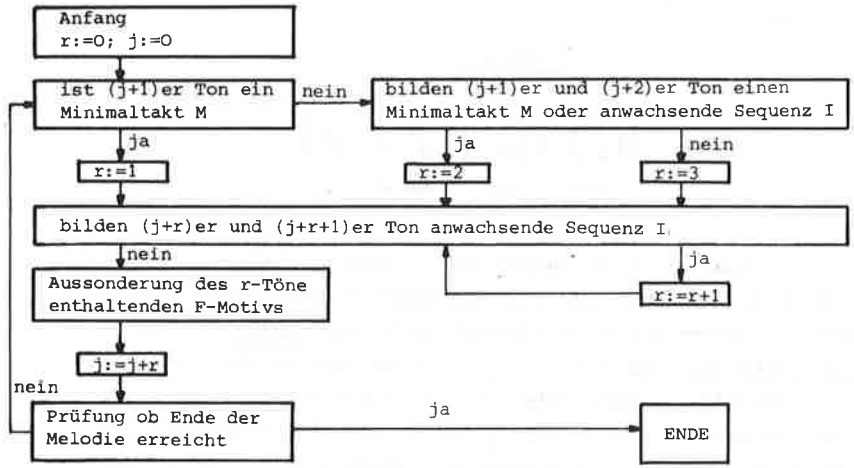


Abb. 1. Das Blockschema des Algorithmus zur Segmentierung einer Melodie, die eine Taktstruktur hat, in F-Motive als ein sukzessives, tonweises Durchmustern dieser Melodie.
j = laufende Nummer des Tones im Text
r = laufende Nummer des Tones im F-Motiv, das gebildet wird
: = zeigt an, daß eine Variable den angegebenen Wert annimmt

Im Block "Ende erreicht?" zeigt sich, ob noch unberücksichtigte Töne in der melodischen Linie übrig geblieben sind, oder ob die Segmentierung in F-Motive abgeschlossen ist.

3.3. In Abb. 2 bringen wir Beispiele der Segmentierung von Melodieabschnitten in F-Motive

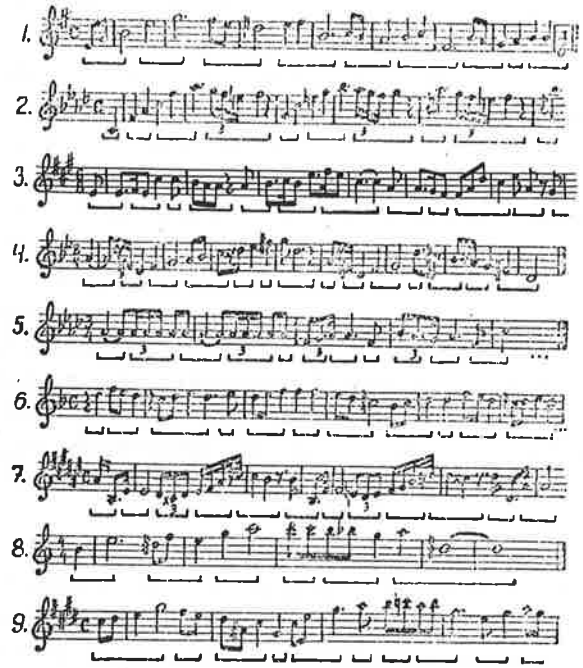


Abb. 2. Segmentierung von Melodien unterschiedlicher Stile in F-Motive (die F-Motive werden mit ⌊‾‾‾⌋ gekennzeichnet)

1. Mozart, Sonate für Geige und Klavier (1. Teil)
2. Beethoven, Sonate op. 2, Nr. 1 (1. Teil)
3. Schubert, Frühlingstraum
4. Chopin, Mazurka op. 24, Nr. 1
5. Čajkovskij, Symphonie Nr. 4 (1. Teil, Einführung)
6. Rachmaninov, Konzert Nr. 3 (1. Teil)
7. Skrjabin, Etüde op. 8, Nr. 12
8. Prokof'ev, Vergänglichkeit Nr. 13
9. Šostakovič Preludia op. 34, Nr. 10

4.1. Von der Definition unter § 3.1 und den Beispielen ausgehend, betrachten wir jetzt das F-Motiv.

4.1.1. Das F-Motiv ist generell: Da das F-Motiv nur mit den Gesetzmäßigkeiten der metrorhythmischen Organisation verbunden ist, die in der Musik verschiedener Stilrichtungen klar zutage tritt, kann es im Prinzip für die Analyse einer melodischen Struktur eines beliebigen Musiktextes mit Taktordnung benutzt werden. Dabei ist eine unterschiedliche Effektivität des F-Motivs im Rahmen unterschiedlicher Stile möglich, jedoch ist eine Bestimmung der stilistischen "Anwendbarkeitsgrenzen" apriori kaum möglich. Wie die Untersuchung der Rekurrenzen des F-Motivs in musikalischen Texten unterschiedlicher Stile gezeigt hat (vgl. Boroda 1974), tritt das F-Motiv als eine aktuelle Elementareinheit in Scarlattis Sonate, Beethovens Rondo, Levitins Sonatine usw. auf. Die Wiederholungsstruktur des F-Motivs folgt in allen untersuchten Texten einheitlichen Gesetzmäßigkeiten. Dies erlaubt uns, das F-Motiv als eine metastilistische Elementareinheit zu betrachten.

4.1.2. Das F-Motiv ist elementar: Aus der Definition und den Beispielen ist ersichtlich, daß das F-Motiv metrorhythmisch elementar ist. Nach der Segmentierung der melodischen Linie in F-Motive zerstört eine weitere Aufteilung ihre metrorhythmische Struktur, da ja für diese Struktur Gruppierungen gleichlanger Töne in Quasitakte (in unserer Terminologie Minimaltakte) und das Prinzip der Tonstrebung zum darauffolgenden längeren Ton (Prinzip der Bildung anwachsender Sequenzen) relevant sind. Das F-Motiv ist ein "metrorhythmisches Atom".

4.1.3. Das F-Motiv ist klein: Das F-Motiv ist nicht nur eine elementare, sondern auch, urteilt man nach den Beispielen, eine hinreichend kleine Einheit. Wie aus Abb. 2 ersichtlich, können die F-Motive aus 1-5 Tönen bestehen. Sogar in kurzen musikalischen Texten enthält die melodische Linie ziemlich viele F-Motive (z.B. in der Prélude op. II, Nr. 2 von Skrjabin sind es mehr als 100 F-Motive, im zweiten der Petrarca-Sonette von Liszt etwa 300). Auf der melodischen Ebene eines musikalischen Textes größeren Umfangs kommen im Durchschnitt etwa 1000 - 1500 F-Motive vor. Wegen der genannten Eigenschaften ist das F-Motiv eine effiziente Einheit für die statistische Analyse musikalischer Texte.

4.1.4. Das F-Motiv hat variable Länge: Wie aus den Beispielen ersichtlich, können die F-Motive sogar innerhalb einer und derselben melodischen Linie aus unterschiedlich vielen Tönen bestehen. Die Längenvariabilität folgt direkt aus der Definition des F-Motivs, da sowohl M* als auch I_c* und M+I_c längenvariabel sind. M* kann 1,2 oder 3, I_c* kann 1,2,3..., usw. Töne enthalten. Längenvariabilität bei gleichzeitig eindeutiger Bestimmbarkeit - das ist eine außerordentlich wichtige Eigenschaft des F-Motivs; denn die Einheiten mit konstanter Länge (Intervalle und ihre Folgen) segmentieren einen musikalischen Text in unnatürlicher Weise, während die bisher in der Musiktheorie bekannten Einheiten von variabler Länge (Motiv, Teilmotiv) weder eindeutig definiert sind, noch einheitlich behandelt werden (vgl. Anm. 1).

Es ist interessant, daß die von Definition 3.1 nicht festgelegte Obergrenze der F-Motiv-Länge faktisch doch scharf bestimmt ist. Wie unsere Untersuchungen musikalischer Texte mit unterschiedlichen Stilen (mit einem Gesamtumfang von mehr als 30000 F-Motiven) gezeigt haben, bestand das längste F-Motiv aus 5 Tönen.[4] Diese außerordentliche Stabilität (deren Ursache aber ziemlich rätselhaft ist) ist eine gute Bestätigung der These, daß das F-Motiv im Grunde eine metastilistische Einheit ist. Es ist

möglich, daß dies mit dem Kurzzeitgedächtnis zusammenhängt: 4 Längen + 5 Höhen im 5-Ton-F-Motiv stimmen anscheinend mit der "Konstante 7±2" überein. Es ist jedoch schwierig, ernstzunehmende Konsequenzen zu ziehen.

Es ist ebenfalls bemerkenswert, daß die mittlere Länge eines F-Motivs im musikalischen Text fast eine Konstante ist und, wie die ersten Beobachtungen zeigten, daß ihre Größe der mittleren Silbenlänge mancher Sprachen (gemessen in Anzahl der Laute) nahe kommt; (vgl. Fucks 1957; Gačečiladze & Cilosani 1971).

4.1.5. Beziehung zur metrorhythmischen Struktur der Melodie.

Aus der Definition des F-Motivs und der Abb. 1 ist ersichtlich, daß die Segmentierung der melodischen Linie eines musikalischen Textes in F-Motive in strenger Übereinstimmung mit der metrorhythmischen Struktur dieser melodischen Linie durchgeführt wird, während sich die Abgrenzung eines konkreten F-Motivs nach der Struktur des betreffenden Abschnitts der Linie richtet. Diese grundsätzliche Lokalität verleiht dem F-Motiv eine beträchtliche "stilistische Flexibilität", d.h. die Fähigkeit, in unterschiedlichen Stilen als eine effiziente Elementareinheit zu funktionieren. Im Unterschied zu Einheiten mit festgelegter Länge wird das Kriterium zur Segmentierung in F-Motive dem Text nicht von außen aufgezwungen, sondern ist mit der inneren Mikrostruktur des Textes eng verbunden.[5)]

4.2. Das F-Motiv und die in der Musikwissenschaft üblichen Einheiten (Motiv, Teilmotiv)

Die Beispiele in Abb. 2 zeigen, daß das F-Motiv eine natürliche "atomare" Elementareinheit ist - nichtsdestoweniger könnte es auf den ersten Blick als eine zu kleine Einheit erscheinen. Es wäre sehr interessant, von diesem Standpunkt aus die

Korrelation der F-Motive mit den in der Musikologie bekannten melodischen Elementareinheiten, dem Motiv und dem Teilmotiv, zu untersuchen. Ein derartiger Vergleich ist, puristisch gesehen, nicht erlaubt, da das Motiv und das Teilmotiv in der gegenwärtigen Musikologie weder formal definiert sind noch eine einfache Interpretation haben (vgl. Anm. 1). D.h. man kann die Definition des F-Motivs und des Motivs (Teilmotivs) nicht vergleichen. Aber man kann die Segmentierungen in Motive oder Teilmotive, wie sie in musikologischen Arbeiten angegeben sind, vergleichen. Einige Vergleiche sind in der Abb. 3 dargestellt.

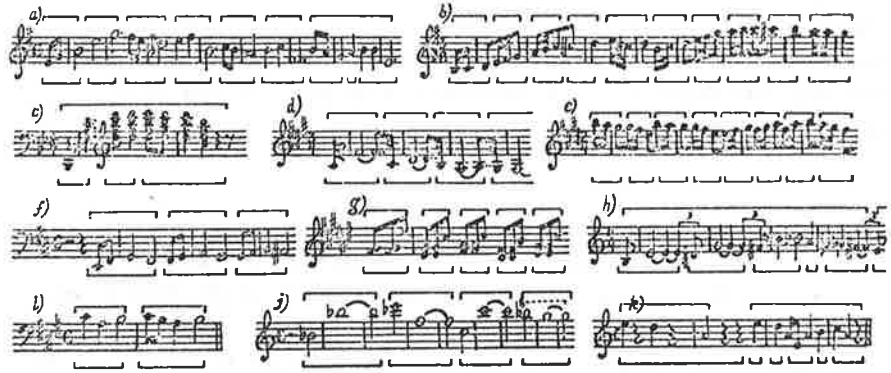


Abb. 3. Vergleich der F-Motive mit Motiven und Teilmotiven, wie sie in neueren musikologischen Arbeiten segmentiert wurden.
Motive (bzw. Teilmotive) werden mit ⌐——¬ (bzw. ⌐- - -¬) F-Motive mit ⌊——⌋ abgegrenzt. Die Segmentierung in Motive (Teilmotive) wurde aus Katuar (1936), Mazel' & Cukkerman (1967, Buckoj (1948), Tjulin (1974), Nikolaeva (1973) u.a. übernommen.

(a) Mozart, Sonate für Geige und Klavier (1. Teil), (b) Beethoven, Sonate op. 14, Nr. 2 (3. Teil), (c) Beethoven, Sonate op. 106 (1. Teil), (d) Schumann, Sonate Nr. 1, fis-moll (1. Teil), (e) Paganini, Caprice Nr. 9, (f) Čajkovskij, Symphonie Nr. 6 (1. Teil, Einführung), (g) Skrjabin, Etude op. 8, Nr. 1, (h) Skrjabin, Extasis poem, (i) Glazunov, Sonate für Klavier, (j) Šostakovič, Symphonie Nr. 5 (1. Teil), (k) Kabalevskij, Sonatine für Klavier.

Die Beispiele in Abb. 3 zeigen, daß das F-Motiv mit dem Motiv und dem Teilmotiv übereinstimmen kann. Diese Übereinstimmungen folgen nicht aus der Definition in § 3.1, sie sind aber auch nicht zufällig: Wie bereits gesagt wurde, sind alle logischen Blöcke des F-Motivs Verallgemeinerungen von elementaren metrorhythmischen Konstruktionen, die in der Musikwissenschaft bekannt sind. Insbesondere verallgemeinert $M+I_c$ eine in der Musik sehr verbreitete Konstruktion, nämlich der "Gesamtrhythmus". Im Ganzen gesehen unterscheidet sich das F-Motiv merklich vom Motiv, ja sogar vom Teilmotiv, durch seine Fähigkeit, "elementar" und vom Umfang her "klein" zu sein (die Motive können auch sehr groß sein und sich über mehrere Takte erstrecken, vgl. Mazel'/Cukkerman 1967, Tjulin 1974), und auch dadurch, daß es eine lückenlose Textsegmentierung von Anfang bis Ende zuläßt, während das Motiv und das Teilmotiv dies prinzipiell nicht zulassen (vgl. Tjulin 1974).

4.2.1. Im Vergleich des F-Motivs mit dem Motiv und dem Teilmotiv fällt auf, daß sich in der Konstruktion des F-Motivs die Prinzipien des Jambus und Choreus gegenseitig nivellieren: der vollständige Minimaltakt wird bestimmt als Choreuskonstruktion (er beginnt auf einem starken, bzw. quasi-starken Taktteil), die anwachsende Tonsequenz wird als jambische festgelegt (es besteht eine Analogie zwischen den Sequenzen "kurzer Ton" - "längerer Ton" und "metrisch schwacher Ton" - "metrisch starker Ton"). Von $M+I_c$ werden sowohl das eine, als auch das andere Prinzip verwendet. Ein ähnlicher Typus der gegenseitigen Nivellierung von Jambus und Choreus läßt sich auch bei Motiven beobachten, wie sie in einer Reihe von modernen Arbeiten segmentiert worden sind; aber dort zeigt sich die Nivellierung auf einer weniger elementaren Ebene, als beim F-Motiv.

4.2.2. Wie man sieht, stimmen das F-Motiv und das Motiv (Teilmotiv) in einigen Punkten klar überein (beide basieren auf dem Metrorhythmus, beide sind von variabler Länge usw.). In konkreten Fällen können ein F-Motiv und ein Motiv auch zusammenfallen. Dies alles erlaubt uns, das F-Motiv als eine Einheit vom Typus des Motivs anzusehen (evtl. kann die Klasse dieser Einheiten noch vergrößert werden).[6] Es ist jedoch klar, daß das F-Motiv kein "formalisiertes Äquivalent" des Motivs bzw. Teilmotivs sein kann. Wie bereits gesagt, ist das F-Motiv wesentlich elementarer als diese Einheiten. In diesem Zusammenhang ergibt sich natürlich die Aufgabe, Regeln für die "Integration" von F-Motiven zu größeren natürlichen Gebilden, etwa von der Größenordnung der Phrasen, zu definieren. Eine Diskussion dieser im allgemeinen sehr komplizierten Frage würde über den Rahmen der vorliegenden Arbeit hinausgehen.[7]

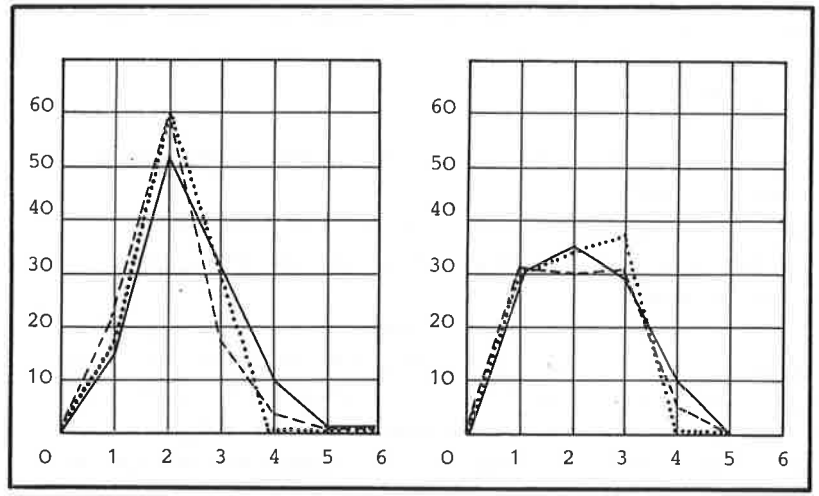


Abb. 4. Die Verteilung der Länge der F-Motive in homophonen und polyphonen Texten unterschiedlicher Stilrichtungen

Links (homophone Texte):
a) G. Tartini: Sonate für Violine und Klavier (durchgehende Linie);
b) W. Mozart: Eine kleine Nachtmusik (gestrichelte Linie);
c) S. Prokof'ev: Sonate für Solovioline (punktiert).

Rechts (polyphone Texte):
a) C. Monteverdi: Madrigal Sofra tenere herbetta (durchgehende Linie);
b) W. Mozart: Fuge d-moll (gestrichelte Linie);
c) D. Šostakovič: Präludium und Fuge Nr. 9 (punktiert).

5. Eine Untersuchung der Verteilung der Länge des F-Motivs anhand von verschiedenen musikalischen Texten zeigte, daß diese Verteilung in musikalischen Texten mit derselben Faktur (homophon bzw. polyphon) trotz unterschiedlicher Stilrichtungen dieselbe Form hat.[8)] Die homophonen und die polyphonen Texte unterscheiden sich jedoch voneinander in dieser Hinsicht, wie aus der Abb. 4 ersichtlich.

6. Fazit

6.1. In dieser Arbeit wurde eine formal bestimmte elementare Melodieeinheit, das F-Motiv, beschrieben. Weiter wurden ein Algorithmusschema und Beispiele für die Segmentierung einer melodischen Linie im musikalischen Text mit einer Taktordnung in F-Motive vorgestellt.

6.2. Die Analyse des F-Motivs zeigte anhand seiner Definition und der Beispiele:

a) Die ein F-Motiv aufbauenden Blöcke sind Verallgemeinerungen von metrorhythmischen Konstruktionen, die in der Musiktheorie bekannt sind.

b) Das F-Motiv ist eine kleine Elementareinheit mit veränderlicher Länge. Sie ist über die verschiedenen Stilformen generalisierbar. Die Segmentierung der melodischen Linie in F-Motive ist gänzlich durch ihre metrorhythmische Struktur bestimmt.

c) Das F-Motiv kann auf verschiedenen Ebenen der Analyse der Melodielinie eines musikalischen Textes verwendet werden, insbesondere ist es zu statistischen Untersuchungen musikalischer Texte geeignet.

d) In Sonderfällen stimmt das F-Motiv mit dem Motiv oder Teilmotiv überein. Im allgemeinen stellt es eine elementare Einheit dar. Seine Übereinstimmung mit dem Motiv in einigen wesentlichen Strukturmerkmalen erlaubt uns, das F-Motiv als eine Einheit vom Typus des Motivs anzusehen.

e) Einige quantitative charakteristische Merkmale des F-Motivs rücken das F-Motiv in die Nähe einer bekannten linguistischen Einheit, der Silbe.

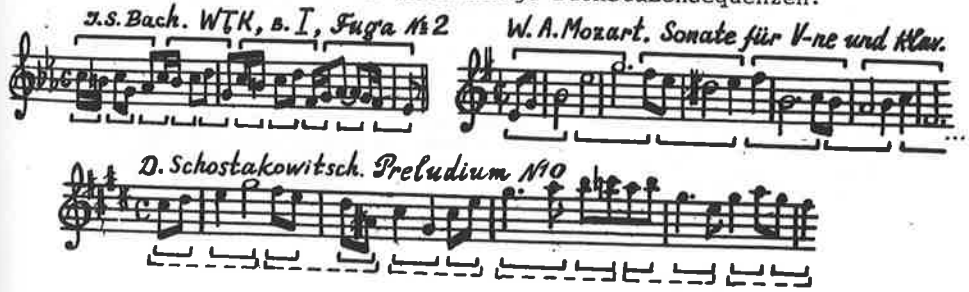
ANMERKUNGEN

1 Zum besseren Verständnis der Ziele dieser Arbeit scheint es uns zweckmäßig, dem Text einige Bemerkungen über das eigentliche Problem der Elementareinheit in der Musik in Form eines kurzen "Vorwortes" vorauszuschicken.

Bei der Untersuchung - besonders einer quantitativen - des musikalischen Textes stößt man ständig an ein durchaus nicht einfaches Problem: wie kann man diesen Text (z.B. auf der melodischen Ebene) in elementare Segmente - elementar in einem bestimmten Sinn - zerlegen? Die Komplexität der Frage besteht darin, daß auf der einen Seite diese Segmente eindeutig definiert werden müssen und die Zerlegung des Textes in diese Elemente völlig frei von subjektiven Entscheidungen sein muß; auf der anderen Seite sollen diese Segmente hinreichend natürlich klingen und die Zerlegung soll irgendwie auf den strukturellen Charakteristiken des gegebenen Textes basieren. Man konnte diese beiden Forderungen lange Zeit nicht in Einklang bringen, zumal wenn man noch einen dritten Umstand berücksichtigen wollte: für quantitative (speziell statistische) Untersuchungen war es äußerst wichtig, den Text lückenlos von Anfang bis Ende in Elemente des gegebenen Typs zu zerlegen, sonst ging ein Teil des Textes für die Analyse verloren.

In der Musikologie waren bislang zwei Typen elementarer (im Sinne von hinreichend kleiner) Einheiten bekannt:

(a) Einerseits melodische Intervalle (definiert als eine Sequenz zweier benachbarter Töne), Sequenzen von 2,3,... Intervallen hintereinander, Töne und ihre Sequenzen, Akkorde u.a. Alle diese Elemente hatten einen unverkennbaren und großen Vorzug: sie wurden völlig klar definiert. So besteht beispielsweise kein Zweifel darüber, was ein melodisches Intervall oder ein Takt usw. ist. Aus diesem Grund waren sie für die quantitative Musikologie sehr attraktiv und wurden in derartigen Untersuchungen breit und systematisch verwendet (vgl. z.B. die bekannten Arbeiten von Fucks und seiner Schule). Jedoch ist aus vielen Beispielen ersichtlich, daß die Zerlegung des musikalischen Textes - speziell seiner Melodie - in Intervalle, ihre Sequenzen usw. völlig unnatürlich, "gewaltsam" ist. Dies betrifft besonders Melodien, die unterschiedlich lange Töne benutzen. Eine solche Zerlegung ähnelt in einem bestimmten Sinn der Segmentierung eines literarischen Textes in gleichlange Buchstabensequenzen:



Die Ursache dieser Unnatürlichkeit liegt darin, daß das Intervall, eine Sequenz von Intervallen, einen Takt usw. mit den rhythmischen Strebungen von Tönen unterschiedlicher Länge, mit den rhythmischen Unterbrechungen, die bei der Gliederung der Melodie eine äußerst wichtige Rolle spielen, nicht verbunden sind. Das melodische Intervall, die Sequenz von Intervallen usw. stellen aus einer allgemeinen Sicht Einheiten mit fixierter Länge dar (das Intervall besteht aus einer fixierten Zahl von Tönen; der Takt, beim gegebenen Maß, aus einer fixierten Zahl von metrischen Taktteilen), sie sind für eine natürliche Zerlegung des musikalischen Textes zu grob, zu unflexibel.

(b) Andererseits sind in der Musikologie gut bekannt die Einheiten mit variabler Länge: das Motiv und das Teilmotiv. Aufgrund der Beispiele, die man in verschiedenen musikologischen Arbeiten findet, kann man schließen, daß diese Einheiten den Text (die Melodie) viel natürlicher zerlegen als die Einheiten mit fester Länge; viele Autoren sind der Meinung, daß das Motiv eine musikalisch sinnvolle Einheit ist, die mit lokalen metrorhythmischen, harmonischen u.a. Tendenzen in der Melodie, mit intonationalen Mikrowiederholungen zusammenhängt. Von diesem Gesichtspunkt aus müßten das Motiv und das Teilmotiv ideale Einheiten für eine strukturelle und quantitative Analyse der Musik darstellen. Dies wäre der Fall, wenn es nicht den Umstand gäbe, daß diese Einheiten strukturell nicht exakt definiert sind und in der Musikologie uneindeutig behandelt werden. Hier sind einige Beispiele solcher Definitionen:

> (i) Ein Motiv ist eine rhythmische, durch einen Hauptakzent verbundene Gruppe von Tönen, die gleichzeitig die kleinste sinnvolle Einheit darstellt. ... Das Motiv oder die Phrase zerfallen manchmal in kleinere melodisch-rhytmische Gruppen von der Länge eines Taktteils (Sposobin 1972: 66-69).
>
> (ii) Ein Motiv ist der kleinste Teil eines musikalischen Gedankens, der eine bedeutungsvolle (semantische) und gleichzeitig konstruktive Einheit darstellt. ... Kleine Teile, die sich innerhalb des Motivs klar trennen lassen oder sich absondern, heißen Teilmotive (Mazel', Cukkerman 1967: 552, 560).
>
> (iii) Motiv (von lat. movere = bewegen): das kleinste sinnvolle Element musikalischer Aussage. Es ist charakteristisch geprägt und voll innerer Spannung, die zu Weiterentwicklung und Wachstum drängt (Seeger, Bd. 2, 1966: 119).

Wie man sieht, erlauben die angeführten Definitionen (ähnliche kann man in verschiedenen Lehrbüchern der musikalischen Form und in Musikwörterbüchern finden) nicht, Regeln der Segmentierung einer Melodie in Motive oder Teilmotive zu formulieren. Außerdem, wie in mehreren neueren Arbeiten bemerkt wird, kann in der Regel eine Melodie, besonders eine lange, nicht lückenlos von Anfang bis Ende in Motive oder Teilmotive zerlegt werden (vgl. z.B. Tjulin 1974: 48). Deswegen können das Motiv und das Teilmotiv bei der quantitativen

Analyse des musikalischen Textes als Ganzem oder seiner Organisation auf der melodischen Ebene als Einheiten nicht benutzt werden.

In dieser Situation erscheint es selbstverständlich, daß man darauf verzichtet, das Motiv oder das Teilmotiv auf eine mehr formale Weise zu definieren, als es die klassische Musikologie getan hat und daß man statt dessen eine neue elementare Einheit (oder sogar Einheiten) aufgrund der Organisationsprinzipien des musikalischen Textes definiert, Einheiten, die allgemein sind und einen metastilistischen Charakter haben. Eine solche Einheit, die wir bereits früher definiert haben (vgl. Boroda 1973), wird in der vorliegenden Arbeit besprochen.

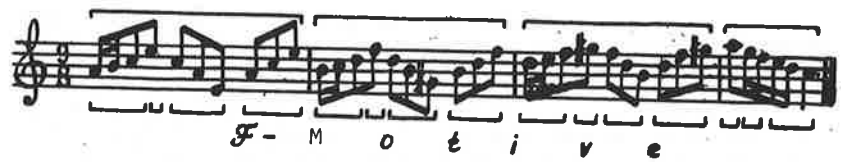
2 Zur Methode des Vergleichs gleichlanger Töne bezüglich ihrer metrischen Stärke ("Gewicht") vgl. unsere Dissertation (Boroda 1979), wo ein axiomatisches System der Taktmetrik in der Musik, das auf den in der Musikologie bekannten Gesetzmäßigkeiten basiert, dargelegt wird (dieses System kann auch als eine Formalisierung des in der Musikologie bekannten Systems der Taktmetrik betrachtet werden). In dieser Arbeit wird strikt nachgewiesen, daß in einer Melodie mit Taktordnung zwei beliebige gleichlange Töne nach ihrer "metrischen Stärke" mit eindeutigem Resultat verglichen werden können.

3 Hierbei kann man die Länge der Pause zur Länge des vorhergehenden Tones hinzuzählen, oder man kann sie ganz weglassen. (Es versteht sich, daß die "Pausenregelung" bei der Segmentierung im voraus abgesprochen werden muß.) Wegen des Phänomens der "Tonspur" ist die erste Regelung vorzuziehen.

4 Weitere Untersuchungen der Charakteristiken des F-Motivs in musikalischen Texten unterschiedlicher Stile (vgl. Boroda 1977, 1979, u.a.) bestätigen diese Schlußfolgerung.

5 Es ist interessant, daß man das F-Motiv auch bei der Analyse der melodischen Phrasierung, einer für den Interpreten wichtigen Aufgabe, verwenden kann. Wir bringen zwei Beispiele:

(a) J.S. Bach, Preludium Nr. 20 aus dem "Wohltemperierten Klavier B. 1" (Thema)



Dieses Thema kann leicht in Takte aufgegliedert werden, und die Gliederung ergibt natürliche Segmente. Wenn man aber auf dieser verhältnismäßig groben Ebene analysiert, so kann man leider nur nacheinanderfolgende Wiederholungen, deren erste Note jedesmal erhöht wird (a-h-d), und eine Wellenförmigkeit der Melodie beobachten. Wenn man diese Melodie auf der Ebene der F-Motive untersucht, so kann man ihr "Intonationsleben" detallierter beobachten und eine Reihe von interessanten Erscheinungen entdecken.

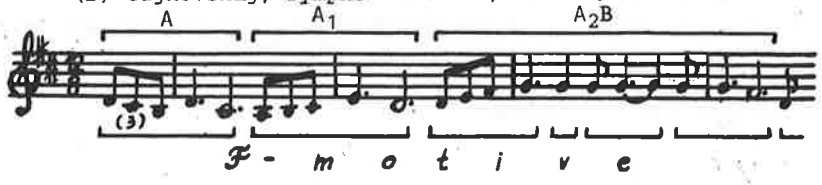
Erstens, die lokalen "Maxima" - die höchsten Punkte - der Melodie werden zum ersten Mal in F-Motiven, die aus 1 Ton bestehen, erreicht; diese F-Motive bilden eine Linie "e-f-gis-a". Auf diese Art enthält die Melodie zwei Schichten: die erste wird durch eine fallende-steigende Bewegung gebildet und enthält Triolen und M/I$_6$-Gruppen, die zweite Schicht ist "e-f-gis-a". Mit anderen Worten, die Analyse in F-Motive enthüllt eine latente Polyphonie der Melodie - eine wichtige Tatsache für den Interpreten.

Zweitens, die Linie "e-f-gis-a", gebildet durch die F-Motive von jeweils 1 Ton, kann als die "Entfaltung" der anfänglichen Sequenz von Sekunden "a-h-c" betrachtet werden, und dies weist auf die starke intonationale Einheitlichkeit der Melodie des gegebenen Themas hin.

Drittens, die Analyse in F-Motive enthüllt drei unterschiedliche - und sich widersprechende! - Tendenzen in der Konstruktion des thematischen Kerns selbst: Aufstieg in Sekunden, kontrastierender Abstieg in drei Tönen und wieder Aufstieg, diesmal in drei Tönen. Dieser Konflikt, dieser "Kampf der Tendenzen" bildet eine äußerst wichtige Bewegungskraft bei der Konstruktion des Thematas und letzten Endes des ganzen Präludiums.

Schließlich, viertens, das Anwachsen der F-Motivlänge am Ende des Themas - die Folge: F-Motiv mit 1 Ton, mit 2 Tönen, mit 3 Tönen (vgl. die letzten drei F-Motive) - erzeugt mit reinen Segmentierungsmitteln einen "Bremseffekt". Es ist offensichtlich, daß bei der Analyse auf einer gröberen Ebene (z.B. der Taktebene) die beschriebenen Effekte nicht zu entdecken sind.

(b) Čajkovskij, Symphonie Nr. 5, Teil 2, 1. Thema



Diese Melodie kann man auch in drei verhältnismäßig große Segmente zerlegen. Die Analyse auf dieser Ebene erlaubt einen bestimmten "Summeneffekt" des dritten Segments zu entdecken. Das ist im Grunde alles. Was liefert die Analyse in F-Motive?

Die ersten zwei F-Motive fallen mit den Segmenten "A" und "A$_1$" zusammen, das dritte gibt "A$_1$" verkürzt wieder. Es ist aber keineswegs nur eine Verkürzung: die "weibliche" (schwache) Endung von "A$_1$" wird durch eine "männliche" (starke) ersetzt. Diese plötzliche Abbremsung scheint ein Vorbote des nachfolgenden Fallens der Melodie zu sein. In der Tat steigt die Melodie nicht mehr über das im dritten F-Motiv erreichte "g". Weiter, es gibt hier drei F-Motive bestehend aus 1, aus 2 und zum Schluß aus 3 Tönen. Als ob die Melodie allmählich "zu Atem kommen würde". Dies ist sehr wichtig, denn ein solcher Verlauf - allmählicher "Atemgewinn" nach einer "Erstarrung" (auf "g") - vertieft den lyrischen Charakter dieses Themas, verleiht ihm eine ungewöhnliche Eindringlichkeit. Schließlich, das F-Motiv aus 3 Tönen bringt erneut die weibliche Endung, die in den ersten zwei F-Motiven vorhanden war und im dritten annulliert wurde. Die Analyse in F-Motive erlaubt also, in dieser Melodie eine Reprisenstruktur zu entdecken. Man kann leicht sehen, daß beim Übergang auf eine Ebene von größeren Elementen viele von den erwähnten Erscheinungen einfach verwischt werden.

6 Es ist interessant, daß das Motiv und das Teilmotiv lückenlos in F-Motive segmentiert werden können (im Falle der Übereinstimmung mit dem F-Motiv - in 1 F-Motiv), wie aus der Abb. 3 ersichtlich. Mit anderen Worten, man kann annehmen, daß das Motiv und das Teilmotiv als eine Sequenz von F-Motiven dargestellt werden können. Wie unsere Untersuchungen zeigen, gilt dies für eine große Zahl von Beispielen (Ausnahmen sind selten). Diese Tatsache kann man leider theoretisch nicht beweisen (nicht einmal für eine bestimmte Klasse von Motiven und Teilmotiven), da weder das Motiv noch das Teilmotiv formal definiert sind. Eine Überprüfung dieser Annahme aufgrund eines umfangreichen experimentellen Materials scheint jedoch äußerst interessant und wünschenswert, da es einen wichtigen Schritt bei der Ausarbeitung des "Problems der Einheiten" in der Musik bedeutet.

7 Vgl. Kap. 8 in diesem Band.

8 Ein Melodieabschnitt eines homophonen Textes deckte sich mit der Hauptstimme dieses Textes. Im polyphonen Text stellte der Melodieabschnitt die Gesamtheit der Melodielinien der Stimmen dar (es wurden Texte untersucht, in denen alle Stimmen melodisch gleichberechtigt sind). Jede der Stimmen wurde unabhängig in F-Motive zerlegt. Danach wurden die Gesamtheiten der F-Motive vereinigt und stellten somit einen polyphonen Text auf der Ebene des F-Motivs dar.

ZUR BESTIMMUNG EINER PHRASENÄHNLICHEN MELODISCHEN INFORMATIONSEINHEIT IN DER MUSIK

M.G. Boroda

Das Problem der Informationseinheiten ist eines der wesentlichen Probleme bei der quantitativen Analyse eines musikalischen Textes. Die üblichen musikalischen Einheiten ergeben im Grunde entweder eine unnatürliche Aufteilung des Textes (wie Intervall, Intervallfolge) - vergleichbar der Unterteilung eines literarischen Textes in Segmente zu n Buchstaben - oder sie werden nicht einheitlich verstanden [wie die in der Musikwissenschaft bekannten Begriffe Teilmotiv, Motiv, Phrase, vgl. (Tjulin (1974); Mazel' & Cukkerman(1967)].[1] Als besonders schwierig erweist sich das Problem der Einheiten bei Aufgaben, die eine Aufteilung des musikalischen Textes in relativ g r o ß e Segmente erfordern. Die Regeln für eine solche Segmentierung müssen natürlich möglichst allgemein und somit für einen großen Kreis von Stilrichtungen anwendbar sein.

In der vorliegenden Arbeit wird der Versuch unternommen, eine relativ große Melodieeinheit des Phrasentyps zu definieren, wobei von der Formalisierung der r h y t h m i s c h e n Aufteilungsprinzipien eines musikalischen Textes ausgegangen wird. Es wird die "Rhythmische Phrase" (R-Phrase) beschrieben, die in einer Melodie auf der Basis der Längenbeziehungen benachbarter Töne bestimmt wird.

Die Definition der R-Phrase stützt sich auf drei Gruppierungsprinzipien von Tönen hinsichtlich ihrer Länge;

1) Ein Ton, der einem anderen längeren Ton vorausgeht, lehnt sich an diesen an und bildet zusammen mit ihm eine rhythmisch in sich geschlossene Gruppierung (Mazel' & Cukkerman 1967).

Dieses Prinzip wirkt auch über 1 Ton und erzeugt Konstruktionen des Typs 

(Wir bezeichnen es als Prinzip 1A).

2) Eine Sequenz von kürzeren Tönen (d.h. kürzer als der Ton, der vor der Sequenz steht) trennt gewöhnlich das Fragment, in dem sie vorkommen, vom vorausgehenden Fragment ab (Mazel' & Cukkerman 1967).

3) In einer Folge von drei Tönen mit abnehmenden Längen schließt sich der zweite Ton eher dem ersten als dem dritten Ton an.[2]

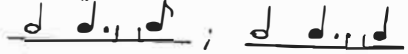
Diese in der Musikwissenschaft bekannten Prinzipien wurden offensichtlich nicht einheitlich definiert. Benutzt man sie jedoch als "qualitative" Prinzipien der rhythmischen Gliederung der Melodie, der rhythmischen Gruppierung ihrer Töne in relativ großen Segmenten, so kann man ganz strikte Regeln dieser Gruppierung festsetzen. Solche Regeln über Einschließung oder Ausschließung eines Tones der Melodie in das nächste zu bildende (relativ große) Segment, werden unten untersucht. Es ist leicht zu sehen, daß diese Regeln, ausgehend von den Prinzipien 1-3, auf der Analyse der Verhältnisse der Tonlängen in Paaren (Prinzip 1), in Dreiergruppen (Prinzip 2 und 3) und Vierergruppen von Tönen (Prinzip 1A) basieren müssen. Wir definieren diese Regeln folgendermaßen:

O1. Wir sprechen von einer R-Verkettung eines gegebenen Tones (im folgenden bezeichnet mit *) mit dem vorausgehenden Ton (Vorgänger), wenn:

a) der gegebene Ton nicht kürzer als sein Vorgänger ist,

usw., oder

b) die Differenz der Länge des Vorgängers und des gegebenen Tones nicht größer ist als die Differenz der Länge des gegebenen Tones und seines Nachfolgers, wobei beide Differenzen größer als Null sind,

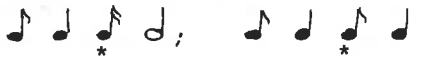


oder

c) die Differenz der Länge des Vorgängers und des gegebenen Tones kleiner ist als die Differenz der Länge des Vorvorgängers und des Vorgängers, wobei beide Differenzen größer als Null sind

oder

d) die Differenz der Länge des Nachfolgers und des gegebenen Tones nicht kleiner ist als die Differenz der Länge des Vorgängers und des Vorvorgängers, wobei beide Differenzen größer als Null sind und der gegebene Ton kürzer ist als sein Vorgänger



(Wir setzen als Konvention fest, daß vor dem ersten (entsprechend hinter dem letzten) Ton einer Melodie ein Ton steht, dessen Länge einen Takt beträgt; wenn auf einen Ton eine Pause folgt, wird dieser Ton um den Wert der Pause verlängert.)[3)]

O2. R-Kette nennen wir eine Folge von Tönen, in der jeder Ton mit dem vorangehenden eine R-Verkettung bildet (Abb. 1).

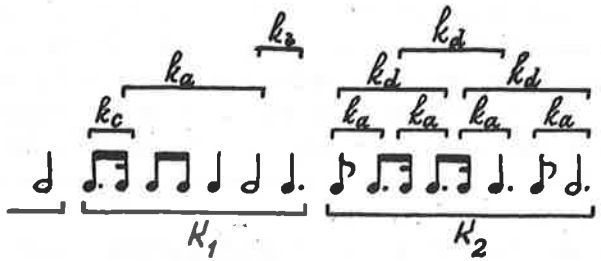


Abb. 1. Die Richtung von R-Ketten in Tonsequenzen. Die R-Ketten werden mit ⌜¬ und L_ gekennzeichnet. Mit k_i wird eine R-Kette bezeichnet, die aufgrund der i-ten Regel (i = a,b,c,d) der R-Verkettung entstand. (Zum Unterschied zwischen den k und K Ketten s. Text)

O3. Wir sagen, daß eine R-Kette A in einer R-Kette B enthalten ist, wenn jeder Ton von A auch in B gehört (z.B., k_c und k_b sind enthalten in K_1).

O4. Wir nennen eine R-Kette vollständig, wenn sie in keiner anderen R-Kette enthalten ist als in sich selbst (K_1, K_2 in Abb. 1).

Es ist offensichtlich, daß sich eine beliebige Folge von Tönen einheitlich und ohne Lücken in vollständige R-Ketten aufteilen läßt.

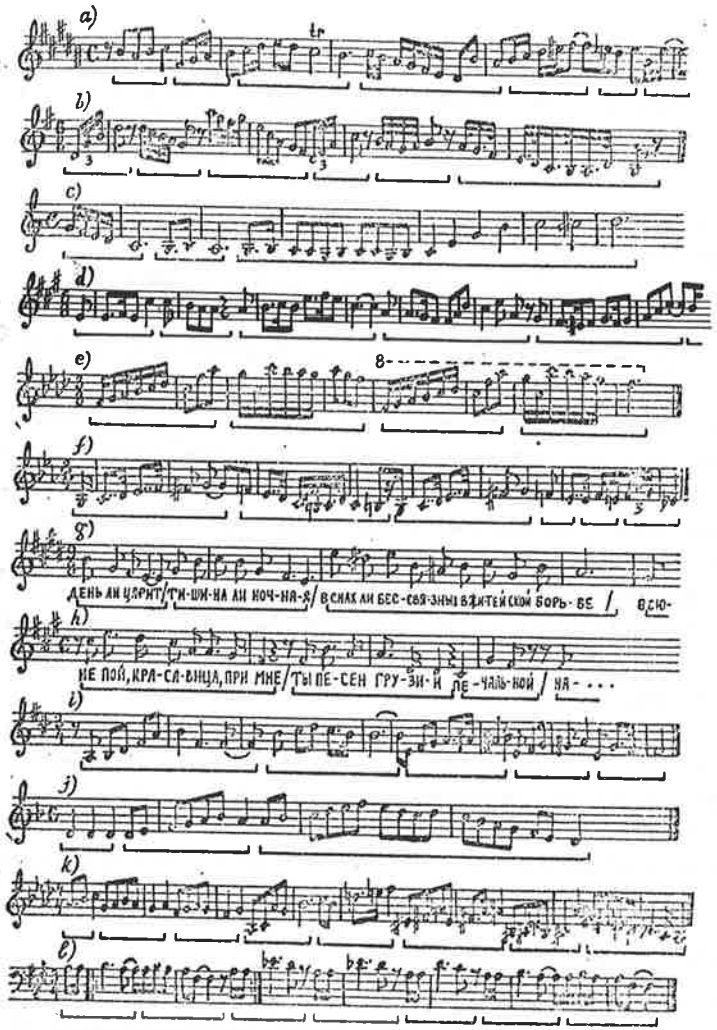


Abb. 2. Segmentierung der Melodie in R-Phrasen (|____|).

(a) Bach, WTK, Fuga Nr. 23, Bd. 1; (b) Haydn, Symphonie Nr. 96; (c) Beethoven, Symphonie Nr. 1; (d) Schubert, Frühlingstraum; (e) Schumann, Schmetterling; (f) Rossini, Moses (4. Akt); (g) Čajkovskij, Den li carit; (h) Rachmaninov, Ne poj, krasavica pri mne; (i) Glier, Romanze für Geige; (j) Prokofjev, Romeo und Julia ("Kampf"); (k) Prokofjev, Sonate für Klavier Nr. 4; (l) Šostakovič, Fuge Nr. 8, Op. 87

Jetzt können wir die R-Phrase definieren.

Definition: Eine rhythmische Phrase (R-Phrase) ist ein Segment einer Melodie, das innerhalb einer vollständigen R-Kette liegt[4] (Abb. 2).

Wie man aus den Beispielen ersieht, ist die R-Phrase eine relativ große und natürliche melodische Einheit.[5]

Die relativ großen Ausmaße der R-Phrasen führen unweigerlich zu der Frage nach den Wechselbeziehungen der R-Phrase mit kleinen, nach rhythmischen Merkmalen bestimmten melodischen Einheiten.

Wir wollen unter diesem Gesichtspunkt den Zusammenhang der R-Phrase mit der metro-rhythmisch elementaren Einheit "Formalmotiv" (F-Motiv) erörtern, die von Boroda (1973) für statistische Untersuchungen von musikalischen Texten definiert wurde.[6]

In Boroda (1973) wird das F-Motiv definiert als Abschnitt einer Melodie, der in den Grenzen einer von vier elementaren metro-rhythmischen Gruppen (vollständiger Minimaltakt, partieller Minimaltakt, anwachsende Tonfolge, minimale metrorhythmische Gruppierung) liegt.[7]

Einige Beispiele der Aufteilung eines Melodieabschnittes in F-Motive und R-Phrasen sind in Abb. 3 dargestellt. Wie aus diesen Beispielen ersichtlich, kann eine R-Phrase in eine Folge von F-Motiven zerlegt werden. Vergleicht man die Definitionen von R-Phrase und F-Motiv, so kann man sehen, daß dies auch im allgemeinen richtig ist. Damit gilt also die folgende Behauptung:

Eine R-Phrase ist eine Folge von F-Motiven.

Und tatsächlich, wenn $A = a_1, a_2, \ldots a_m$ und $B = b_1, b_2, \ldots b_n$ (a_i und b_i sind Töne bestimmter Länge) zwei aufeinanderfolgende R-Phrasen sind, dann ist der Ton b_1 auf jeden Fall kürzer als a_m. Aus der Definition des F-Motives geht jedoch hervor, daß ein Ton nur in dem Falle zu einem F-Motiv gehört, wenn er nicht kürzer als der vorangehende ist. Deshalb muß ein F-Motiv, das in der R-Phrase A beginnt, auch in ihr enden. Daraus folgt, daß jede R-Phrase in eine Folge von F-Motiven unterteilt werden kann.

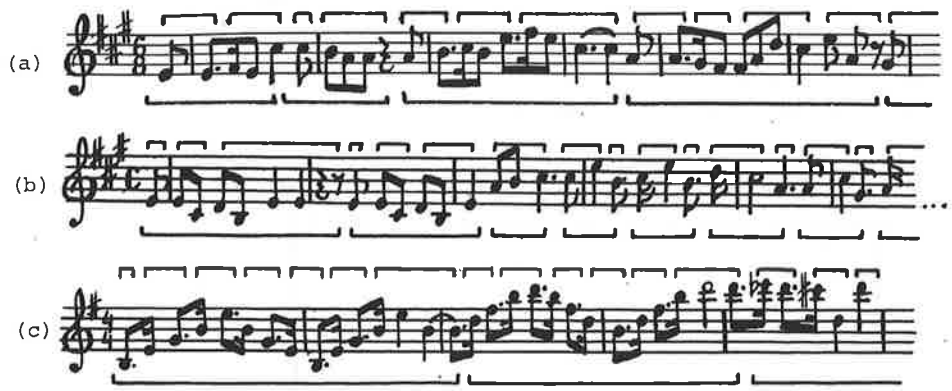


Abb. 3. Beziehung der R-Phrasen und F-Motive. Die F-Motive werden mit ⌜¬⌝, die R-Phrasen mit ⌞_⌟ gekennzeichnet
(a) Schubert, Frühlingstraum
(b) Čajkovskij, I bol'no i sladko
(c) Prokofjev, Romeo und Julia.

Ganz offensichtlich wird durch diese Tatsache, eine Hierarchie der rhythmischen Segmentierungs- und Strukturebenen eines musikalischen Textes festgesetzt sowie zwei wichtige Ebenen dieser Struktur definiert. Von besonderem Interesse ist die Festlegung der Bedingungen, unter denen F-Motive als elementare melodische Einheiten aufgrund nicht-rhythmischer Merkmale in R-Phrasen integriert werden. Die relative Homogenität der inneren Melodiestruktur von R-Phrasen zum Beispiel (jede R-Phrase nimmt gewöhnlich entweder in den Grenzen einer melodischen, z.B. einer tonleiter- oder arpeggioartigen Struktur, ab oder besteht aus einer kleinen Anzahl solcher Strukturen, die "das Ende der gegebenen mit dem Anfang der folgenden" zu einem F-Motiv verketten)

legt den Gedanken nahe, die Regeln der rein melodischen Verkettung der F-Motive zu größeren Komplexen zu bestimmen, und diese Komplexe mit R-Phrasen zu vergleichen. Dies würde einen tieferen Einblick in die Rolle der rein melodischen Prinzipien der Organisation eines musikalischen Textes gewähren.

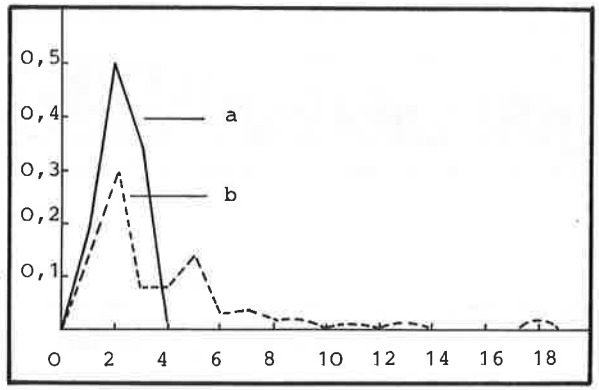


Abb. 4. Verteilung der R-Phrasenlängen
a) ČAJKOVSKIJ: Die Träne bebt (Vokalpart);
b) ČAJKOVSKIJ: Betrachtung (Violinenpart)

Unter dem Gesichtspunkt der Analyse eines musikalischen Textes auf der Ebene der R-Phrasen wäre es interessant zu untersuchen, wie die Längen der R-Phrasen (in Anzahl von Tönen oder F-Motiven) in musikalischen Texten unterschiedlicher Stilrichtungen verteilt sind. Eine vorläufige Untersuchung zeigte zum Beispiel, daß die Verteilung der Längen von R-Phrasen, gemessen in Anzahl der F-Motive, in einem Vokalwerk anders ist als in einem Instrumentalwerk (Abb. 4).

Anmerkungen

1) Ausführlich zum "Problem der Einheiten" in der Musik vgl. Boroda (1979).

2) Dieses Prinzip wurde in der Musikwissenschaft nicht explizit formuliert. Eine implizite Beschreibung (als "weibliche Kadenzform") wurde jedoch z.B. in Katuar (1934) gegeben.

3) Man sieht leicht, daß die Bedingungen (a) bis (d) der R-Verkettung ein unabhängiges System bilden, d.h. keine Bedingung kann von einer anderen abgeleitet werden. Bezeichnen wir mit t_0 die Länge des gegebenen Tones, mit t_{-1} die des Vorgängers, mit t_{-2} die des Vorvorgängers und mit t_1 die des Nachfolgers. Weiter führen wir folgende Bezeichnungen ein

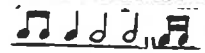
$$\begin{aligned} \Delta_0 &= t_0 - t_{-1} \\ \Delta_{-1} &= t_{-1} - t_{-2} \\ \Delta_1 &= t_1 - t_0. \end{aligned} \qquad (1)$$

Man kann die Bedingungen (a) bis (d) der R-Verkettung des gegebenen Tones mit dem Vorgänger folgendermaßen schreiben:

(a) $\Delta_0 \geq 0$
(b) $\Delta_1 \leq \Delta_0 < 0$
(c) $\Delta_{-1} < \Delta_0 < 0$
(d) $\Delta_0 < 0 < \Delta_{-1} \leq \Delta_1$.

Offensichtlich bildet keine dieser Regeln einen Spezialfall einer anderen, so daß sie ein unabhängiges System bilden.

4) Wir bringen hier eine "inhaltlich klare" - jedoch weniger formale Definition der R-Phrase:
Wir bezeichnen als R-Phrase ein Fragment der Melodie im Rahmen einer der folgenden rhythmischen Gruppierungen:
(a) Nichtabnehmende rhythmische Sequenz, in der kein Ton kürzer ist als sein Vorgänger, z.B.



(b) Sequenz mit "weiblicher Endung". Hier ist der letzte Ton kürzer als der vorletzte, der nächste Ton ist noch kürzer, aber der Längenunterschied des vorletzten und des letzten Tones ist nicht größer als der des letzten Tones und seines Nachfolgers, z.B.

(c) Sequenz mit "schwachem Anfang". Ihr erster Ton ist länger als der zweite, der Vorgänger des ersten Tones ist länger als der erste, aber die Längendifferenz des ersten

Tones und seines Vorgängers ist größer als die Längendifferenz des ersten Tones und seines Nachfolgers, z.B.

(d) Sequenz mit einer Endung des Typs  oder

 , d.h. eine Sequenz, die mit vier Tönen endet,

wobei der erste kürzer als der zweite und der dritte kürzer als der vierte ist und der Längenunterschied des vierten und des dritten ist nicht kleiner als der des zweiten und des ersten.

(e) Gemischte Sequenz, die Fragmente des Typs (a), (b), (c) oder (d) enthält.

5) Es ist interessant, daß in der Vokalmusik (Lieder, die zu dichterischen Texten komponiert wurden) die Endungen der R-Phrasen mit starken Verszäsuren übereinstimmen. Dies kann man in den Romanzen von Čajkovskij und Rachmaninov sehen (vgl. Abb. 2, wo die Verszäsuren mit "/" bezeichnet sind). Mit anderen Worten, wie eine vorläufige Analyse zeigte, fallen die R-Phrasen mit relativ beendeten Textsegmenten zusammen. Das Problem solcher Übereinstimmungen, ihrer Ausgeprägtheit und Ursachen verdient eine separate Untersuchung und ist sowohl für das bekannte Problem "Sprache und Musik" als auch für die Analyse der allgemeinen Prinzipien des musikalischen Rhythmus von Bedeutung.

In diesem Zusammenhang berühren wir noch ein Detail, das in den musikalischen Beispielen in Abb. 2 ersichtlich ist: Die R-Phrase ist eine natürliche melodische Einheit unter der Bedingung, daß in der Melodie Töne verschiedener Längen benutzt werden (zusammen mit Gruppen gleichlanger Töne), d.h. unter den Bedingungen rhythmischer Heterogenität und starker rhythmischer Tendenzen. Falls in der Melodie nur gleichlange Töne vorkommen (z.B. in Stücken des Typs von moto perpetuo u.ä.), so verliert die R-Phrase als Einheit ihre Effizienz. Dies ist aber auch natürlich, da unter diesen Umständen der elementare Rhythmus der Längenverhältnisse benachbarter Töne der Melodie seine organisierende Rolle verliert.

6) Die Untersuchung der Organisation der Wiederholung von F-Motiven in einem musikalischen Text in Boroda (1977,1979) und Boroda & Nadarejšvili & Orlov & Čitasvili (1977) zeigte, daß die Prinzipien dieser Organisation auf Texte der unterschiedlichsten Stilrichtungen zutreffen.

7) Vgl. Kap. 6 in diesem Band.

Häufigkeitsstrukturen musikalischer Texte

M.G. Boroda, Tbilisi

Die vorliegende Arbeit[1] untersucht anhand von musikalischem Textmaterial (d.h. Musikwerken als Textgebilden) aus verschiedenen Stilen die Prinzipien der Rekurrenz kleiner motivartiger melodischer Elemente im Text. Diese Organisationsprinzipien, die für Texte verschiedenster Stile *allgemeingültig* sind, werden mithilfe von Methoden beschrieben, die in der quantitativen Linguistik entwickelt worden sind. Es wird gezeigt, daß zwischen diesen Prinzipien und dem musikalischen Text *als Ganzem* ein Zusammenahng besteht und daß hier analoge Verhältnisse wie bei den Organisationsprinzipien der Rekurrenz von Wörtern in literarischen Texten vorliegen.

1.

Die Untersuchungen musikalischer Texte aus verschiedenen Stilen zeigen, daß die Wiederholung einen der Hauptfaktoren der musikalischen Formenbildung und überhaupt das wichtigste Mittel zur Gestaltung des "musikalischen Inhalts" darstellt. Das Prinzip der Wiederholung durchzieht den musikalischen Text auf sämtlichen formalen Ebenen - von ganzen Sätzen musikalischer Werke bis hin zu Intervallen und rhythmischen Einheiten.

Aus dieser Sicht kann die musikalische Form auf jeder beliebigen Ebene betrachtet werden als eine organisierte Aufeinanderfolge von "neuen", d.h. vom Textanfang an erstmalig vorkommenden Elementen und solchen Elementen, die bereits aufgetretene wiederholen. Die Erforschung der Gesetzmäßigkeiten, nach denen sie organisiert ist, stellt für die Musiktheorie eine wichtige Aufgabe dar. Wie die bisherigen Ergebnisse zeigen, ist es allerdings mit herkömmlichen musikwissenschaftlichen Methoden nur in unterschiedlichem Ausmaße möglich, die Organisation der im Text auftretenden Wiederholungen und Alternationen von *großen* Teilstücken (in der Größenordnung von Sätzen) einerseits und *kleinen* Teilstücken (vom Range eines Motivs) andererseits zu analysieren. Während für Großeinheiten viele Prinzipien ihrer Wiederholung und Alternation im Text wohlbekannt sind (teilweise sind sie in Formschemata fixiert), sind für *kleine* Elemente derartige Prinzipien noch nicht nachgewiesen.[2] Es ist unklar, ob die Rekurrenz kleiner

Elemente im musikalischen Text innerhalb des jeweiligen Stils (einer Epoche) einheitlich organisiert ist, ob sie bei jedem Komponisten individuell ausgeprägt ist, oder ob es letztlich ein für jeden Text gültiges durchgängiges Organisationsprinzip gibt. Eine Antwort hierauf wird möglich durch eine konkrete Untersuchung der Rekurrenzstruktur kleiner musikalischer Elemente in musikalischen Texten verschiedener Stile.

Für eine derartige Untersuchung ist offenbar folgendes notwendig:

a) die Abgrenzung einer bestimmten Textebene (z.B. der melodischen),
b) die Segmentierung des Textes auf dieser Ebene in Elemente eines bestimmten Typs;
c) bestimmte Kriterien zur Unterscheidung der Elemente voneinander;
d) die Auswertung des Textes auf der abgegrenzten Ebene bezüglich der Rekurrenz dieser Elemente.

Im Untersuchungsergebnis läßt sich dann jedem der ermittelten distinkten Textelemente ein bestimmter Zahlenwert zuordnen - seine Vorkommenshäufigkeit im Text. Dementsprechend kann dem ganzen Text eine Tabelle der Form

$$\begin{pmatrix} a_1, & a_2, & a_3, & \dots, & a_n \\ p_1, & p_2, & p_3, & \dots, & p_n \end{pmatrix} \qquad (1)$$

zugeordnet werden, in der p_i die Vorkommenshäufigkeit des Elementes a_i im Text ist.

Die ermittelten Daten können unter zwei Gesichtspunkten analysiert werden. Einerseits lassen sich, bei einer quantitativen Interpretation der Elemente a_i, anhand einer Tabelle der genannten Form (1) statistische Charakteristiken berechnen, und ihre Werte können dann für unterschiedliche Texte und Textgruppen verglichen werden. Das ist das Verfahren der "statistischen Musikanalyse", das bei der Untersuchung der stilistischen Besonderheiten der Werke verschiedener Komponisten durchaus effektiv angewandt werden kann (vgl. Detlovs 1968, Fucks 1975, Rojterštejn 1973). Hingegen können metastilistische Prinzipien, nach denen die Rekurrenz von Elementen im musikalischen Text organisiert ist, bei einem solchen Verfahren schwerlich festgestellt werden, denn die Vorkommenshäufigkeiten von Elementen, und somit die Werte der statistischen Charakteristiken, weichen in der Regel von Stil zu Stil erheblich voneinander ab. Darüber hinaus führt die Notwendigkeit einer zuverlässigen Bewertung dieser Häufigkeiten dazu, daß diese nicht jeweils für einzelne Texte berechnet werden, sondern für stilistisch homogene *Mengen* von Texten oder Textausschnitten, daß also die Geschlossenheit des Textes außer acht gelassen wird.

Es besteht andererseits jedoch auch die Möglichkeit, die Daten einer Tabelle (1) dahingehend auszuwerten, daß nicht Paare "Element-Häufigkeit" betrachtet werden, sondern die Häufigkeiten selbst und die Beziehungen zwischen ihnen. Dabei wird davon abstrahiert, welche Häufigkeit welchem Element entspricht. Ein derartiges Vorgehen erlaubt es, die Struktur der Elementwiederholungen im Text detailliert zu untersuchen: die Beziehungen zwischen "häufigen" und "seltenen" Elementen, zwischen rekurrenten und nichtrekurrenten Elementen etc. Hierbei entsteht nicht das Problem der zuverlässigen Bewertung der "Wahrscheinlichkeit" dieses oder jenes konkreten Elementes. Zudem können schließlich auch bei identischen Mengen von Vorkommenshäufigkeiten von Elementen diese Elemente selbst voneinander gänzlich verschieden sein. Daher ist es sehr wahrscheinlich, daß mittels der Analyse von Häufigkeitsmengen $\{p_i\}$ und der in ihnen geltenden Beziehungen *allgemeingültige Organisationsprinzipien der Elementrekurrenz im musikalischen Text als einem Ganzen* nachgewiesen werden können.

Zugunsten einer derartigen Hypothese sprechen, wenn auch indirekt, die Ergebnisse der Untersuchung eines ähnlich gelagerten Problems in der quantitativen Linguistik. Dort nämlich werden Mengen von Vorkommenshäufigkeiten von Elementen in lexikalischen Stichproben (man bezeichnet sie gewöhnlich als *statistische* oder *Häufigkeitsstruktur* einer Stichprobe) schon seit langem eingehend untersucht. Die Analyse der Häufigkeitsstruktur von Stichproben machte es möglich, Organisationsprinzipien der Wortrekurrenz nachzuweisen, die für verschiedenste Stichproben allgemeingültig sind (Frumkina 1964, Orlov 1970).

So ließ sich zum Beispiel in jeder Stichprobe eine große Anzahl wenig frequenter Wörter feststellen und demgegenüber eine geringe Anzahl häufiger Wörter. Je kleiner dabei die Häufigkeit eines

Wortes in der Stichprobe war, desto größer war die Anzahl der Wörter mit dieser Häufigkeit. Insbesondere die einmal vorkommenden, d.h. in der Stichprobe nicht wiederholten Wörter bildeten annähernd die Hälfte ihres Lexikons. Wenn man nun weiterhin die Menge $\{p_i\}$ der relativen Vorkommenhäufigkeiten von Wörtern in der Stichprobe nach abnehmenden Werten ordnete, konnte der allgemeine Verlauf der Häufigkeitsabnahme durch den Ausdruck

$$p_i = \frac{K}{(B+i)^{\gamma}}, \qquad i=1,2,3,\ldots,n \qquad (2)$$

näherungsweise berechnet werden. Hierbei ist p_i die Häufigkeit i-ter Ordnung in der geordneten Häufigkeitsmenge, n der Umfang des Lexikons der Stichprobe, und K, B und γ sind Konstanten. Den Ausdruck (2) bezeichnet man als Zipf-Mandelbrotsches Gesetz (vgl. Frumkina 1964, Orlov 1970).

In jüngeren Arbeiten (Orlov 1970, 1975) wurde gezeigt, daß derartige Gesetzmäßigkeiten an vollständigen Texten in sich abgeschlossener literarischer Werke am genauesten erfüllt werden. Dabei gilt $\gamma=1$, K und B sind Funktionen der Textlänge[3] N_o und der Frequenz des im Text am häufigsten vorkommenden Wortes p_{max}, und der ganze Ausdruck für p_i nimmt damit folgende Form an[4]:

$$p_i = \frac{\frac{1}{\ln(N_o p_{max})}}{\frac{1}{p_{max}\ln(N_o p_{max})} - 1+i} \qquad (3)$$

Der Umfang $n(N_o)$ des Lexikons des Textes, auf den das Gesetz (3) angewandt wird, und die Zahl $n_m(N_o)$ der einzelnen Wörter, von denen jedes im Text m-mal gebraucht ist, wird nach Orlov (1975) durch die Beziehungen

$$n(N_o) = \frac{N_o - \frac{1}{p_{max}}}{\ln(N_o p_{max})} \qquad (4)$$

$$n_m(N_o) = \frac{n(N_o)}{m(m+1)} \qquad (5)$$

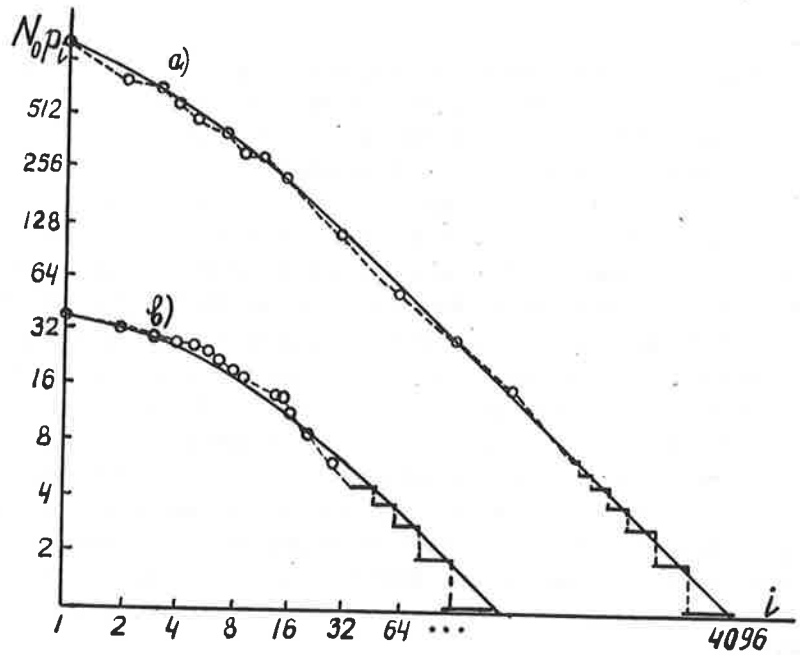


Abb. 1: Graphische Darstellung der Häufigkeitsstruktur literarischer Texte.

Experimentell ermittelte Kurve und theoretische Beschreibung nach dem verallgemeinerten Zipf-Mandelbrotschen Gesetz (3) (entnommen aus Orlov 1975)
a) A.S. Puškin, Kapitanskaja dočka (Die Hauptmannstochter)
b) altruss. Byline Avdot'ja Rjazanočka
Die durchgezogene Linie verbindet die Punkte, die gemäß der theoretischen Beschreibung berechnet wurden, die gestrichelte Linie die experimentell ermittelten. Die "Stufen" im rechten Teil der graph. Darstellung sind bedingt durch die große Anzahl wenig frequenter, d.h. einmal, zweimal etc. vorkommender, Wörter im jeweiligen Text, der bezüglich der Geltung des Gesetzes (3) untersucht wurde.

definiert. Demnach ist die Struktur der Wiederholungen von Wörtern im literarischen Text abhängig von der Länge dieses Textes und von der Frequenz des in ihm am häufigsten vorkommenden Wortes. Je län-

ger ein Text ist, desto größer ist, bei annähernd gleichen Werten für p_{max}, der Umfang seines Lexikons. Einen bedeutenden Anteil am Lexikon eines Textes stellen die wenig frequenten Wörter, d.h. einmal, zweimal, dreimal etc. vorkommende. Die Zahl der m-mal vorkommenden Wörter nimmt ab mit dem Anwachsen von m.

Wie bei Orlov (1970, 1975) gezeigt wird, beschreiben die Beziehungen (3), (4) und (5) die Struktur von Wortwiederholungen in vollständigen literarischen Texten verschiedener Stile (vgl. Abb.1). Für eine derartige Beschreibung ist die Vollständigkeit des untersuchten Textes von wesentlicher Bedeutung: für Textausschnitte und Textkorpora sind die Beziehungen (3), (4) und (5) in der Regel nicht erfüllt. So beträgt beispielsweise die Abweichung der tatsächlichen Werte für den Lexikon-Umfang aus (4) 100-150 %. Somit gehorcht die Rekurrenzorganisation von Wörtern im Text einer Reihe von allgemeinen Prinzipien und steht im Zusammenhang mit der vollen Textlänge. Die Analyse der Häufigkeitsstruktur hat es ermöglicht, diese Prinzipien nachzuweisen und sie mathematisch zu beschreiben.

Diese anhand von literarischen Texten ermittelten Ergebnisse legten den Gedanken nahe, mithilfe der Analyse von Häufigkeitsstrukturen auch die Organisation der Rekurrenz von Elementen in *musikalischen* Texten zu untersuchen, die zu diesem Zweck auf einer bestimmten Ebene jeweils in ihrer Gesamtheit, d.h. von Anfang bis Ende, betrachtet werden müßten.

2.

Wir waren also vor die Aufgabe gestellt, unter Berücksichtigung verschiedener Stile die Struktur von Elementwiederholungen im musikalischen Text zu untersuchen, und zwar auf der melodischen Ebene (in homophonen Texten anhand der "Oberstimme", in polyphonen anhand der Gesamtheit der Stimmen).

Bei einer derartigen Untersuchung tritt jedoch folgende Schwierigkeit auf. Wählt man als Elementareinheit eine Einheit mit fixierter Länge (ein melodisches Intervall, eine Aufeinanderfolge von Intervallen u.ä.), so führt das zu einer unnatürlichen Segmentierung des musikalischen Textes, die vergleichbar wäre mit der Zerstückelung eines literarischen Textes in Segmente von n Buch-

staben Länge. Einheiten mit variabler Länge hingegen, wie sie in der Musikwissenschaft schon bekannt sind (Motiv, Submotiv), sind nicht eindeutig definiert und werden uneinheitlich behandelt (vgl. Kac 1972, Tjulin 1974). Angesichts dieser Lage ergibt sich die Notwendigkeit, exakt definierte Elementareinheiten mit variabler Länge abzugrenzen, die auf bekannten musikalischen Gesetzmäßigkeiten beruhen.

Eine Einheit dieser Art ist das "formale Motiv" (F-Motiv), das wir schon in einer früheren Arbeit (Boroda 1973) definiert haben. Dabei sind wir von den in der Musikwissenschaft bekannten metrisch-rhythmischen Organisationsprinzipien taktgebundener Musik ausgegangen. Auf die Eigenschaften des F-Motivs und sein Verhältnis zu Motiv und Submotiv sind wir in einer weiteren Arbeit (Boroda 1977) auch bereits eingegangen. Im folgenden führen wir eine kurze Definition des F-Motivs an:

Als "formales Motiv" (F-Motiv) wird ein Segment der melodischen Linie bezeichnet, das innerhalb einer der nachstehenden vier elementaren metrisch-rhythmischen Gruppierungen einzugrenzen ist:

a) innerhalb eines vollständiges Minimaltaktes, d.h. von zwei bzw. im Triolenrhythmus drei gleichlangen Tönen, deren erster metrisch stärker ist als die übrigen (vgl. Abb. 2a);

b) innerhalb eines partiellen Minimaltaktes, d.h. eines Tones bzw. im Triolenrhythmus zweier gleichlanger Töne, die zusammen mit dem unmittelbar folgenden Ton noch keinen vollständigen Minimaltakt ergeben, da dieser Ton sich entweder von der Länge her unterscheidet oder metrisch stärker ist als die unmittelbar voraufgehenden (vgl. Abb. 2a);

c) innerhalb einer anwachsenden Tonfolge, d.h. einer Tonfolge, bei der jeweils der folgende Ton länger ist als der voraufgehende (vgl. Abb. 2b);

d) innerhalb einer metrisch-rhythmischen Minimalgruppierung, d.h. einer Verbindung aus einem (vollständigen oder partiellen) Minimaltakt und einer anwachsenden Tonfolge, die mit dem letzten Ton des Minimaltaktes beginnt (vgl. Abb. 2c).[6]

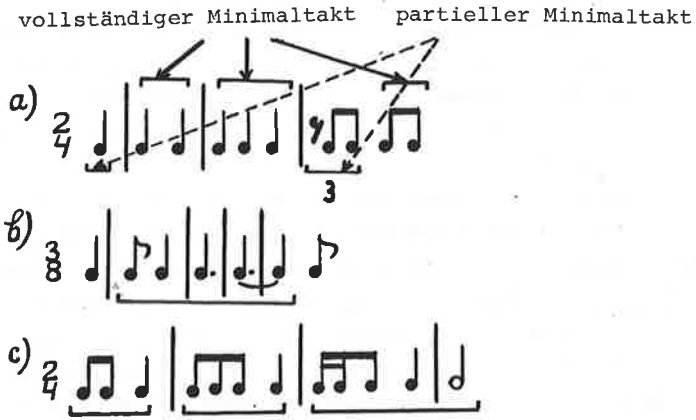
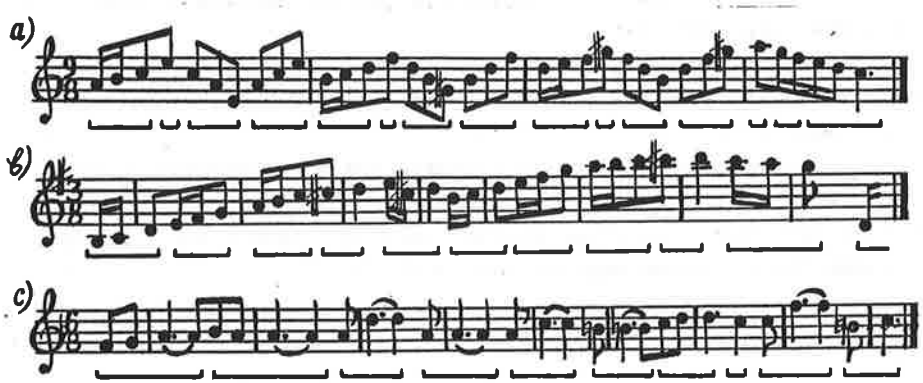


Abb. 2

Wie in den erwähnten Arbeiten (Boroda 1973, 1977) gezeigt wird, kann die melodische Linie eines musikalischen Textes mit Taktstruktur auf diese Weise von Anfang bis Ende eindeutig in F-Motive segmentiert werden. Einige Beispiele einer solchen Segmentierung zeigt Abb. 3.



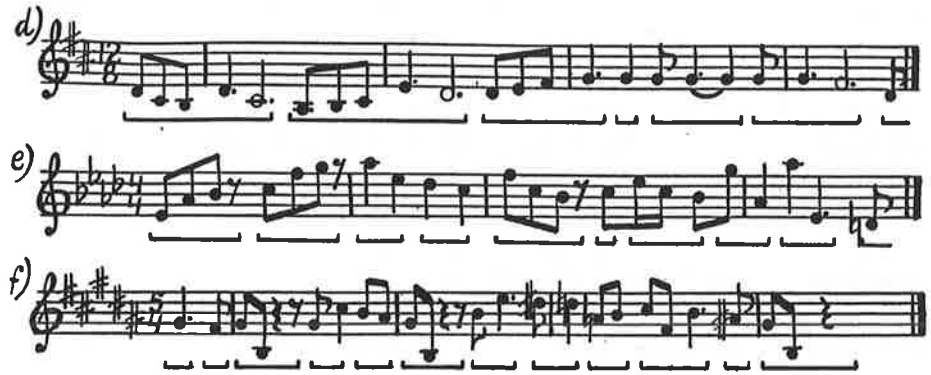


Abb. 3: F-Motiv-Segmentierung von Melodieausschnitten musikalischer Texte aus verschiedenen Stilen.

(Die einzelnen F-Motive sind durch Klammern der Art ⌐¬ kenntlich gemacht.)

a) J.S. Bach. Präludium Nr. 20 aus dem Wohltemperierten Klavier, 1. Teil BWV 865
b) L. van Beethoven. Sonate für Klavier Nr. 10
c) F. Schubert. Gretchen am Spinnrade D 118
d) P.I. Čajkovskij. Symphonie Nr. 5 e-moll op. 64
e) S.S. Prokof'ev. Romeo und Julia
f) D.D. Šostakovič. Fuge für Klavier op. 87 Nr. 12

Aus diesen Beispielen und den genannten Definitionen geht hervor, daß das F-Motiv eine Elementareinheit mit variabler Länge ist. Die Segmentierung der melodischen Linie in F-Motive steht in engem Zusammenhang mit deren metrisch-rhythmischer Struktur. Im Ausnahmefall kann ein F-Motiv sich auch mit einem Motiv oder Submotiv decken (wie in den Beispielen b) und c) der Abb. 3).

Nachdem wir nun die Elementareinheit, das F-Motiv, festgelegt haben, können wir unsere Aufgabe exakter bestimmen: sie besteht also darin, für musikalische Texte aus verschiedenen Stilen die Organisation der Rekurrenz von F-Motiven im melodischen Text-"Schnitt" zu untersuchen, d.h. die Häufigkeitsstruktur und andere Charakteristiken dieser Organisation zu analysieren.

Als Untersuchungsmaterial wurden hierzu homophone Texte (mit einer signifikanten "Oberstimme") und polyphone Texte (mit untereinander melodisch gleichwertigen Stimmen) aus dem Stil-Spektrum vom 18. bis zum 20. Jahrhundert herangezogen. Der melodische "Schnitt" eines jeden Textes, d.h. die "Oberstimme" eines homophonen bzw. die Gesamtheit der Stimmen eines polyphonen Textes, wurde in F-Motive segmentiert,[7] und es wurden für ihn folgende Charakteristiken bestimmt:

1) die Anzahl *aller* im melodischen "Schnitt" gebrauchten F-Motive (im folgenden bezeichnet als *Textlänge*);

2) die Anzahl *aller* nach einem bestimmten Kriterium *distinkten* F-Motive (im folgenden bezeichnet als *Motivinventar des Textes*[8]). Dabei galten zwei F-Motive ausschließlich für den Fall als identisch, wenn das eine aus dem anderen durch eine exakte sequenzartige Transposition (also eine Parallelverschiebung in der Tonhöhe) unter Beibehaltung der jeweiligen Tonlängen gewonnen werden konnte;[9]

3) die *Vorkommenshäufigkeiten* eines jeden distinkten F-Motivs. Die so ermittelte Menge von Häufigkeiten wurde nach abnehmenden Werten geordnet und stellte somit die *Häufigkeitsstruktur des Textes* dar.

Die Analyse der Beziehungen zwischen diesen Charakteristiken für die verschiedenen Texte ließ folgende Gesetzmäßigkeiten erkennen:

1) In jedem Text ließ sich einerseits eine geringe Anzahl häufig wiederholter F-Motive feststellen und andererseits eine beträchtliche Anzahl seltener F-Motive, die in diesem Text nur einmal, zweimal etc. vorkamen. Die Anzahl solcher m-mal vorkommenden F-Motive wuchs mit abnehmendem m.

2) Es zeigte sich, daß das Motivinventar eines Textes mit der Länge dieses Textes in Zusammenhang steht. In Texten mit unterschiedlicher Länge war auch das Motivinventar unterschiedlich umfangreich (je länger der Text, desto umfangreicher).

Diese beiden eben genannten Gesetzmäßigkeiten gelten, wie wir feststellen konnten, für vollständige Texte. Bei Textausschnitten, selbst bei in sich relativ geschlossenen wie Teilen zyklischer Werke, kamen oft Abweichungen vor: beispielsweise hatten verschieden lange Ausschnitte ein annähernd gleich umfangreiches Motivinventar und umgekehrt.[10]

Ähnliche Erscheinungen wie die oben gezeigten wurden an literarischen Texten festgestellt, für deren Wortwiederholungsstruktur das verallgemeinerte Zipf-Mandelbrotsche Gesetz galt. Deshalb kam es zu der Hypothese, daß die Organisation der Rekurrenz von F-Motiven im melodischen "Schnitt" musikalischer Texte auch für Texte verschiedenster Stile allgemeingültigen Charakter besitzt, daß zwischen ihr und dem Text als Ganzem ein Zusammenhang besteht und daß für sie das Zipf-Mandelbrotsche Gesetz gilt.

3.

Betrachten wir zunächst das jeweilige Motivinventar der untersuchten Texte und überprüfen, inwieweit für vollständige Texte und Textausschnitte das tatsächliche Motivinventar mit der theoretischen Prognose gemäß dem Ausdruck (4), d.h. der Folgerung aus dem verallgemeinerten Zipf-Mandelbrotschen Gesetz, übereinstimmt.

In der Abb. 4 sind die Graphen der Funktion (4) für P_{max}=0,04 und p_{max}=0,20 angegeben, also für den Bereich, in dem die Werte p_{max} für die untersuchten Texte und Textausschnitte liegen. (Als Textausschnitte wurden einzelne Teile der untersuchten zyklischen Werke, also der Sonaten, der Präludien und Fugen etc., herangezogen.) Als schwarze Punkte sind die jeweiligen Motivinventare der vollständigen Texte eingezeichnet, als kleine Kreise die der Textausschnitte. Wie aus der Abbildung ersichtlich wird, liegen nahezu alle schwarzen Punkte *innerhalb* des Streifens zwischen den Kurven I und II, während die kleinen Kreise größtenteils *außerhalb* dieses Streifens gelegen sind. Die Distribution der kleinen Kreise läßt nur schwer den Schluß auf irgendeine Gesetzmäßigkeit zu. Somit zeigt die Abb. 4, daß das Motivinventar von Textausschnitten im allgemeinen nicht mithilfe des verallgemeinerten Zipf-Mandelbrotschen Gesetzes in der Form der Beziehung (4) beschrieben werden kann, wohingegen das Motivinventar von vollständigen Texten durch dieses Gesetz hinreichend genau beschrieben wird.

In weiteren Einzelheiten sind die Beziehungen zwischen dem "theoretisch" bestimmten und dem "experimentell" ermittelten Motivinventar der vollständigen Texte in der Tabelle 1 (Spalten 1 - 7) dargestellt, in der für jeden Text die Werte für seine

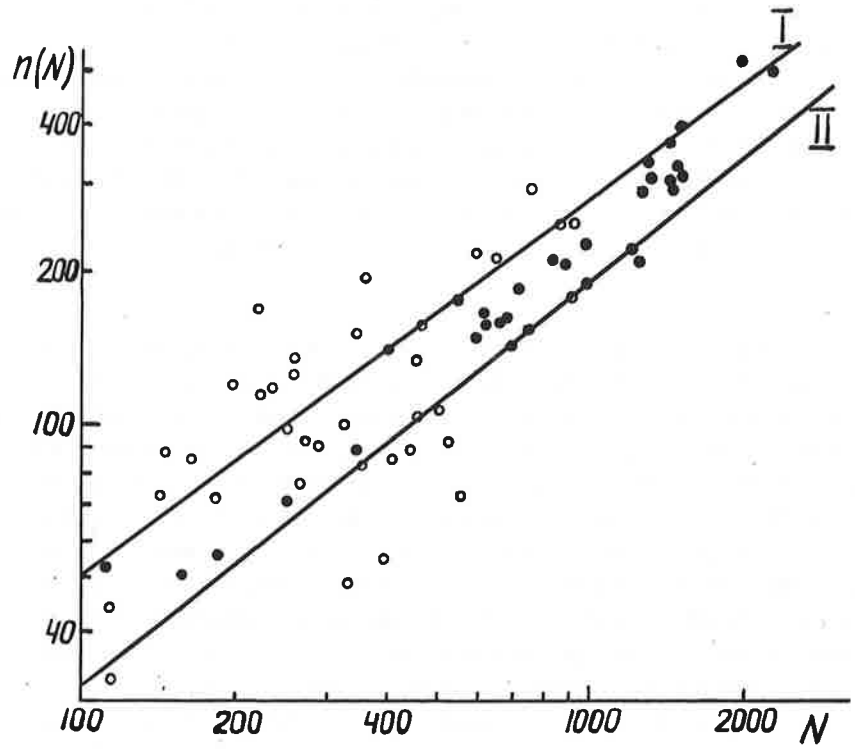


Abb.4

Länge N_o, für die maximale Vorkommenshäufigkeit eines F-Motivs p_{max}, für das tatsächliche Motivinventar n und für dessen Prognose gemäß (4), n*, angegeben sind. Wie man dort erkennen kann, sind bei den meisten Texten die Abweichungen von n gegenüber der Prognose n* nicht größer als ±20 %, und bei einer Reihe von Texten (so bei den Sonaten Nr. 1 und Nr. 9 von Scarlatti (6. und 8.), bei der Mozart-Sonate (15.), bei der Fuge von Mjaskovskij (28.) etc.) stimmt das tatsächliche Motivinventar mit seiner Prognose praktisch überein.

Demnach weisen Textausschnitt und vollständiger Text schon aufgrund dieser einen Charakteristik der Wiederholungsstruktur von F-Motiven, dem Motivinventar, einen wesentlichen Unterschied

auf. Der Vergleich zwischen dem Motivinventar vollständiger Texte und dem von Textausschnitten mit den entsprechenden Prognosen gemäß (4) zeigt, daß für die Wiederholungsstruktur von F-Motiven bei Textausschnitten insgesamt das verallgemeinerte Zipf-Mandelbrotsche Gesetz nicht gilt, wohingegen die Wiederholungsstruktur von F-Motiven im vollständigen Text diesem Gesetz gehorcht. Diese Feststellung erlaubt es, mithilfe des Gesetzes (3) zu einer detaillierteren Analyse der Organisation der Rekurrenz von F-Motiven in vollständigen musikalischen Texten überzugehen.

Wir werden die Häufigkeitsstrukturen von Texten betrachten, und wir werden den tatsächlichen Verlauf der Häufigkeitsabnahme von F-Motiven im Text mit seiner Beschreibung gemäß (3) vergleichen. In der Abb. 5.1 sind die experimentell und theoretisch ermittelten Graphen der Häufigkeitsstrukturen einiger Texte angegeben:

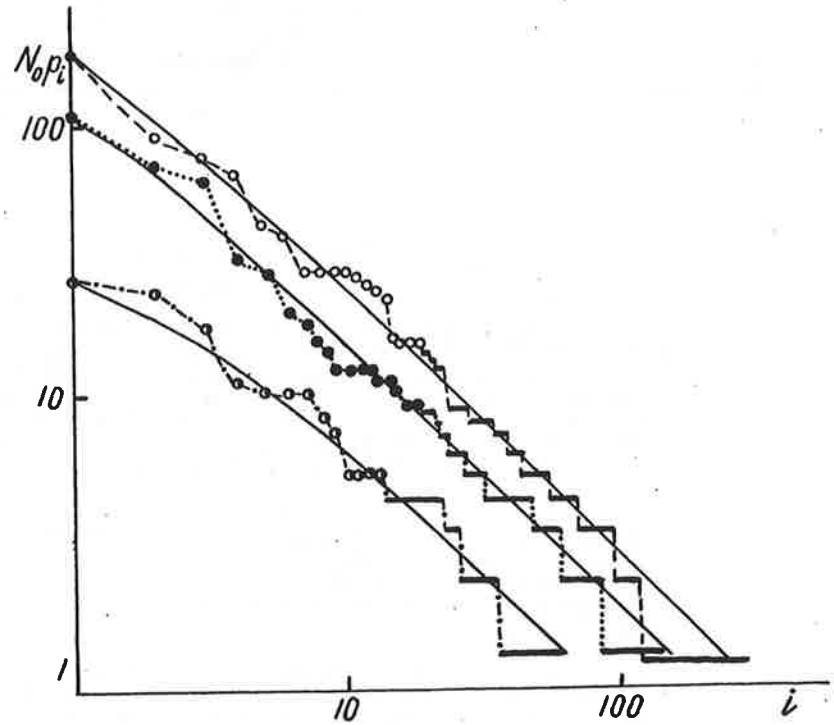


Abb. 5.1.1

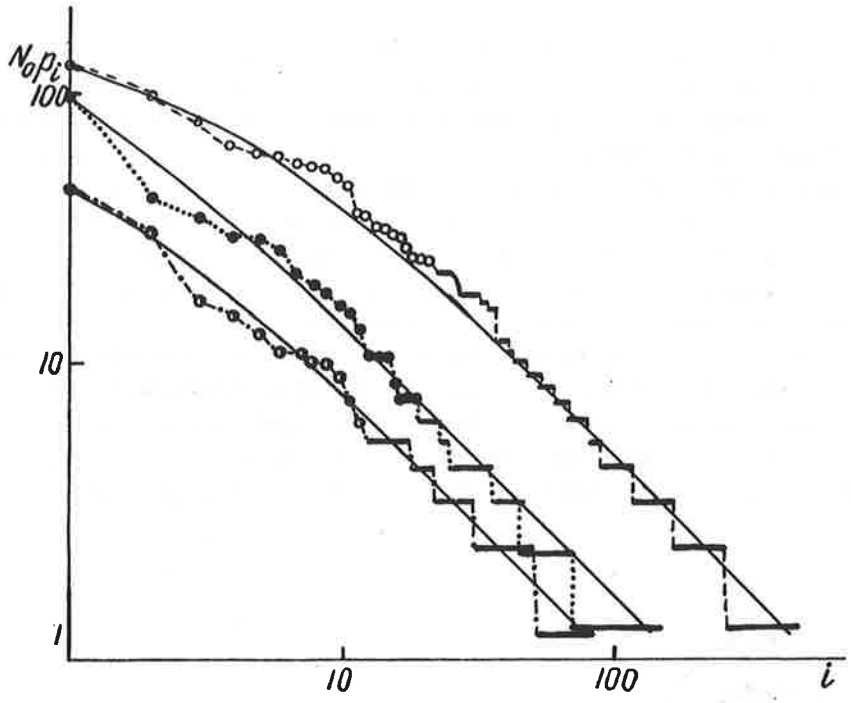


Abb. 5.1.2

Abb. 5.1: Graphische Darstellung der Häufigkeitsstrukturen vollständiger musikalischer Texte.

Auf der Ordinatenachse sind die absoluten Vorkommenshäufigkeiten der F-Motive im Text angetragen, auf der Abszissenachse die Ränge (Ordnungsnummern) der Häufigkeiten in einer nach abnehmenden Werten geordneten Reihenfolge. Auf beiden Achsen ist der Maßstab logarithmisch. Die theoretischen Kurven (gemäß Formel (3) konstruiert) erscheinen in Form einer durchgezogenen Linie.

Abb. 5.1.1:
a) J.S. Bach. Präludium und Fuge Nr. 22 aus dem Wohltemperierten Klavier, 2. Teil BWV 891 (O--O--O)
b) J.A. Levitin. Sonatine für Flöte solo (O...O...O)
c) D. Scarlatti. Sonate für Klavier Nr, 1 (Edition Peters) (●-·-●)
Abb. 5.1.2:
d) F. Chopin. Sonate Nr. 3 h-moll op. 58 (O--O--O)
e) J.S. Bach. Präludium und Fuge Nr. 2 aus dem Wohltemperierten Klavier, 2. Teil BWV 871 (●...●...●)
f) L. van Beethoven. Sonatine für Klavier F-Dur (●-·-●)

Wie die Abbildung zeigt, stimmen die experimentell ermittelten Kurven mit ihren theoretischen Beschreibungen gemäß (3) gut überein, sogar für einen relativ kurzen Text wie die Sonate Nr. 1 von Scarlatti (mit einer Länge von nur 250 F-Motiven). Gleichzeitig sind dabei die Texte ihrem Stil, ihrer Kompositionsform und ihrem satztechnischen Typ (homophone und polyphone Texte) nach unterschiedlich.

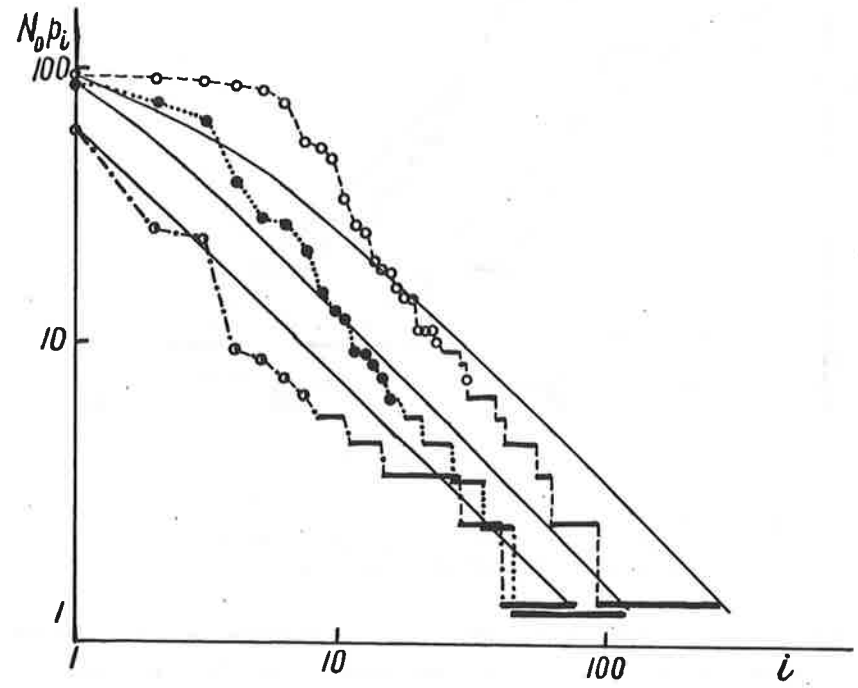


Abb. 5.2.1

Abb. 5.2: Graphische Darstellung der Häufigkeitsstrukturen von Textausschnitten (Teilen zyklischer Werke).

Die Benennungen der Achsen und der Maßstab sind dieselben wie in Abb. 5.1. Die theoretischen Kurven erscheinen in Form einer durchgezogenen Linie.

Abb. 5.2.1: g) J.S. Bach. Contrapunctus 8 aus der Kunst der Fuge (O--O--O)
h) J.S. Bach. Präludium aus BWV 891 (●...●...●)
i) Ju.A. Levitin. Sonatine für Flöte solo, 1. Satz (●-·-●)

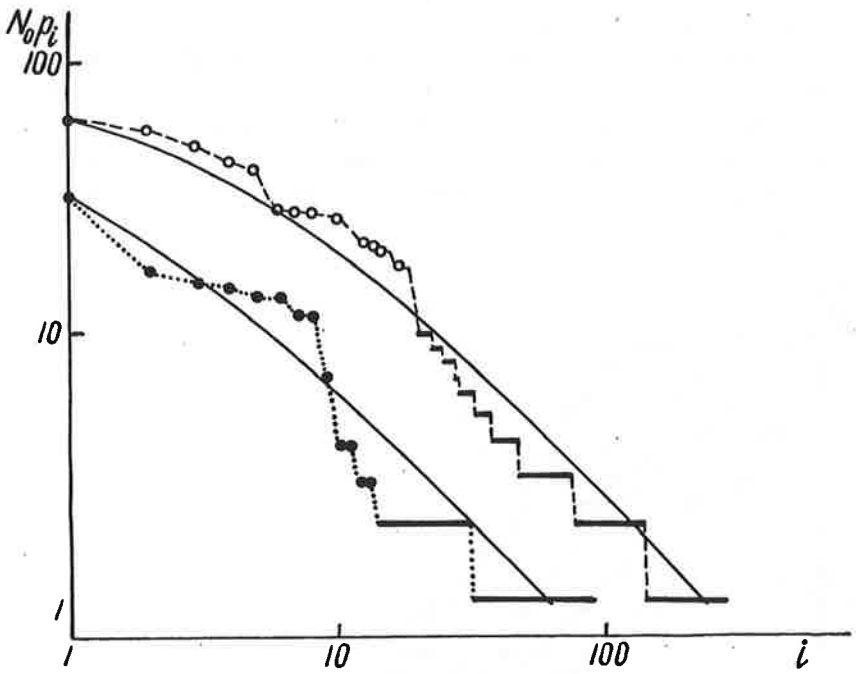


Abb. 5.2.2

Abb. 5.2.2:
j) F. Chopin. Sonate Nr. 3, 1. und 2. Satz (O--O)
k) D.D. Šostakovič. Präludium aus op. 87 Nr. 4 (●...●)

Zum Vergleich sind in der Abb. 5.2 die Graphen der Häufigkeitsstrukturen von *Textausschnitten* angegeben, also von einzelnen Teilen zyklischer Werke (darunter von solchen, für die die Häufigkeitsstrukturen der vollständigen Texte in der Abb. 5.1 dargestellt sind).

Die theoretischen Werte der Häufigkeiten der F-Motive sind nach der Formel (3) berechnet für N_o gleich der Länge des Textausschnittes und p_{max} gleich der Häufigkeit seines häufigsten F-Motives.

Wie aus der Abb. 5.2 klar ersichtlich wird, weichen die experimentell ermittelten Häufigkeitskurven ganz erheblich von den zugehörigen theoretischen Kurven ab, und zwar im Bereich kleiner Häufigkeiten

nach unten, im Bereich großer Häufigkeiten gewöhnlich nach oben, und bei einigen Graphen zeigen sich auch im Bereich mittlerer Häufigkeiten Abweichungen ("Durchbiegungen") der experimentellen von der theoretischen Kurve (so z.B bei dem Satz der Levitin-Sonatine).

Dabei ist es sehr wichtig, darauf hinzuweisen, daß die Längen der Textausschnitte, deren Häufigkeitsstrukturen in Abb. 5.2 (1+2) dargestellt sind, *durchaus vergleichbar* sind *mit den Längen der vollständigen Texte*, die Abb. 5.1 (1+2) zugrundeliegen. So beträgt beispielsweise die Länge des Ausschnittes aus Bachs Variationenzyklus "Die Kunst der Fuge" (Contrapunctus 8, vgl. Abb. 5.2.1) 1350 F-Motive und ist damit annähernd gleich der Länge von Präludium und Fuge Nr. 22 aus dem Wohltemperierten Klavier, 2. Teil (BWV 891, vgl. Abb. 5.1.1). Beim Vergleich von Abb. 5.1.1 und Abb. 5.2.1 läßt sich jedoch ein deutlicher Unterschied zwischen der Häufigkeitsstruktur der F-Motive des vollständigen Textes und der des Textausschnittes erkennen. Ein ähnliches Bild ergibt sich, wenn man den Ausschnitt aus Präludium und Fuge Nr. 22 (BWV 891, vgl. Abb. 5.2.1) dem annähernd gleichlangen vollständigen Text von Präludium und Fuge Nr. 2 aus dem Wohltemperierten Klavier, 2. Teil (BWV 871, vgl. Abb. 5.1.2) gegenüberstellt, annähernd gleichlange vollständige Texte von Chopin Textausschnitten etc. In allen diesen Fällen lassen sich bei den Textausschnitten erhebliche Abweichungen vom Zipf-Mandelbrotschen Gesetz feststellen: Der vollständige Text erfüllt dieses Gesetz wesentlich besser als der annähernd gleichlange Textausschnitt, selbst bei Werken desselben Komponisten. Ein ähnliches Bild wie in Abb. 5.1 und 5.2 läßt sich auch für die anderen untersuchten vollständigen Texte und Textausschnitte feststellen.

All diese Beobachtungen lassen die Aussage zu, daß in den untersuchten musikalischen Texten aus einem breiten Spektrum von Stilen die Organisation der Rekurrenz von F-Motiven in der Tat allgemeingültigen Abhängigkeitsverhältnissen unterliegt. Sie kann mithilfe des Zipf-Mandelbrotschen Gesetzes beschrieben werden, doch darüber hinaus steht die Erfüllung dieses Gesetzes in unmittelbarem und wesentlichem Zusammenhang mit dem musikalischen Werk als Ganzem, mit seiner Vollständigkeit und Abgeschlossenheit. Infolgedessen steht die Wiederholungsstruktur von F-Motiven im Text für alle untersuchten musikalischen Texte in einem *gleichartigen Zusammenhang mit der vollen Textlänge*.

4.

Die letztgenannte Schlußfolgerung hat eine wesentliche Konsequenz: Wenn die Wiederholungsstruktur von F-Motiven in einem für alle untersuchten Texte gleichartigen Zusammenhang steht, dann müssen in annähernd gleichlangen Texten (mit gleichzeitig annähernd gleichen Werten für die Häufigkeit ihres jeweiligen häufigsten F-Motivs p_{max}) auch *ähnliche* bzw. *annähernd gleiche* Wiederholungsstrukturen von F-Motiven vorliegen. Demgegenüber müssen sie sich in Texten mit erheblichen Längenunterschieden voneinander unterscheiden. Dabei spielt es keine Rolle, ob die Texte ihrem Stil nach grundsätzlich verschieden oder annähernd gleich sind.

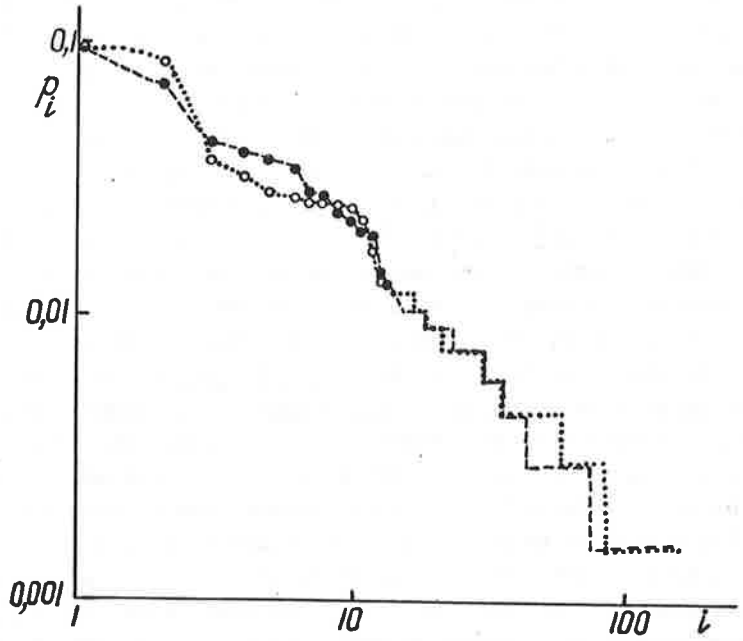


Abb. 6.1: Graphen der Häufigkeitsstrukturen musikalischer Texte mit annähernd gleicher Länge (bei annähernd gleichen Werten für p_{max}).

a) D.B. Kabalevskij. Rondo für Klavier op. 59 (Länge: 625 F-Motive) (O...O)
b) L. van Beethoven. Rondo C-Dur für Klavier op. 51 Nr. 1 (Länge: 624 F-Motive) (●--●)

Wie die Analyse zeigt, liegt diese maßgebende Erscheinung tatsächlich vor. So sind in der Abb. 6.1 beispielsweise die Graphen zweier Texte angegeben, deren Länge und ebenso deren Werte für p_{max} praktisch gleich sind.

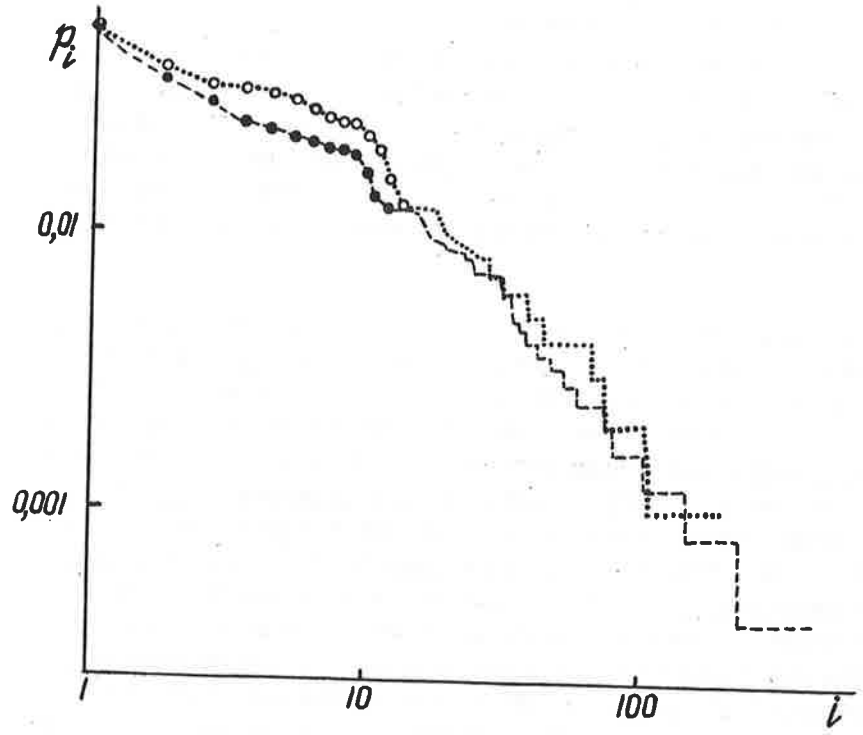


Abb. 6.2: Graphen der Häufigkeitsstrukturen musikalischer Texte mit unterschiedlicher Länge (bei annähernd gleichen Werten für p_{max}).

c) F. Chopin. Phantasie f-moll op. 49 (Länge: 987 F-Motive) (O...O)
d) F. Chopin. Sonate Nr. 3 h-moll op. 58 (Länge: 2364 F-Motive) (●--●)

Wie aus der Abbildung ersichtlich wird, ist trotz aller grundlegenden stilistischen Unterschiedlichkeit der Texte erwartungsgemäß der allgemeine Verlauf der Häufigkeitskurve für beide Rondos annähernd gleich, es zeigt sich weder ein systematisch höherer noch ein systematisch niedrigerer Verlauf einer Häufigkeitskurve im Vergleich zur anderen.

Im Gegensatz dazu sind in der Abb. 6.2 die Graphen der Häufigkeitsstrukturen von Texten angegeben, die zwar auch annähernd gleiche Werte für p_{max} haben, sich aber bezüglich ihrer Länge im Verhältnis 1:2,5 unterscheiden.

In diesem Falle ist, wie man klar erkennen kann, der Verlauf der Häufigkeitsabnahme in den beiden Texten unterschiedlich: Die Kurve für die Phantasie (den kürzeren Text!) verläuft insgesamt höher und fällt flacher als die Kurve für die Sonate. Ähnliche Verhältnisse, wie sie in Abb. 6.1 und 6.2 dargestellt sind, lassen sich auch für einige andere der untersuchten Texte aufzeigen.

5.

Die angeführten Beispiele werfen aber folgendes Problem auf: Wenn bei annähernder Gleichheit der Werte für p_{max} die Häufigkeitskurve für den längeren Text insgesamt *unter* der Kurve für den längeren Text verläuft, dann muß der längere Text relativ weniger häufig vorkommende F-Motive als der kürzere Text enthalten und andererseits relativ mehr selten, insbesondere nur einmal vorkommende. Infolgedessen müssen neue (vom Textanfang an erstmalig vorkommende) F-Motive im längeren Text durchschnittlich häufiger auftreten als im kürzeren. Und folglich muß, beim Vergleich von gleichlangen Textausschnitten aus einem längeren und einem kürzeren Text, das Motivinventar des Ausschnittes aus dem längeren durchschnittlich größer sein. (Der Zusatz "durchschnittlich" soll auf die Notwendigkeit hinweisen, die Ungleichmäßigkeiten im Anwachsen des Motivinventars im Text zu berücksichtigen, die durch die Eigenart der jeweiligen musikalischen Form bedingt sind). Mit anderen Worten: Der längere Text muß an F-Motiven (in der Terminologie der vorliegenden Arbeit:bezüglich seines Motivinventars) stärker gesättigt sein als der kürzere.

Diese Schlußfolgerung erscheint, wenn man die grundlegenden stilistischen Unterschiede der untersuchten Texte bedenkt (vgl. Tabelle 1), bei weitem nicht offenkundig. A priori vermutet man, daß bei einem musikalischen Text der "motivische Sättigungsgrad" in erster Linie durch den Stil und die Entstehungszeit des betreffenden Werkes bedingt ist. Es mutet wahrscheinlich an, daß die Werke eines einzelnen Komponisten (jedenfalls die größeren) alle ungefähr denselben motivischen Sättigungsgrad besitzen, zumindest auf der Ebene der F-Motive. Gerade deshalb ist es notwendig, den Zusammenhang zwischen dieser Charakteristik und der Textlänge experimentell zu überprüfen.

In der Abb. 7 sind die Graphen für die Akkumulation des Motivinventars, d.h. der Anzahl der distinkten F-Motive im Text, für einige verschieden lange Texte angegeben, von denen drei vom selben Komponisten (Chopin) stammen:

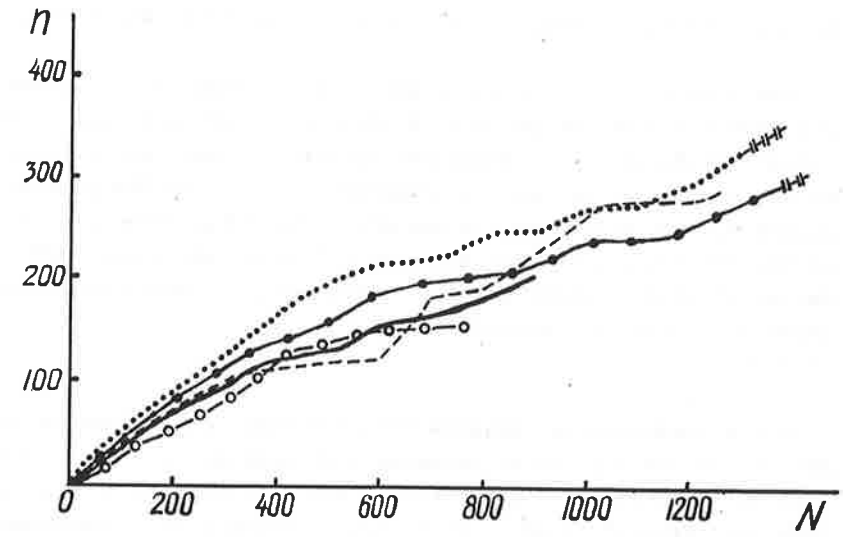


Abb. 7: Graphische Darstellung des Anwachsens des Motivinventars im musikalischen Text.

(Der weitere Verlauf der Kurven a) und c) über den Rand der Abbildung hinaus ist durch ⊢‖ ‖⊣ angedeutet).

a) F. Chopin. Sonate Nr. 3 h-moll op. 58 (• • •)
b) W.A. Mozart. Sonate für Klavier Nr. 10 (- - -)
c) F. Chopin. Sonate Nr. 2 b-moll op. 35 (•—•—•)
d) F. Chopin. Ballade Nr. 1 g-moll op. 23 (————)
e) Ju.A. Levitin. Sonatine für Flöte solo (o-o-o)

Die Abbildung zeigt, daß sich ungeachtet der auffälligen Unregel-

mäßigkeit im Anwachsen des Motivinventars im Text[11] deutlich die Tendenz beobachten läßt, daß der längere Text motivisch stärker gesättigt ist. Die Kurve des Anwachsens des Motivinventars im längeren Text verläuft im allgemeinen *über* der entsprechenden Kurve des kürzeren Textes. Dies ist besonders auffällig zu erkennen bei den drei Chopin-Texten: der Ballade Nr. 1 (Kurve d); Länge: ca. 900 F-Motive), der Sonate Nr. 2 (Kurve c); Länge: ca. 1500 F-Motive) und der Sonate Nr. 3 (Kurve a); Länge ca. 2400 F-Motive). Dort verläuft nämlich die Kurve des Anwachsens des Motivinventars für die Ballade insgesamt unter der Kurve für die Sonate Nr. 2, und diese wiederum verläuft unter der Kurve für die Sonate Nr. 3.

Somit wird für die untersuchten musikalischen Texte das verallgemeinerte Zipf-Mandelbrotsche Gesetz "im großen und ganzen" erfüllt (man denke an den übereinstimmenden Verlauf der experimentellen und theoretischen Häufigkeitskurven und die Übereinstimmungen mit der Beziehung (4) bezüglich des Motivinventars), und darüber hinaus ergibt sich aus diesem Gesetz eine wesentliche Folgerung in Form einer deutlichen Tendenz: Je länger ein Text ist, desto höher ist sein motivischer Sättigungsgrad.

6.

Die experimentellen Graphen der Häufigkeitsstrukturen in den Abb. 5. zeigen noch einen weiteren wichtigen Aspekt der Organisation der Rekurrenz von F-Motiven in den untersuchten Texten. Bei jedem der Graphen lassen sich in seinem rechten Teil charakteristische "Stufen" erkennen, die mit dem Vorhandensein umfangreicher Gruppen von gleich häufig vorkommenden F-Motiven mit kleiner Häufigkeit in Zusammenhang stehen. Wie aus den Abbildungen ersichtlich, wird der allgemeine Verlauf der Häufigkeitsabnahme von F-Motiven durch das verallgemeinerte Zipf-Mandelbrotsche Gesetz (3) im allgemeinen gut beschrieben. Jedoch wird die theoretische Kurve von den genannten "Stufen" vielfach geschnitten, und teilweise liegen diese auch unterhalb der Kurve, ohne sie überhaupt zu berühren. Dies ist ein Anzeichen dafür, daß die reale Wiederholungsstruktur von F-Motiven in musikalischen Texten in bestimmter Weise vom Gesetz (3) abweicht. In Anbetracht dessen, daß der

Bereich kleiner Häufigkeiten einen bedeutenden Teil der Häufigkeitsstruktur ausmacht, war es also von Interesse nachzuprüfen, wie groß diese Abweichungen sind und wie exakt die Zahl der m-mal vorkommenden F-Motive durch den Ausdruck (5), d.h. durch die Folgerung aus dem Gesetz (3) beschrieben wird. Die entsprechenden Daten sind in der Tabelle 1 (Spalten 8 - 16) aufgelistet. Die tatsächliche Anzahl der m-mal vorkommenden F-Motive ist dort mit n_m bezeichnet, die zugehörige Prognose gemäß (5) mit n_m^* .

Wie man erkennen kann, sind die Abweichungen von n_m gegenüber der Prognose n_m^* im Durchschnitt erheblich größer als die Abweichungen beim Motivinventar. Da aber der allgemeine Verlauf der Häufigkeitsabnahme und das Motivinventar von Texten mit dem verallgemeinerten Zipf-Mandelbrotschen Gesetz in Übereinstimmung stehen, läßt sich vermuten, daß im Bereich der kleinen Häufigkeiten gegenseitige Kompensationsmechanismen wirksam sind, bei denen ein Defizit an Elementen mit der einen Häufigkeit durch einen Überschuß an Elementen mit einer anderen sozusagen wieder ausgeglichen wird. In der Tat kann man anhand der Daten der Tabelle 1 feststellen,daß in einem Text, bei dem beispielsweise die Anzahl der einmal vorkommenden F-Motive kleiner ist als ihre Prognose gemäß (5), die Anzahl der zweimal oder dreimal vorkommenden F-Motive entsprechend größer ist als die zugehörige Prognose. Ähnliches gilt auch in anderen Fällen. Offensichtlich gewährleisten eben derartige "Umgruppierungen" die insgesamt gute Übereinstimmung der experimentellen Häufigkeitskurven mit dem verallgemeinerten Zipf-Mandelbrotschen Gesetz.

Anhand der Daten der Tabelle 1 läßt sich auch die allgemeine Ursache angeben, durch die wahrscheinlich diese "Umgruppierungen" zustandekommen. Wenn man die Zahl der m-mal im Text vorkommenden F-Motive mit der zugehörigen Prognose gemäß (5) vergleicht, kann man erkennen, daß die Zahl der *zweimal vorkommenden* F-Motive in der Regel *größer* ist als ihre theoretische Prognose. In drei Texten (in der Sonate von Tartini, in der Klaviersonate und in der "Kleinen Nachtmusik" von Mozart) ist die Zahl der zweimal vorkommenden F-Motive sogar *größer als die der einmal vorkommenden*. Im Durchschnitt des ganzen Textkorpus ist die Abweichung der Zahl der zweimal vorkommenden F-Motive gegenüber der Prognose gemäß (5) ungefähr dreimal so groß wie die entsprechenden Abweichungen

für die einmal bzw. dreimal vorkommenden F-Motive (bei letzteren dem Betrage nach). All diese Beobachtungen lassen die Annahme zu, daß in den untersuchten Texten ein eigenartiges "Prinzip des zweimaligen Vorkommens eines F-Motivs" wirksam ist, die Tendenz, ein einmal vorgekommenes F-Motiv zu *wiederholen*. Dieses Prinzip ist offenbar spezifisch für musikalische Texte (als "Prinzip des zweimaligen Vorkommens eines Elementes"), denn bei literarischen Texten, die bezüglich der Geltung des Gesetzes (3) untersucht worden sind, weicht, wie sich den Daten der Arbeit von Orlov (1975) entnehmen läßt, die Zahl der zweimal vorkommenden Wörter nur unwesentlich von (5) ab, und im Durchschnitt sind diese Abweichungen annähernd gleich Null. Überdies kann die übergroße Zahl der zweimal vorkommenden F-Motive in Zusammenhang gebracht werden mit der für die Musik charakteristischen Tendenz, ein Element *unmittelbar* oder *fast unmittelbar nach seinem ersten Vorkommen zu wiederholen*, so wie es bei der Wiederholung eines Themas nach seiner ersten Formulierung, der Wiederholung der Exposition einer Sonate oder Symphonie, der Motiv-Wiederholung bei der Formulierung eines Themas etc. geschieht. Offenbar ist dasselbe Prinzip auch auf der Ebene der F-Motive (vgl.auch Boroda 1976) gültig. Unter diesen Umständen kann das Defizit an einmal vorkommenden F-Motiven in einer Reihe von Texten darauf zurückzuführen sein, daß die Tendenz zur Elementwiederholung derartig stark in Erscheinung tritt, daß sie teilweise auf "potentiell nur einmal vorkommende" F-Motive übergreift. Es ist durchaus wahrscheinlich, daß das verallgemeinerte Zipf-Mandelbrotsche Gesetz und das "Prinzip des zweimaligen Vorkommens eines F-Motivs" für die untersuchten Texte gewissermassen *fundamentale Gegebenheiten* darstellen. Die "Kompensationserscheinungen" und generell die erheblichen Abweichungen im Bereich kleiner Häufigkeiten kommen dann zustande infolge der Kollision dieser beiden fundamentalen Gegebenheiten miteinander.

Fassen wir also unsere Ergebnisse zusammen. Wir haben anhand von musikalischem Textmaterial aus den Stilen des 18. - 20. Jahrhunderts die Organisation der Rekurrenz von kleinen melodischen Elementen, den F-Motiven, im musikalischen Text untersucht. Diese Analyse ergab

folgendes:

1. Die Wiederholungsstruktur von F-Motiven im Text steht in Zusammenhang mit der Textlänge N_o (der Summe aller gebrauchten F-Motive) und mit der Vorkommenshäufigkeit seines häufigsten F-Motivs p_{max}. Diese Struktur läßt sich mithilfe einer Form des in der quantitativen Linguistik bekannten Zipf-Mandelbrotschen Gesetzes beschreiben. Dabei ist es wesentlich, den Text als Ganzes zu betrachten, denn für Textausschnitte, selbst für in sich relativ geschlossene wie Teile von Werken in zyklischer Form, wird dieses Gesetz nicht erfüllt.

2. Je länger ein musikalischer Text ist, desto umfangreicher ist, bei gleichen Werten für p_{max}, sein Motivinventar (die Anzahl der distinkten F-Motive) und desto höher ist sein motivischer Sättigungsgrad. Demnach läßt sich bei einem Ausschnitt aus einem langen Text ein umfangreicheres Motivinventar feststellen als bei einem gleichlangen Ausschnitt aus einem kürzeren Text.

3. In jedem Text gibt es eine beträchtliche Anzahl selten vorkommender F-Motive. Die Zahl der $\underline{m}$-mal vorkommenden F-Motive wächst mit abnehmendem $\underline{m}$. Die Analyse des Bereichs kleiner Häufigkeiten ($\underline{m}$ = 1, 2, 3) ließ erkennen, daß in den Texten ein "Prinzip des zweimaligen Vorkommens eines F-Motivs" wirksam ist, das mit der für die Musik charakteristischen Tendenz in Zusammenhang steht, ein einmal vorgekommenes Element zu wiederholen, es ein zweites Mal vorkommen zu lassen. Diese Tendenz, die bei der ansonsten im allgemeinen dem Zipf-Mandelbrotschen Gesetz entsprechenden Wiederholungsstruktur von F-Motiven gleichzeitig in Erscheinung tritt, ruft offenbar die an vielen Texten festgestellten "Kompensationserscheinungen" hervor, bei denen der Überschuß an zweimal vorkommenden F-Motiven durch ein Defizit an dreimal, und in einigen Fällen auch an einmal vorkommenden F-Motiven ausgeglichen wird.

Tabelle 1

1	2	3	4	5	6
lfd. Nr.	Materialquelle	N_o	p_{max}	n	n*
1.	J.S. Bach (1685-1750). Präludium und Fuge Nr. 13 BWV 858; aus dem Wohltemperierten Klavier, 1.Teil (abgek.: WTK 1).	615	0.0910	186	153.0
2.	ders. Präludium und Fuge Nr. 20 BWV 865; WTK 1.	1422	0.1958	296	252.0
3.	ders. Präludium und Fuge Nr. 2 BWV 871; WTK 2.	671	0.1430	168	145.0
4.	ders. Präludium und Fuge Nr. 22 BWV 891; WTK 2.	1430	0.1350	310	272.0
5.	G.F. Händel (1685-1756). Fuge für Orgel Nr. 2 G-Dur.	765	0.1070	201	173.0
6.	D. Scarlatti (1685-1757). Sonate Nr. 1 (Zählung der Edition Peters).	250	0.1190	72	72.0
7.	ders. Sonate Nr. 5.	182	0.0659	60	67.0
8.	ders. Sonate Nr. 9. (Ausgabe von B. Holdenweiser).	156	0.1090	53	52.0
9.	ders. Sonate Nr. 13. (Ausgabe von L. Nikolaev).	568	0.0986	190	139.0
10.	ders. Sonate Nr. 25 (Edition Peters).	107	0.0655	51	47.2
11.	G. Tartini (1692-1770). Sonate g-moll für Violine und Klavier.	828	0.0652	218	204.0
12.	J. Haydn (1732-1809). Symphonie Nr. 45 fis-moll "Abschieds-Symphonie".	1304	0.0437	340	317.0
13.	W.A. Mozart (1756-1791). "Eine kleine Nachtmusik" KV 525.	1260	0.1143	205	252.0
14.	ders. Fuge g-moll für Klavier.	731	0.1370	199	157.0
15.	ders. Sonate für Klavier Nr. 10.	1249	0.0509	296	297.0
16.	J.G. Albrechtsberger (1736-1809). Fuge c-moll.	693	0.2160	168	138.0
17.	L. van Beethoven (1770-1827). Sonate für Klavier Nr. 19.	584	0.0565	151	163.0

7	8	9	10	11	12	13	14	15	16	17
δ	n_1	n_1^*	δ_1	n_2	n_2^*	δ_2	n_3	n_3^*	δ_3	$\frac{n_1}{n}$
21.6	105	93.0	12.9	31	31.0	0.0	12	15.5	-22.6	0.565
17.45	192	126.0	58.4	39	42.0	-7.2	16	21.0	-23.8	0.648
15.8	96	72.5	32.4	26	24.2	8.3	9	12.1	-25.8	0.572
13.6	191	136.0	41.3	26	45.0	-42.0	21	22.7	-7.5	0.616
16.2	120	86.5	38.8	28	28.8	-2.8	8	14.0	-42.9	0.598
0.0	37	36.0	2.8	9	12.0	-25.0	3	6.0	-50.0	0.515
-10.5	28	33.5	-16.2	16	11.2	42.9	1	5.6	-46.6	0.467
1.9	20	26.0	-23.1	11	8.7	27.0	8	4.3	84.5	0.378
36.8	100	69.5	43.9	39	23.2	68.2	17	11.6	46.6	0.526
8.1	25	23.6	5.9	15	7.9	90.7	3	3.4	-12.5	0.490
6.9	60	102.0	-41.2	85	34.0	150.0	11	17.0	-35.3	0.276
7.3	164	158.5	3.5	65	52.8	23.1	25	28.5	-6.0	0.483
-18.7	60	126.0	-52.4	61	42.2	44.6	13	21.1	-38.4	0.293
26.7	119	78.5	51.6	33	25.8	16.3	10	12.9	-14.7	0.548
0.0	95	149.0	-36.2	98	49.5	98.0	14	24.8	-43.6	0.321
20.9	111	69.8	70.0	17	23.0	-21.8	8	11.5	-30.8	0.661
-7.4	60	81.5	-26.4	36	27.2	32.4	13	13.6	-4.4	0.400

(Fortsetzung Tabelle 1)

1	2	3	4	5	6
18.	ders. Rondo C-Dur für Klavier op. 51 Nr. 1.	624	0.0960	162	150.0
19.	ders. Sonatine für Klavier F-Dur.	333	0.1320	85	86.6
20.	R. Schumann (1810-1856). Konzert für Violoncello und Orchester a-moll op. 129.	2023	0.0537	532	428.0
21.	F. Chopin (1810-1849). Ballade Nr. 1 g-moll op. 23.	902	0.0932	206	201.0
22.	ders. Sonate Nr.2 b-moll op.35.	1484	0.0458	330	346.0
23.	ders. Sonate Nr.3 h-moll op.58.	2364	0.0532	504	485.0
24.	ders. Phantasie f-moll op. 49.	987	0.0589	209	239.0
25.	F. Mendelssohn-Bartholdy (1809-1847). Präludium und Fuge für Orgel e-moll.	1393	0.1098	282	276.0
26.	C. Saint-Saëns (1835-1921). Introduction et Rondo capriccioso für Violine und Orchester.	1200	0.0675	222	270.0
27.	A.N. Skrjabin (1872-1915). Valse op. 38.	402	0.0746	155	117.0
28.	N.J. Mjaskovskij (1881-1950). Fuge für Klavier e-moll.	214	0.1635	60	58.5
29.	S.S. Prokof'ev (1891-1953). Sonate für Violine solo op. 115.	1519	0.0745	398	318.0
30.	P. Hindemith (1895-1963). Sonate für Violine solo Nr. 1	1452	0.0584	374	325.0
31.	D.D. Šostakovič (1906-1975). Präludium und Fuge für Klavier op. 87 Nr. 4	983	0.1078	237	208.0
32.	ders. Präludium und Fuge für Klavier op. 87 Nr. 9	677	0.2160	146	135.0
33.	ders. Präludium und Fuge für Klavier op. 87 Nr. 12	1292	0.0844	316	274.0
34.	D.B. Kabalevskij (geb. 1904). Rondo für Klavier op. 59.	625	0.0950	171	150.0
35.	S.M. Taktakišvili (geb. 1900). Rondo für Violine und Klavier.	512	0.1268	103	120.0
36.	Ju.A. Levitin (geb. 1912). Sonatine für Flöte solo.	760	0.1466	159	160.0

7	8	9	10	11	12	13	14	15	16	17
8.0	83	75.0	9.0	33	25.0	32.0	9	12.5	-28.0	0.512
-1.9	31	43.3	-28.4	23	14.4	59.6	9	7.2	25.0	0.365
24.3	280	214.0	30.8	145	71.4	103.0	20	35.7	44.0	0.527
2.5	108	100.5	7.5	36	33.4	7.8	9	16.7	-46.1	0.525
-4.6	117	174.0	-32.8	84	58.0	44.8	25	29.0	-13.8	0.355
3.9	237	242.5	-2.3	100	80.7	23.9	48	40.4	18.9	0.470
-12.6	92	119.5	-23.0	36	39.8	-9.6	8	19.9	-59.9	0.440
2.2	153	138.0	10.2	48	46.0	4.4	18	23.0	-21.8	0.539
-17.8	104	135.0	-23.0	47	45.0	4.5	14	22.5	-37.8	0.469
27.4	76	57.8	37.4	43	19.6	119.6	5	9.8	-38.8	0.516
2.6	35	29.3	19.9	10	9.8	2.0	3	4.9	-32.2	0.584
26.2	204	158.0	28.4	90	52.6	66.1	28	26.3	1.9	0.514
15.7	206	162.5	27.4	70	54.0	29.6	27	27.0	0.0	0.553
13.9	144	104.0	38.5	29	34.0	-14.8	12	17.4	-30.8	0.608
8.2	74	67.5	9.6	24	22.5	6.7	11	11.2	-1.8	0.506
15.3	212	137.0	54.8	33	45.6	-27.7	18	22.8	-21.1	0.671
14.0	84	75.0	12.0	29	25.0	16.0	19	12.5	52.0	0.492
-14.1	48	60.0	-20.0	19	20.0	-5.0	3	10.0	-70.0	0.466
-0.6	76	80.0	-5.0	25	26.5	-5.7	13	13.3	-2.3	0.478

Anmerkungen:

1 Die Untersuchungen zu dieser Arbeit wurden am Lehrstuhl für Ästhetik und Kunstwissenschaft des Staatlichen V. Saradžišvili-Konservatoriums Tbilisi (Kafedra estetiki i iskusstvovedenija Tbilisskoj gosudarstvennoj konservatorii im. V. Saradžišvili) durchgeführt. Der Aufsatz, dessen überarbeitete und ergänzte Fassung hier in Übersetzung erscheint, wurde in russischer Sprache unter dem Titel "Častotnye struktury muzykal'nych tekstov" in dem Sammelband: Sbornik statej posvjaščennyj 60-letiju Velikoj Oktjabr'skoj socialističeskoj revoljucii (Sammelband aus Anlaß der 60-Jahrfeier der Oktoberrevolution). (ed. A. Saverzašvili et al.). Tbilisi: Mecniereba, 1977 veröffentlicht. Der Autor dankt Ju.K. Orlov, R.Ch. Zaripov und E.M. Dumanis für ihre freundlichen Hinweise bei der Durchsicht der vorliegenden Arbeit.

2 Gemeint sind die Organisationsprinzipien für einen Text in seiner G e s a m t h e i t , von Anfang bis Ende. Die Gesetzmäßigkeiten der Wiederholung und Alternation kleiner Elemente in weniger umfangreichen Teileinheiten, insbesondere in thematischen Abschnitten, sind in der Musikwissenschaft auch in Einzelheiten hinreichend bekannt (vgl. Mazel', Cukkerman 1967).

3 Unter der Textlänge versteht man die Zahl der im Text gebrauchten Wortformen.

4 In Anlehnung an Orlov (1975) werden wir den Ausdruck (3) im folgenden auch als verallgemeinertes Zipf-Mandelbrotsches Gesetz bezeichnen.

5 Die Namen russischer bzw. sowjetischer Autoren (Schriftsteller und Komponisten) werden hier nicht in der sonst üblichen Weise transkribiert, sondern sie erscheinen, ebenso wie die von Verfassern wissenschaftlicher Arbeiten, in wissenschaftlicher Transliteration (also z.B. Šostakovič - nicht Schostakowitsch). (Anm. d. Übers.).

6 Eine detailliertere und enger gefaßte Definition für a) - d) findet sich in unseren erwähnten Arbeiten (Boroda 1973, 1977).

7 In polyphonen Texten wurde zunächst jede einzelne Stimme s e p a r a t in F-Motive segmentiert, und daraufhin wurden die jeweiligen Summen der 1., 2. etc. Stimme in einer einzigen Summe vereinheitlichend zusammengefaßt, anhand derer dann die unten aufgeführten Charakteristika für den Text bestimmt wurden.

8 Der im russ. Original gebrauchte Terminus, den wir hier mit "Motivinventar" übersetzen, lautet "intonacionnyj zapas" (wörtlich: intonatorisches Inventar). In der russ. musikwissenschaftlichen Terminologie hat "intonacija" (wörtlich: Intonation) einerseits dieselben Bedeutungen wie der deutsche Terminus "Intonation", andererseits wird es aber auch in bestimmten, im hier gegebenen Zusammenhang vorliegenden, Fällen synonym für "Motiv" gebraucht. Da dies in der deutschen Terminologie nicht möglich ist, mußte auf die Übersetzung "Motivinventar" zurückgegriffen werden, auch wenn dabei die terminologischen Unterschiede zwischen "intonacija" und "motiv" bzw. zwischen "F-Motiv" und "Motiv" in der gängigen Definition nicht deutlich gemacht werden. Darüber hinaus ist der Gebrauch von "intonacija" hier insbesondere auch dadurch motiviert, daß in metaphorischer Bedeutung "intonacija" als "musikalisches Wort" (muzykal'noe slovo) verstanden werden kann. Somit ist die hier aufzuzeigende Analogie zu literarischen Texten, an denen das Lexeminventar (slovarnyj zapas) untersucht wird, durch den Terminus "intonacionnyj zapas" für "Motivinventar" im Sinne von "Inventar an musikalischen Wörtern" zusätzlich verdeutlicht. Zu den Bedeutungen von "motiv" und "intonacija" in der russ. Terminologie, besonders auch zum Verhältnis von "intonacija" als musikalischem Wort zum Wort als lexikalischer Einheit vgl. das neueste sowjetische musikwissenschaftliche Standardwerk: Muzykal'naja enciklopedija. (ed. Ju.V. Keldyš). Bd. 1, Moskva, 1973 . "motiv": Bd. 3, 1976, Spalte 696-698; "intonacija": Bd. 2, 1974, Sp. 550-557, bes. Sp. 553-554. (Anm. d. Übers.).

9 Dabei war natürlich die tongetreue Wiederholung eines F-Motivs als eine Sequenzbildung mit Null-Intervall mitverstanden.

10 Die Textlänge, das Motivinventar und die Vorkommenshäufigkeiten von F-Motiven wurden für Textausschnitte in der gleichen Weise bestimmt wie für vollständige Texte.

11 Zu einigen Erscheinungsformen und Ursachen dieser Unregelmäßigkeit vgl. Boroda (1976).

Erläuterungen zu Tabelle 1:

N_o - Textlänge; p_{max} - relative Häufigkeit des häufigsten F-Motivs im Text; n - tatsächliches Motivinventar des Textes; n* - Prognose für das Motivinventar nach der Formel (4); δ - relative Abweichung von n gegenüber n*; n_m (m = 1, 2, 3) - tatsächliche Anzahl der einmal, zweimal bzw. dreimal im Text vorkommenden F-Motive; n_m* - theoretische Prognose für die Anzahl der m-mal im Text vorkommenden F-Motive nach der Formel (5); δ_m - relative Abweichung von n_m gegenüber n_m*. Ebenso wird in der Tabelle für jeden Text der Wert des Verhältnisses n_1/n der Anzahl der einmal vorkommenden F-Motive zum Motivinventar angegeben.

Die Korrespondenzen der hier angegebenen Numerierung zu der in neuerer Zeit üblich gewordenen Numerierung der Werke D. Scarlattis nach Kirkpatrick konnten für die in Frage kommenden Sonaten mangels Material nicht ermittelt werden. (Anm. d. Übers.)

DAS VERALLGEMEINERTE ZIPF-MANDELBROTSCHE GESETZ UND DIE VERTEILUNG DER ANTEILE VON FARBFLÄCHEN IN DER MALEREI

B.A. Vološin, Ju.K. Orlov

Wenn universale Gesetze oder Mechanismen der Informationsaufnahme existieren, dann müssen sich diese in der Universalität von Nachrichtenstrukturen zeigen, die in bestimmter Weise für die menschliche Wahrnehmung optimal sind. Eine solche universale Struktur, die sich sowohl in verschiedenen literarischen Werken wie auch im musikalischen Material von "hohem künstlerischen Wert" besonders exakt zeigt, ist das verallgemeinerte Zipf-Mandelbrotsche Gesetz (vgl. Orlov, 1970a,b; Boroda & Orlov 1970). Die in den Arbeiten von Orlov (1969: 250, 255) durchgeführte Analyse dieses Gesetzes, wie auch die experimentelle Überprüfung ihrer Ergebnisse (vgl. Nadarejšvili & Orlov 1971: 549) haben gezeigt, daß die Verwirklichung dieses Gesetzes nicht durch das Wirken universaler statistischer Mechanismen erklärt werden kann, sondern zielgerichtete Bemühungen seitens des Autors vorausgesetzt werden müssen, die ihm selbst anscheinend völlig unbewußt sind. Die Tatsache, daß solche Bemühungen notwendig sind, bestätigt die große Bedeutung dieser Gesetzmäßigkeit in der Kommunikation. Die Annahme liegt nahe, daß ihre Verwirklichung in einer Nachricht mit der optimalen Ausnutzung der Mechanismen des menschlichen probabilistischen Prognostizierens im Augenblick der Nachrichtenaufnahme zusammenhängt.

In der vorliegenden Arbeit wird eine vorläufige Überprüfung des verallgemeinerten Zipf-Mandelbrotschen Gesetzes berichtet. Untersucht wurde die nach abnehmender Häufigkeit geordnete Reihenfolge der Flächen eines Bildes, die von einer bestimmten Farbe eingenommen werden (eine Fläche ist die Summe aller Teilflächen derselben Farbe). Es muß betont werden, daß ein grundlegender Unterschied zwischen der vorliegenden Untersuchung

und den bekannten Untersuchungen der Anteile von Farbflächen besteht: in der vorliegenden Arbeit wird die konkrete Farbe der Flächen völlig unberücksichtigt gelassen. Daher ist die erhaltene Reihenfolge ebenso unabhängig von einer konkreten Farbe, wie die Worthäufigkeitsreihe einer lexikalischen Stichprobe von der nominellen Häufigkeit eines jeden Einzelwortes.

Für die Untersuchung wurden Farbreproduktionen der Kunstwerke verwendet. Die Einteilung und Identifizierung der Farbflächen wurde von Experten durchgeführt,[1)] zu denen auch einer der Autoren gehörte. Die Reproduktion wurde in die jeweiligen Farbflächen zerschnitten und zur Identifikation der Farben wurden die ausgeschnittenen Teile paarweise vor schwarzem Hintergrund verglichen. Durch dieses Vorgehen wurde der Einfluß der andersfarbigen Umgebung einer Fläche ausgeschlossen. Die Gesamtfläche einer Farbe wurde durch Abwiegen festgestellt; vorher wurde die Homogenität des Papiers, auf dem die Reproduktion gedruckt wurde, überprüft.

Es stellte sich heraus, daß die nach abnehmender Häufigkeit geordnete Reihenfolge der Farbflächenanteile dieser Bilder (im folgenden werden wir diese Reihenfolge "Farbreihe" nennen) - zumindest im Anfangsteil - der Formel folgt, die sich aus dem verallgemeinerten Zipf-Mandelbrotschen Gesetz ergibt:

$$f(p) = Ap^{-\alpha} \qquad (1)$$

[dies ist Formel (7) aus Orlov (1970b)]; hierbei ist f(p) die bedingte Dichte der Verteilung der Farbflächenanteile, α ist ein unabhängiger Parameter, A ist eine Konstante, die sich aus der Normierung

$$\int_{p_v}^{p_1} f(p)\, dp = 1$$

ergibt, wobei p_1 und p_v jeweils die größte und die kleinste von einer Farbe eingenommene relative Fläche darstellen.

Aus (1) erhält man bei unterschiedlichen Werten von α verschiedene Approximationen für die Farbreihe. Insbesondere erhalten wir für $\alpha > 0$ die Mandelbrotsche Formel

$$p_i = K(B + i)^{-1/\alpha} \qquad (2)$$

wobei

$$K = \left(\frac{1-\alpha}{p_1^{1-\alpha} - p_v^{1-\alpha}}\right)^{1/\alpha} \quad ; \quad B = \left(\frac{K}{p_1}\right)^{\alpha} - 1, \text{ wenn } \alpha \neq 1 \qquad (3)$$

und

$$K = \frac{1}{\ln \frac{p_1}{p_2}} \quad ; \quad B = \frac{K}{p_1} - 1; \text{ wenn } \alpha = 1. \qquad (4)$$

Bei $\alpha = 0$ erhalten wir die Exponentialapproximation

$$p_i = p_1 \exp\left[(1-i)(p_1 - p_v)\right]. \qquad (5)$$

Ein typischer Kurvenverlauf, wie wir ihn bei der Analyse des Bildes von Levitan "Ewige Ruhe" erhielten, ist in der Abbildung 1 dargestellt. Auf der Abszissenachse sind in logarithmischem Maßstab die Rangplätze nach der Liste der Farbanteile (in abnehmender Reihenfolge), auf der Ordinatenachse - in gleichem Maßstab - die Anteile selbst aufgetragen. Man kann leicht sehen, daß der Anfangsteil der Farbreihe zufriedenstellend mit einer Geraden approximiert werden kann, was dem Ausdruck (2) bei

$$B = 0 \qquad (6)$$

entspricht, wobei der Steigungskoeffizient der angepaßten Geraden gleich $-1/\alpha$ ist.

Obwohl der auf diese Weise approximierte Abschnitt der Rangreihe in der Regel deren kleineren Teil ausmacht, ist der gesam-

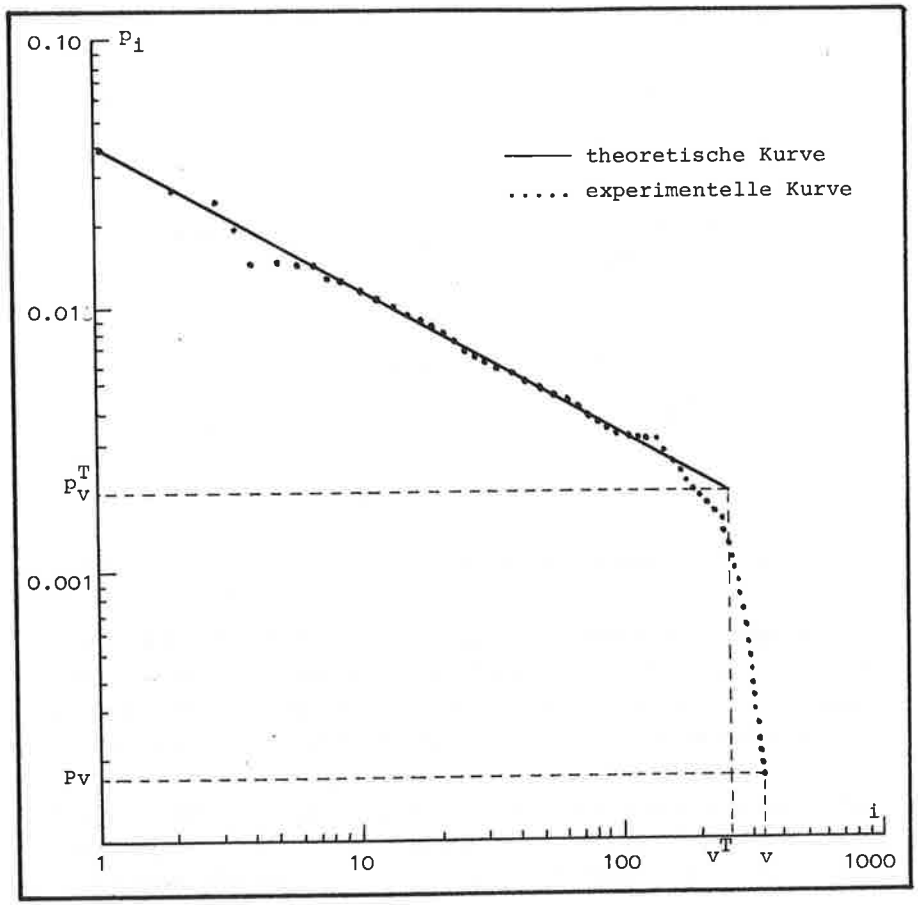


Abb. 1. Verteilung der Farbflächen in Levitans "Ewige Ruhe"

te Flächenanteil, der von den entsprechenden Farben eingenommen wird, normalerweise beträchtlich größer als die Hälfte der Gesamtfläche eines Bildes (~ 80-95%). Das bedeutet, daß die Gesetzmäßigkeit (1) durch die "gewichtigsten", dominierenden Farben eines Bildes erfüllt wird. Kleine Farbflächen jedoch, deren Flächenanteile stark von der angepaßten Kurve nach unten abweichen, kommen in viel größerer Anzahl vor, als nach Formel (1) zu erwarten wäre. Und tatsächlich, wenn wir aus der Formel (6) die Größe p_v^T bestimmen - den Wert des Gliedes der Reihe nach (2), bis zu dem die Summe der ersten v^T ihrer Glieder Eins erreicht (da der Ausdruck für B aus der Normierungsbedingung für f(p) erhalten worden war) - sehen wir, daß der "Schweif" der empirischen Graphik sich deutlich weiter ausdehnt.

Ein solches Bild erhielten wir - mit einigen Variationen - im gesamten untersuchten Material (s. Tabelle 1). Die Werte des Koeffizienten α und der Hilfskoeffizienten K und B sind von Bild zu Bild unterschiedlich und hängen von dem jeweiligen Experten ab, der die Farbidentität der verschiedenen Flächen bestimmt hat. Jedoch sind die Schwankungen der fundamentalen Größen p_1, P_v^T (diese Größe kann man als eine Grenze betrachten, unter der wesentliche Abweichungen von den Formeln des verallgemeinerten Zipf-Mandelbrotschen Gesetzes beginnen) und α für das gleiche Bild Levitans, die anhand von Auswertungen verschiedener Experten festgestellt wurden, recht gering.

Sehr gering sind die Abweichungen des "Schweifes" bei dem Bild "Ukrainische Nacht" von A. Kuindži, das praktisch überhaupt keine kleinen Farbflächen und Details aufweist und dessen Farbreihe außerordentlich abrupt abfällt (α = 0, die Reihe wird durch den Ausdruck (5) approximiert). In dem Plakat von O. Džiškariani "Blühe auf, schönes Land" (α = 0.3) gibt es überhaupt keinen "Schweif".

Eine "stückweise"[2] Analyse des Bildes von Levitan zeigte, daß in allen ausreichend großen Teilen (d.h. 1/4 und 1/2 der Gesamtfläche) des Bildes die gleiche Gesetzmäßigkeit auftritt: eine gute Approximation (jedoch etwas schlechter als für das ganze Bild) des Anfangsteiles der Farbreihe durch den Ausdruck

Tabelle 1.

Nr.	Material (in Klammern der Name des Experten)	v	$\alpha = \frac{1}{\gamma}$	p_{max}	p_{min}	p^{T}_{min}	K	B	Umfang des "Schweifes" in %	Gewicht des "Schweifes" in %
1	Kuindži "Ukrainische Nacht" (Vološin)	30	0	0,1061	0,0001	-	-	-	3,3	0,6
2	Levitan "Ewige Ruhe" (Vološin)	61	2,5	0,0892	0,0004	0,0232	0,0074	0	77,0	33,6
3	---------- " ------- (Čchaidze)	98	2,3	0,0565	0,0004	0,0101	0,0086	0	69,5	24,4
4	---------- " -------- I (Nadarejšvili)	94	2,0	0,0347	0,0006	0,0022	0,0173	0	43,4	8,47
5	---------- " -------- II	109	2,26	0,0526	0,0012	0,0074	0,0183	0	54	21,8
6	---------- " -------- III	143	2,72	0,0409	0,0003	0,0086	0,0074	0	72	39
7	---------- " -------- IV	79	1,165	0,1257	0,0096	0,0071	0,1353	0	50	14,2
8	---------- " --- I u. III	232	2,26	0,0268	0,0002	0,0015	0,0302	0	61	28,6
9	---------- " --- II u. IV	205	1,61	0,0556	0,0002	0,0034	0,0136	0	61	20
10	---------- " ---- I u. II	159	2,35	0,0281	0,0002	0,0063	0,0334	0	61	29,8
11	---------- " ---III u. IV	173	1,33	0,0842	0,0003	0,0017	0,0150	0	30,6	4,5
12	---------- " ----------	341	1,84	0,0398	0,0001	0,0020	0,0287	0	32,6	4,6
13	Džiškariani "Blühe auf schönes Land" (Nadarejšvili)	93	0,3	0,1333	0,0001	-	812,2	65,9	0	0

(2) und eine darauffolgende abrupte Abweichung nach unten, wobei das relative Gewicht des abweichenden Teiles anwächst.

Aus den Arbeiten von Jarbus (1966) ist bekannt, daß sich die Augen beim Betrachten von Bildern vorzugsweise auf die Konturen und die kleinen Elemente eines Bildes fixieren. Im Zusammenhang damit kann man tatsächlich vermuten, daß Farbflächen, deren Anteile den Anfangsteil der Kurve darstellen (der gut durch den Ausdruck (1) approximiert ist), hauptsächlich peripher wahrgenommen werden, während die kleinen Farbflächen und Details, deren Anteile den unteren Kurvenabschnitt, den "Schweif" bilden, vorzugsweise die foveale Wahrnehmung auf sich lenken.

Der mittlere Sehwinkel der Farbflächen, die den Anfang des "Schweifes" bilden (die Biegung der Kurve im Bereich p^{T}_{v}) beträgt ungefähr $0.3\text{-}0.6^{o}$ (es wurde die mittlere Zahl der Farbflächen berechnet, die im Bereich der Biegung lagen, sowie der Abstand, aus dem die Versuchspersonen ein Bild vorzugsweise betrachtet haben, unter der Annahme, daß die Form der Farbfläche annähernd rund ist). Berücksichtigt man, daß die Mehrheit der Flächen große Ausmaße und bei weitem keine runde Form besitzt, so kann man sagen, daß diese Größe mit dem Winkel der zentralen Netzhautgrube (Fovea), der 1.3^{o} (Glezer & Cukkerman 1961) beträgt, gut übereinstimmt.

So kann man zu der Meinung gelangen, daß ein beträchtlicher Teil der Verteilung der Farbflächen dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgt. Hierbei handelt es sich genau um den Teil eines Bildes, auf den die Aufmerksamkeit eines Betrachters nicht vorrangig gerichtet ist und der peripher wahrgenommen wird. Kleine Farbflächen und Details jedoch, deren Projektionen bei gewöhnlicher Betrachtungsweise auf der Fovea abgebildet werden, weisen eine wesentliche Abweichung von diesem Gesetz auf. Dies zeugt von einer Feinabstimmung zwischen der beobachteten Verteilung und den Sehmechanismen. Man kann nicht ausschließen, daß die Organisation der großen Farbflächen in Übereinstimmung mit dem verallgemeinerten Zipf-Mandelbrotschen Gesetz zu einem gewissen Grad mit dem Phänomen des "Kolorits" zusammenhängt.

Die Autoren danken der Mitarbeiterin des Institutes für Kybernetik der Akademie der Wissenschaften der GSSR, I.Š. Nadarejšvili, und dem Studenten der Kunstakademie von Tiffis, M.JU. Čchaidze, die für die vorliegende Untersuchung als Experten tätig waren.

ANMERKUNGEN

1) Das Verfahren der subjektiven Expertise wurde gewählt, weil es unumgänglich ist, die Verteilung der Farbflächen vom "Gesichtpunkt" des menschlichen Auges her zu untersuchen. Die Grenzen nämlich, die Meßgeräte angeben würden, entsprechen nicht der Empfindlichkeit, mit der das menschliche Auge eine Farbänderung wahrnimmt. Die Verfälschung der Farbnuance, die gewöhnlich bei mehrfarbigen Reproduktionen auftritt, spielte in der vorliegenden Untersuchung keine Rolle, da im Grunde genommen nur relative Farbänderungen untersucht wurden. Ein Vergleich der einzelnen Reproduktionen mit den Originalen zeigte, daß im Original unterschiedlich gefärbte Teile in den Reproduktionen ebenso erschienen.

2) Das Bild wurde in vier gleich große Teile zerschnitten. Das linke obere Rechteck wird in der Tabelle mit I bezeichnet, das rechte obere mit II, das linke untere mit III und das rechte untere mit IV.

Über die Verwendung der Wörter unterschiedlicher Häufigkeiten in Rustavelis Gedicht "Der Held im Tigerfell".

I.Š. Nadarejšvili, Ju.K. Orlov

Die Verteilung der Wörter in Rustavelis Gedicht "Der Held im Tigerfell" ist in der Tabelle 1 angegeben (Rustaveli 1951). Das Vokabular des Gedichtes wurde je nach Häufigkeit des Wortvorkommens in Klassen aufgeteilt[1]. In die erste Klasse fallen diejenigen Wörter, die nur jeweils 1 mal vorkommen; in die zweite Klasse diejenigen, die 2 oder 3 mal vorkommen usw., wobei jede weitere Klasse doppelt so breit ist wie die vorangehende.

Tabelle 1. Verteilung der Worthäufigkeiten

Klasse Nr.	Häufigkeitsklassen	Zahl unterschiedlicher Wörter in der Klasse	Zahl der Wortverwendungen in der Klasse
	Wörter, die		
1	1 mal vorkommen	2995	2995
2	2 bis 3 mal vorkommen	1329	3050
3	4 bis 7 mal vorkommen	767	3891
4	8 bis 15 mal vorkommen	389	4239
5	16 bis 31 mal vorkommen	246	5507
6	32 bis 63 mal vorkommen	146	6594
7	64 bis 127 mal vorkommen	52	4507
8	128 bis 255 mal vorkommen	26	4984
9	256 bis 511 mal vorkommen	11	3839
10	512 bis 880 mal vorkommen	4	2940
		5965	42120

Wie ersichtlich, enthält das Gedicht 5965 unterschiedliche Wörter; ungefähr die Hälfte von ihnen kommt nur 1 mal vor. Weiter kann man die folgende Gesetzmäßigkeit beobachten: jede Klasse hat etwa den doppelten Umfang wie die nächstgrößte Klasse. Gleichzeitig sind aber die Wortverwendungen in allen Klassen ungefähr gleich (die letzte Spalte). Auf diese Weise kann man bei einer zufälligen Wahl eines Wortes aus dem Gedicht mit gleicher Wahrscheinlichkeit ein Wort aus beliebiger Klasse nehmen. Es bestehen gute Gründe zur Annahme, daß eine ähnliche Gesetzmäßigkeit einen allgemeinen Charakter hat.

Da die Struktur des ganzen Gedichtes sehr deutlich und einheitlich ist, stellte sich das Problem der Beziehung der Wortverwendungen einzelner Häufigkeitsklassen zu dieser Struktur. Das Gedicht ist in vierzeilige Strophen aufgeteilt. Alle Verse der Strophe reimen sich (aaaa). Jeder Vers besteht aus 16 Silben, wobei in der Mitte des Verses (nach der achten Silbe, die immer die letzte Silbe des Wortes ist) eine deutliche Zäsur vorhanden ist:

romelman šekmna samḳaro/ʒzalita mit ʒzlierita,
zegardmo arsni sulita/ḳuna zecit monaberita,
čven, ḳacta, mogvca kveḳana,/gvakvs utvalavi perita,
misgan ars ḳovli xelmcipe/saxita mis mierita.

Deswegen war es sinnvoll, alle Wörter des Gedichtes nach ihrer Position im Gedicht zu klassifizieren. Es wurden 5 strukturelle Wortklassen aufgestellt:

1) Wörter am Ende des Verses (E)
2) Wörter am Ende der ersten Vershälfte (vor der Zäsur) (E')
3) Wörter am Versanfang (A)
4) Wörter am Anfang der zweiten Vershälfte (A')

Auf jede von diesen Klassen entfallen 15.4% aller Wörter des Gedichts.

5) Die restlichen Wörter (R), die 38.4% aller Wörter ausmachen (vgl. Abb. 1a)

Wenn die Wörter beliebiger Häufigkeitsklassen keiner bestimmten Tendenz folgen, d.h. völlig zufällig unter den Strukturklassen verteilt sind, dann müßte das Histogramm ihrer Verteilung annähernd so sein wie in Abb. 1a. Für die Untersuchung der statistichen Übereinstimmung der strukturellen Klassen und

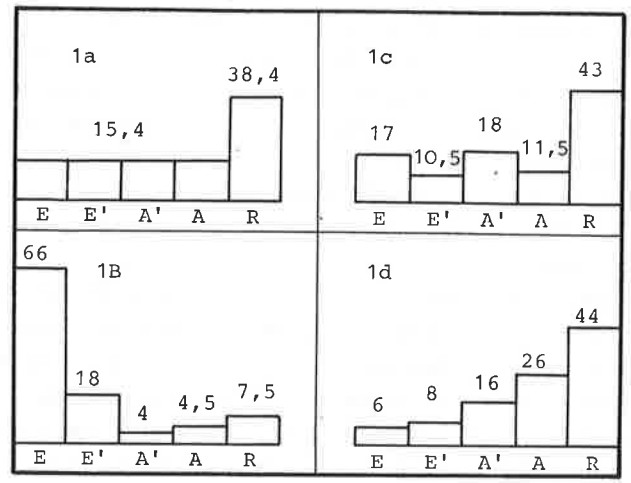


Abb. 1. Häufigkeiten struktureller Wortklassen

der Häufigkeitsklassen haben wir Wörter aus drei Häufigkeitsklassen erhoben: aus der ersten, die nur einmal vorkommende Wörter enthält (wir haben 538 solcher Wörter erhoben), aus der fünften - dies ist irgendeine "mittlere Klasse", die 16 bis 31 mal vorkommende Wörter enthält (wir erhoben 29 solche Wörter, die insgesamt 694 mal vorkamen), aus der neunten - dies ist die Klasse der "häufigen" Wörter, die 256 bis 511 mal vorkamen. Aus der neunten Klasse haben wir 2 Wörter mit voller Bedeutung genommen: *tkma* (sagen), das 447 mal vorkam und *guli* (Herz), das 317 mal vorkam. Diese beiden Wörter wurden im Gedicht 764 mal verwendet. Wörter aus der zehnten, der häufigsten Klasse haben wir nicht erhoben, da sie nur aus Hilfswörtern besteht. So wurde jede aus den gewählten Klassen mit über 500 Wortverwendungen repräsentiert.

Die Resultate der Verteilung der Wörter aus der ersten Häufigkeitsklasse in unterschiedliche strukturelle Klassen sind in % in der Abb. 1b dargestellt. Die Abbildungen 1c und 1d stellen die analogen Resultate für die fünfte bzw. neunte Klasse dar.

Wie man sieht, nur die Wörter der fünften Häufigkeitsklasse haben keine bemerkenswerte Tendenz, da sie auf die strukturellen Klassen zufällig aufgeteilt sind (vgl. Abb. 1a und 1c). Die erste und die neunte Häufigkeitsklasse haben sehr markante, auch wenn gegensätzliche Tendenzen. Während sich die seltenen Wörter (1. Häufigkeitsklasse) vor allem am Versende und auch am Ende der Halbverse konzentrieren (insgesamt entfallen auf E und E' 84.2% der Wortverwendungen), meiden häufige Wörter diese strukturelle Klasse.

Wenn also bei vollkommen zufälliger Wahl eines Wortes aus dem Text alle Häufigkeitsklassen in stabilem Gleichgewicht stehen, sind die extremen Klassen in der Struktur des Textes sehr ungleichmäßig verteilt. Es besteht Grund zur Annahme, daß die Verteilung der zweiten, dritten und vierten Häufigkeitsklasse zwischen 1b und 1c liegen, während die der sechsten, siebten und achten zwischen 1c und 1d liegen.

Die Konzentration seltener Wörter am Versende bezeugt die Sorgfältigkeit bei der Reimwahl und Abneigung gegen Reimwiederholung. Dieser Umstand wurde merkwürdigerweise von Majakovskij (1958) bei seiner eigenen poetischen Erfahrung entdeckt: „Der Reim verbindet die Verse, deswegen muß sein Material kompakter sein als das des Versrestes. Ich selbst stelle das charakteristischste Wort an das Versende und finde den Reim trotz aller Schwierigkeiten".

Wie aus unseren Resultaten folgt, scheint diese Selbstbeobachtung des großen Dichters nicht nur seine eigene Arbeitsmethode zu sein, sondern offensichtlich ein allgemeines Prinzip der Organisation eines poetischen Textes. Es ist wichtig hervorzuheben, daß man die Genauigkeit dieser Organisation mit statistischen Methoden auswerten kann.

Anmerkung

1. Als unterschiedliche Wörter haben wir diejenigen bezeichnet, die entweder unterschiedliche Bedeutung haben, oder diejenigen, die zwar dieselbe Bedeutung haben, aber zu unterschiedlichen Wortarten gehören. Als identisch haben wir diejenigen Wörter betrachtet, die sowohl dieselbe Bedeutung haben und gleichzeitig zu derselben Wortart gehören. Eine Ausnahme aus diesem gemischten semantisch-morphologischen Kriterium bilden Synonyma, die man als unterschiedliche Wörter betrachtet. Die Wortzählung erfolgte nach Šanidze (1956).

ÜBER EINIGE STATISTISCHE BESONDERHEITEN MUSIKALISCHER NACHRICHTEN

M.G. Boroda, Ju.K. Orlov

Die modernen Methoden der statistischen Musikanalyse (Hiller & Isaacson 1963; Pierce 1967) nehmen an, daß es möglich ist, Melodien durch ein Urnenmodell zu simulieren. In der vorliegenden Untersuchung wird gezeigt, daß einige statistische Besonderheiten von Melodien nicht durch sukzessive Ziehungen aus einer Urne modelliert werden können.

Es wurden die Häufigkeiten der Melodieintervalle und ihrer Kombinationen (Folgen) zu 2, 3, 4 und 5 hintereinander in den Melodien F. Chopins gezählt. (Es wurden Themen aus ungefähr 3/4 des gesamten Werkumfanges des Komponisten ausgewählt; der Umfang der Stichprobe macht etwa 10^4 Melodieintervalle aus.) Es stellte sich heraus, daß die nach abnehmender Häufigkeit geordnete Reihe dem verallgemeinerten Zipf-Mandelbrotschen-Gesetz gehorcht (Orlov 1970). Unter Verwendung dieser Angaben untersuchten wir die Verteilung der Intervalle und ihrer Folgen mit unterschiedlichen Häufigkeiten hinsichtlich des Metrums (d.h. der Strukturen des Taktes) in den Melodien. Jedem Melodieton wurde die Häufigkeit der Intervallfolge, die mit dem betreffenden Ton endet, gegenübergestellt. So wurde die Melodie als eine Zahlenfolge der Häufigkeiten der sie bildenden Töne angesehen, wobei die Taktstruktur erhalten blieb, d.h. für jede dieser Häufigkeiten wurde gezeigt, in welchem Teil des Taktes sich der ihr entsprechende Ton befindet.

Es wurde festgestellt, daß seltene Töne überwiegend in den ersten (starken) Teil eines Taktes fallen. Diese Beobachtung wurde an den Mazurken Chopins überprüft. Die Ergebnisse sind in der Tabelle dargestellt.

Um deutliche Unterschiede zu erhalten, wurden die Töne in drei Gruppen eingeteilt: die relativ seltenen (mit den relativen Häufigkeiten von π_1 bis π_1'), die relativ häufigen (mit den Häufigkeiten von π_3 bis π_3') und die Gruppe der Töne mit mittleren Häufigkeiten.

Für jeden einzelnen Taktteil wurde für alle Gruppen die relative Häufigkeit $\bar{p}$ berechnet, mit der ein Ton in den gegebenen Taktteil fällt. Die Zählung wurde in drei Varianten durchgeführt: in der ersten Variante gilt als Häufigkeit eines Tones die unabhängige Häufigkeit des entsprechenden Melodieintervalles, in der zweiten die Häufigkeit einer Folge aus zwei Intervallen, die ihm vorangehen, und in der dritten die Häufigkeit einer Folge aus drei Intervallen. Es ist ganz offensichtlich, daß in jeder Variante der Zählung sich die Grenzen der Häufigkeiten für die Intervallgruppen ändern müssen.

Taktteil	Häufigkeit des Tones im Taktteil	Zufällige Verteilung	seltene Gruppe			mittlere Gruppe			häufige Gruppe		
			Variante I	II	III	I	II	III	I	II	III
			$\pi_1=10^{-1}$	$\pi_1=10^{-1}$	$\pi_1=\pi_1=10^{-1}$	$\pi_2=6,4\cdot10^{-3}$	$\pi_2=3,2\cdot10^{-3}$	$\pi_2=8\cdot10^{-4}$	$\pi_3=10^{-4}$	$\pi_3=2,5\cdot10^{-3}$	$\pi_3=3,2\cdot10^{-3}$
			$\pi_1=3,1\cdot10^{-3}$	$\pi_1=7\cdot10^{-4}$		$\pi_2=2,5\cdot10^{-3}$	$\pi_2=6,3\cdot10^{-3}$	$\pi_2=1,5\cdot10^{-3}$	$\pi_3=1,5\cdot10^{-4}$	$\pi_3=5\cdot10^{-3}$	$\pi_3=2,5\cdot10^{-3}$
stark	p_1	0,3571	0,4110	0,4078	0,4460	0,3106	0,3257	0,3903	0,3198	0,2647	0,2389
stark	p	0,3723	0,5150	0,4550	0,5020	0,3520	0,3661	0,4422	0,3450	0,2940	0,2700
stark	p_2	0,3869	0,6190	0,5022	0,5580	0,3934	0,4063	0,4937	0,3702	0,3233	0,3011
rel.stark	p_1	0,3412	0,1611	0,2866	0,2081	0,3164	0,2398	0,1895	0,3316	0,4060	0,3815
rel.stark	p	0,3563	0,2150	0,3333	0,2573	0,3580	0,2732	0,2338	0,3570	0,4380	0,4160
rel.stark	p_2	0,3708	0,2629	0,3734	0,3059	0,3996	0,3062	0,2781	0,3824	0,4700	0,4505
schwach	p_1	0,2445	0,2217	0,1773	0,1936	0,2580	0,3208	0,2752	0,2735	0,2380	0,2815
schwach	p	0,2714	0,2700	0,2150	0,2417	0,2900	0,3607	0,3240	0,2980	0,2680	0,3140
schwach	p_2	0,2983	0,3285	0,2527	0,2884	0,3220	0,4012	0,3728	0,3225	0,2850	0,3475

Wenn sich die Töne einer beliebigen Gruppe völlig zufällig auf die Taktteile verteilen würden (ein solches Ergebnis muß man für eine zufällige Folge erwarten, die man entweder durch unabhängige Experimente erhalten hat, oder durch Experimente, die von Resultaten vorangegangener Experimente abhängig sind, jedoch nicht von der Position des generierenden Elementes in

einem bestimmten äußeren Schema) dann würden sie auf jeden Taktteil ungefähr zu gleichen Teilen fallen ($\bar{p} \approx 1/3$). Die Berücksichtigung der Synkopen (Fehlen einer Note in einem Taktteil) erlaubte es, für diesen Fall eine genauere Verteilung zu erstellen (Spalte "Zufällige Verteilung" in der Tabelle). Für jede Verteilung wurden 95% Konfidenzintervalle (p_1, p_2) nach Student gebildet. Wie aus der Tabelle ersichtlich ist, unterscheidet sich die Verteilung der Gruppen von Tönen auf die Taktteile auffallend von einer Zufallsverteilung und besonders bemerkenswert ist die Abweichung vom erwarteten Wert bei den Häufigkeiten, mit denen die Töne aus der seltenen Gruppe in den starken (ersten) Taktteil fallen ($p = 0.6190$).

Daher kann die beschriebene Besonderheit von Melodien nicht durch ein gewöhnliches Urnenmodell dargestellt werden, da die Wahrscheinlichkeit der Wahl des nächst folgenden Elementes einer zufälligen Folge von Tönen von seiner Position (d.h. vom Taktteil) abhängt. Einen solchen zufälligen Prozeß kann man sich vorstellen als das Ergebnis aufeinanderfolgender zufälliger Ziehungen aus speziellen Urnen, wobei die Reihenfolge der Urnen nicht zufällig ist.

Eine analoge Abhängigkeit der Wahrscheinlichkeit des Elementes von seiner Lage wurde im literarischen Material (vgl. Nadarejšvili & Orlov 1969) gezeigt. Dies erlaubt, Informationsprozessen ähnlichen Typs eine größere Allgemeinheit einzuräumen.

Die Autoren danken L.G. und M.G. Kevlišvili und Ju.G. Boroda für ihre Hilfe bei der Codierung des Notenmaterials für die EDV.

ÜBER DEN CHARAKTER DER VERTEILUNG VON INFORMATIONSEINHEITEN GERINGER HÄUFIGKEIT IN KÜNSTLERISCHEN TEXTEN

M.G. Boroda, I.Š. Nadarejšvili,
Ju.K. Orlov, R.Ja. Čitašvili

1

Bei der Erstellung von statistischen Textmodellen wird gewöhnlich von der Hypothese der Homogenität dieser Texte ausgegangen, d.h. der Unabhängigkeit der Wahrscheinlichkeit des Erscheinens einer gegebenen Informationseinheit von ihrer Lage im Text (Kalinin 1964, 1965). Jedoch zeigte die Überprüfung dieser Hypothese an literarischen Texten (Tokarev/Jakubajtis 1969, Bektaev/Luk'Janenkov 1971), daß sie zumindest für die relativ häufigen Wörter nicht haltbar ist: die Häufigkeiten dieser Wörter lassen statistisch signifikante Abweichungen von dem vorgeschlagenen statistisch homogenen Modell erkennen. Die in diesen Arbeiten angewandte Methodik erlaubt es lediglich, ein negatives Resultat zu formulieren, gibt jedoch nicht die Möglichkeit, irgendwelche bestimmten Konsequenzen über den tatsächlichen Charakter der Verteilung der Wörter im Text zu ziehen. In der vorliegenden Arbeit wird eine spezielle Methodik für die Untersuchung des Charakters der Verteilung der am seltensten auftretenden Elemente in künstlerischen Texten vorgeschlagen. Die Wichtigkeit dieser Frage hängt damit zusammen, daß die seltenen Wörter gewöhnlich einen bedeutenden Teil des Wortschatzes eines beliebigen zusammenhängenden Textes ausmachen.

Wir gehen von der Hypothese aus, daß die selten vorkommenden Wörter eine Tendenz zur Häufung zeigen müssen, die darin begründet ist, daß die Verwendung dieser Wörter auf lokale Episoden des Textes beschränkt ist. Diese Hypothese kann man einfach so überprüfen, daß man die beobachtete Verteilung der Abstände zwischen den Elementen einer bestimmten Häufigkeit in dem gegebenen Text

mit der theoretischen Verteilung dieser Abstände vergleicht, die man unter der Annahme der statistischen Homogenität erhalten hatte.

Als Gesamtcharakteristik der Verteilung aller Wörter, von denen jedes in einem gegebenen Text n-mal vorkommt, wählt man einfach die zufällige Größe τ, den Abstand zwischen der ersten und der letzten Verwendung jedes einzelnen Wortes aus der betrachteten Gruppe von Wörtern mit derselben Häufigkeit. In einem statistisch homogenen Text ist die Wahrscheinlichkeit, daß $\tau = l$ ist (wenn ein Element m-mal in einer Stichprobe des Umfangs N vorkommt) proportional zur Zahl der möglichen Permutationen von m gleichen Elementen auf N Positionen bei einem Abstand l zwischen den äußeren Elementen:

$$P\{\tau = l\} = \frac{N - l}{C_N^m} C_{l-1}^{m-2} . \qquad (1)$$

Da N für künstlerische Texte große Werte annimmt, erschien es zweckmäßig, zu der normierten Größe $x = \frac{\tau}{N}$ überzugehen und die Dichte der Verteilung von x bei großen N zu berechnen. Aus der Formel (1) kann man unter Verwendung der Stirlingschen Formel einen Ausdruck für die Grenzverteilung erhalten:

$$f(x) = m(m - 1)x^{m-2}(1 - x) . \qquad (2)$$

Die mathematische Erwartung und die Varianz der Verteilung (2) sind wie folgt:

$$E(x) = \frac{m - 1}{m + 1} \qquad (3)$$

$$\sigma^2(x) = \frac{2(m - 1)}{(m + 1)^2(m + 2)} . \qquad (4)$$

Die beiden letzten Ausdrücke erlauben es, ein Konfidenzintervall nach der "Drei-Sigma"-Regel aufzustellen:

$$\bar{x} \pm 3 \frac{\sigma(x)}{\sqrt{k}} ,$$

wobei k die Zahl der verschiedenen Wörter ist, die im Text m-mal vorkommen und $\bar{x}$ das arithmetische Mittel der beobachteten Werte $x = \frac{\tau}{N}$ für die Wörter aus der betrachteten Gruppe (der m-mal vorkommenden Wörter).

2

Als Versuchsmaterial wurde ein Auszug aus "Die rechte Hand des Großen Meisters" von K. Gamsachurdia (im georgischen Original) verwendet, nämlich die ersten 10000 Wortverwendungen vom Textanfang an. Dieser Auszug wurde auf durchnumerierte Kärtchen herausgeschrieben. Dadurch war es möglich, die Positionen des ersten und des letzten Auftretens eines gegebenen Wortes festzustellen. Die Differenzen der äußersten Positionsnummern, geteilt durch den Umfang der Stichprobe, ergaben die Werte x, für die dann der Mittelwert für alle k der m-mal vorkommenden Wörter gebildet wurde. Die so erhaltenen Resultate sind in der Tabelle 1 den entsprechenden theoretischen Werten gegenübergestellt.

Tabelle 1

m	k	$\bar{x}$	E(x)	$\sigma(x)$	$3\frac{\sigma(x)}{\sqrt{k}}$	$\frac{E(x)-\bar{x}}{3\frac{\sigma(x)}{\sqrt{k}}}$
2	174	0.243	0.333	0.236	0.0535	1.68
3	244	0.346	0.500	0.224	0.0430	3.58
4	129	0.232	0.600	0.200	0.0527	3.18
5	82	0.552	0.657	0.178	0.0590	1.25
6	55	0.574	0.715	0.160	0.0646	2.18
7	47	0.676	0.750	0.144	0.0632	1.17
8	33	0.639	0.778	0.132	0.0638	2.06
9	25	0.587	0.800	0.121	0.0724	2.94
10	15	0.660	0.818	0.111	0.0857	1.84

Man sieht, daß für alle m = 2,3,...,10 die Größe $\bar{x}$ wesentlich kleiner als der erwartete Wert E(x) ist, d.h. die Wörter geringer Häufigkeit haben in der Tat eine Tendenz zur Häufung innerhalb

des Textes. Dies erschüttert eine der Schlußfolgerungen von Tokarev und Jakubajtis (1969):

"... in dem Maße, wie wir uns in den Bereich immer kleinerer Häufigkeiten begeben, treten immer mehr Wörter auf, für die die Poisson-Hypothese zuzutreffen scheint." Man kann sich vorstellen, daß eine solche Schlußfolgerung durch die unzureichende Empfindlichkeit der Methodik bedingt ist, die in jener Arbeit bei der Untersuchung des Bereiches geringer Häufigkeiten verwendet wurde.

Analoge Ergebnisse erhielt man auch an musikalischem Material. Auf der Basis der elementaren Melodieeinheit "F-Motiv" (vgl. Boroda 1973) wurde die Melodielinie - die Hauptstimme vom Anfang bis zum Ende des Textes - in musikalischen Texten unterschiedlicher Stilrichtungen untersucht.[1)]

Die Melodielinie jedes Textes wurde in F-Motive eingeteilt und diese wurden auf durchnumerierte Kärtchen notiert, so wie man es in literarischen Texten mit den Wörtern gemacht hatte. Als identisch gelten zwei F-Motive dann und nur dann, wenn die Intervallfolgen beider Tonfolgen gleich sind und die Tonlängen unverändert bleiben (d.h. man kann ein F-Motiv aus dem anderen durch Transponieren erhalten).

Die Anzahl aller unterschiedlichen F-Motive bildete das F-Motivinventar (Intonationsbestand) der Melodielinien. Für jedes F-Motiv des Intonationsbestandes wurde die Häufigkeit seines Vorkommens in der gegebenen Melodielinie bestimmt sowie die Positionsnummern seiner ersten und letzten Verwendung in ihr. Dann wurde - gesondert für die zwei-, drei-, vier- und fünfmal vorkommenden F-Motive - in jedem der untersuchten Texte die Differenz der Nummern der ersten und letzten Verwendung des gegebenen F-Motivs bestimmt. Von diesen Größen wurde - wie bei den literarischen Texten - der Mittelwert (für ein gegebenes m) gebildet. Weiterhin wurden die Werte $E(x)$, $3\sigma(x)/\sqrt{k}$ bestimmt (s. Tabelle 2).

Wie aus der Tabelle 2 ersichtlich, ist in den untersuchten musikalischen Texten der mittlere Abstand $\bar{x}$ zwischen den ersten und letzten Verwendungen der F-Motive wesentlich kleiner als

Tabelle 2

m	k	$\bar{x}$	$E(x)$	$\sigma(\bar{x})$	$3\frac{\sigma(x)}{\sqrt{k}}$	$\frac{E(x)-\bar{x}}{3\frac{\sigma(x)}{\sqrt{k}}}$
			W. MOZART: Sonate Nr. 10 (für Klavier)			
2	98	0.125	0.333	0.236	0.0716	2.9
3	14	0.208	0.500	0.224	0.1795	1.63
4	39	0.250	0.600	0.200	0.0962	3.64
5	7	0.244	0.667	0.178	0.202	2.14
			F. CHOPIN: Ballade Nr. 1			
2	35	0.0684	0.333	0.236	0.12	2.2
3	9	0.347	0.500	0.224	0.224	0.645
4	13	0.224	0.600	0.200	0.167	2.26
5	3	0.564	0.667	0.178	0.308	0.334
			P. HINDEMITH: Sonate für Solovioline Nr. 1			
2	66	0.0444	0.333	0.236	0.088	3.32
3	24	0.0436	0.500	0.224	0.137	3.34
4	12	0.0596	0.600	0.200	0.174	3.12
5	15	0.0845	0.667	0.178	0.138	4.12

der Erwartungswert $E(x)$. Zum Beispiel ist in der Ballade Chopins $E(x)/\bar{x}$ für die zweimal vorkommenden F-Motive gleich 4.87, in der Sonate von Hindemith gleich 7.50 usw. Anders gesagt, die F-Motive, die in einem musikalischen Text selten vorkommen, zeigen - wie auch die seltenen Wörter in literarischen Texten - eine eindeutige Tendenz zur Häufung in bedeutend engeren Textabschnitten, als es aus dem Modell eines statistisch homogenen Textes folgen würde. Analoge Erscheinungen wurden auch in einer Reihe anderer homophoner Texte unterschiedlicher Stilrichtungen des 17. bis 20. Jahrhunderts beobachtet ("Die Abschiedssinfonie" von Haydn, das "Rondo C-Dur" von Beethoven, die "Sonatine für Flöte" von Levitin u.a.).

3

Das bei den seltenen Elementen beobachtete Phänomen der Häufung muß in erster Linie einen Einfluß auf die Zunahme der Anzahl unterschiedlicher seltener Elemente mit der Zunahme der Textlänge ausüben.

Diese Häufungserscheinung muß - zumindest in den Anfangsteilen von Texten - das Tempo der Zunahme der Anzahl unterschiedlicher Elemente ein wenig verringern. Offensichtlich erklärt sich gerade hierdurch der Unterschied im Anwachsen des Wortschatzes zwischen dem untersuchten Auszug aus dem Text Gamsachurdias und einer Stichprobe aus demselben Text, die jedoch nicht mehr aus einem zusammenhängenden Text bestand. Stattdessen wurde jede vierte Seite vom Textanfang an ausgewählt (Tabelle 3). Eine solche Stichprobenerhebung mußte den Häufungseffekt neutralisieren und das Wachstumstempo des Wortschatzes beschleunigen. Die Berechnung der Konfidenzintervalle nach der Formel für die Varianz des Wortschatzes laut Kalinin (1965: Formel 29) zeigte, daß der beobachtete Unterschied statistisch signifikant ist.

Tabelle 3

Umfang der Stichprobe	Zahl der verschiedenen Wörter	
	in den ersten 10000 Wörtern vom Textanfang an	auf jeder vierten Seite vom Textanfang an
1000	625	671
2000	1066	1126
3000	1437	1545
4000	1735	1890
5000	2035	2243
6000	2315	2661
7000	2551	2956
8000	2817	3256
9000	3102	3529
10000	3320	3799

Für musikalische Texte wurde ein anderes Verfahren zur Überprüfung des Einflusses der Häufung seltener Elemente auf das "Wortschatzwachstum", d.h. auf den Intonationsbestand, benutzt. Wie schon erwähnt, kann man in homophonen Texten jedem F-Motiv der Hauptstimme des Textes die Nummer des ersten Auftretens des gegebenen Motivs in dieser Stimme zuordnen. Dies erlaubt es, für eine beliebige Stichprobe eines Textes die Zahl der verschiedenen F-Motive in dieser Stichprobe anzugeben und eine experimentelle Kurve des Anwachsens ihrer Anzahl mit zunehmender Entfernung vom Textanfang zu erstellen.

Unter Verwendung der Formel Kalinins, die es erlaubt, den Wortschatz für eine gegebene Stichprobe zu berechnen, und ausgehend von der Hypothese der statistischen Homogenität eines Textes [Formel (10) aus Kalinin (1965)], erstellten wir für jeden der untersuchten homophonen Texte die theoretische und die experimentelle Kurve des Anwachsens des Intonationsbestandes. In der Abb. 1 ist eine der typischen Kurven dargestellt. Anhand der Grafik läßt sich klar verfolgen, daß die beobachteten Werte des Intonationsbestandes verglichen mit den theoretischen Werten in den ersten 2/3 des Umfanges der Hauptstimme des Textes kleiner sind (Abb. 1). In den übrigen Texten beobachtet man auch eine analoge Abweichung der experimentellen Kurve von der theoretischen, was durch die statistische Heterogenität des Textes hervorgerufen wird.

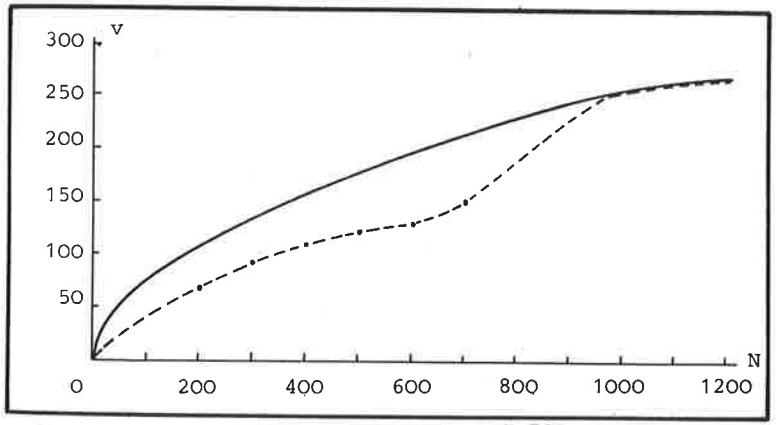


Abb. 1. MOZART: Sonate Nr. 10 für Klavier

Anmerkung

1) Für die Untersuchung wurden homophone Texte ausgewählt, da in ihnen die "Hauptstimme" vom Anfang des Textes bis zum Ende klar abgegrenzt werden konnte. In solchen Texten konnte man jedem der verschiedenen F-Motive, die in die "Hauptstimme" eingehen, die Nummer seines ersten und letzten Auftretens in dieser Stimme zuschreiben - in Analogie zu der Vorgehensweise, die bei literarischen Texten angewendet wurde.

Nichtstationäre Erscheinungen im Prozess der Text-generierung

G.Š. Nadarejšvili, I.Š. Nadarejšvili, Ju.K. Orlov

Die in Orlov (1977) dargelegte Übereinstimmung der Häufigkeitsstrukturen des Wortschatzes eines literarischen Werkes mit seiner Länge wirft eine ganze Reihe von Fragen auf. In erster Linie stellt sich die Frage: w i e kommt eine solche Übereinstimmung zustande? Da ja die Notwendigkeit dieser Übereinstimmung die Wahl einer bestimmten Kurve der Vokabularzunahme festlegt, so bedeutet dies also, daß gleich von Anfang an eine festgelegte Strategie der Zunahme neuer Wörter im Text ausgewählt werden muß.

Es sind zwei Fälle möglich. Entweder werden erst die Wahrscheinlichkeiten aller Wörter des potentiellen Wortschatzes eines Textes festgelegt und dann die Wörter in Übereinstimmung mit diesen Wahrscheinlichkeiten gewählt, oder es müssen bestimmte Kontroll-Mechanismen existieren, die das tatsächliche Wachsen des Wortschatzes eines Textes mit dem vorgeschriebenen vergleichen und die notwendige Korrektur vornehmen. Im ersten Fall muß die Wahrscheinlichkeit des Auftretens irgendeines Wortes nicht von seiner Stellung im Text abhängen; die Distribution der Positionen, die von dem jeweiligen Wort im Text eingenommen werden, muß vollständig zufällig sein.

Die von Tokarev & Jakubajtis (1969) und Bektaev & Luk'janenkov (1971) durchgeführte Überprüfung dieser Vermutung bestätigte diese Hypothesen nicht, zumindest nicht im Bereich der häufigen Wörter. Aber da ja die häufigen Wörter einen verhältnismäßig geringen Teil des Wortschatzes eines Textes ausmachen, spielen die seltenen Wörter die Hauptrolle im Prozeß der Vokabularzunahme. In Boroda & Nadarejšvili & Orlov & Čitašvili (1977) wurde die Homogenität eines Textes im Bereich der seltenen Wörter, die zweimal, dreimal, usw. in der untersuchten Stichprobe vorkommen, überprüft. Es wurde der Abstand zwischen

dem ersten und dem letzten Auftreten eines jeden Wortes im Text als die Zahl der Wortverwendungen zwischen diesen Vorkommen gemessen; dann wurde der mittlere Abstand für alle Wörter der gegebenen Häufigkeit berechnet (zum Beispiel für alle zweimal vorkommenden Wörter) und mit dem erwarteten Abstand verglichen, der unter der Annahme der statistischen Homogenität eines Textes (d.h. der Unabhängigkeit des Wortgebrauchs) errechnet worden war. Da es vorteilhaft ist, mit der normierten Größe x zu arbeiten, die das Verhältnis des Abstandes zum Umfang der Stichprobe darstellt, so ergibt sich die mathematische Erwartung des (normierten) mittleren Abstandes E(x) für die Wörter mit der Häufigkeit m als

$$E(x) = \frac{m - 1}{m + 1} \qquad (1)$$

und die Varianz als

$$\sigma^2(x) = \frac{2(m - 1)}{(m + 1)^2(m + 2)} \qquad (2)$$

Diese zwei Ausdrücke erlauben, ein Konfidenzintervall

$$\bar{x} \pm 3 \frac{\sigma(x)}{\sqrt{k}}$$

nach der "Drei-Sigma-Regel" zu konstruieren, wobei k die Zahl der verschiedenen m-mal vorkommenden Wörter ist und $\bar{x}$ das arithmetische Mittel der beobachteten x-Werte für ein gegebenes m.

Es wurden sowohl literarische als auch musikalische Texte untersucht, und in allen Fällen wurden statistisch zuverlässige Häufungen der selten auftretenden Wörter festgestellt, da sich der mittlere Abstand als wesentlich geringer erwies als der erwartete, der nach der Formel (1) berechnet worden war. Diese Erscheinung wurde für die Wörter mit den Häufigkeiten bis m = = 10 - 15 beobachtet (die höheren Zahlen k der Wörter mit der gegebenen Häufigkeit erwiesen sich als unzuverlässig für eine

statistische Aussage).

Im Bereich der selten vorkommenden Wörter fehlt also die statistische Homogenität, da der Gebrauch eines gegebenen Wortes in einer bestimmten Position eine Erhöhung der Wahrscheinlichkeit des Vorkommens dieses Wortes in der Umgebung dieser Position hervorruft.

Analoge Berechnungen wurden auch auf der Ebene der Silben durchgeführt. In dem Gedicht "Der Totengräber" von G. Tabidze zeigte sich eine deutliche Häufung (Tabelle 1):

Tabelle 1

m	k	$\bar{x}$	E(x)	σ(x)	$3\frac{\sigma(x)}{\sqrt{k}}$	$\frac{E(x)-\bar{x}}{3\frac{\sigma(x)}{\sqrt{k}}}$
2	97	0.209	0.333	0.236	0.072	1.73
3	35	0.450	0.500	0.224	0.113	0.44
4	23	0.413	0.600	0.200	0.125	1.49
5	22	0.500	0.667	0.178	0.113	1.48

Die Hauptcharakteristik erscheint in der letzten Spalte, die das Verhältnis der Differenz zwischen dem erwarteten und dem beobachteten Wert zum Konfidenzintervall darstellt. Wenn diese Größe im absoluten Wert Eins übersteigt, dann ist die beobachtete Divergenz signifikant. Wie man aus der Tabelle 1 ersieht, kann nur im Falle der dreimal vorkommenden Silben die beobachtete Häufigkeit $\bar{x}$ = 0.450 gegenüber dem erwarteten Wert E(x) = = 0.5 durch zufällige Ursachen erklärt werden.

In dem gegebenen Fall kann man tatsächlich die Häufung durch die Gruppierung der Reime, der Assonanzen und der Alliterationen in dem poetischen Text erklären. Deshalb wurde eine Kontrollberechnung an einem Prosaauszug gleichen Umfangs (N = = 1762 Silben) desselben Autors durchgeführt (die Erzählung "Schlachten"; alle Berechnungen wurden am georgischen Originaltext durchgeführt). Die Ergebnisse der Berechnungen sind in der Tabelle 2 dargestellt. Obwohl auch in diesem Fall die mittleren Werte des Abstandes niedriger sind als die erwarteten, kann

man jedoch nicht von einer statistischen Signifikanz dieser Erscheinung sprechen.

Tabelle 2

m	k	$\bar{x}$	E(x)	σ(x)	$3\frac{\sigma(x)}{\sqrt{k}}$	$\frac{E(x)-\bar{x}}{3\frac{\sigma(x)}{\sqrt{k}}}$
2	65	0.309	0.333	0.236	0.088	0.275
3	26	0.373	0.500	0.224	0.130	0.980
4	23	0.552	0.600	0.200	0.125	0.384
5	14	0.593	0.667	0.178	0.143	0.518

Ebensolche Berechnungen in den Texten A.S. Puškins (das Gedicht "Napoleon" mit einem Umfang von 1017 Silben; Auszüge gleichen Umfangs aus "Die Kapitänstochter" und "Die Geschichte des Pugačevskij-Aufstandes") zeigten einen ähnlichen Unterschied zwischen Prosa und Poesie nicht. Alle mittleren Abstände zwischen den ersten und letzten Verwendungen von Silben der gegebenen Häufigkeit befanden sich in den Grenzen des Konfidenzintervalls. Jedoch zeigte eine detailliertere Untersuchung der Abstände zwischen den zweimal vorkommenden Silben in dem Gedicht "Napoleon", daß die Anzahl der nahe beieinanderliegenden Silben (zwischen denen der Abstand 200 Silben nicht übersteigt) signifikant größer als die erwartete ist. Aber dieser Effekt wird durch die entgegengesetzte Erscheinung, nämlich, durch die erhöhte Anzahl der weit auseinander gelegenen Silben, verschleiert. In der Prosa jedoch unterscheidet sich die Verteilung der Intervallängen zwischen den Silben nicht von einer zufälligen Verteilung. So macht sich auch in Texten A.S. Puškins der Einfluß der allgemeinen Textorganisation auf seine statistische "Mikrostruktur" bemerkbar.

Auf der lexikalischen Ebene kann man in den Fällen, wo die Positionen der Wörter klar strukturiert sind, eine Neigung der Wörter irgendeiner Häufigkeit zu festgelegten Positionen bemerken. So zeigt sich zum Beispiel im "Der Held im Tigerfell" eine erhöhte Konzentration der seltenen Wörter an den Zeilenenden (Nadarejšvili & Orlov 1969); eine ähnliche Erscheinung wurde auch in den Walzern F. Chopins festgestellt (Boroda & Orlov 1970) - die seltensten Töne von Melodien konzentrieren sich statistisch signifikant in den starken Taktteilen.

Alle diese Angaben bestätigen nicht nur, daß die Verteilung der Positionen der Wörter auf den gesamten Text sich stark von einer zufälligen Verteilung unterscheidet; sie bestätigen ebenso das Vorhandensein einer Steuerung nach der Häufigkeit im Prozeß der Textgenerierung, d.h. ein Wort wird für eine gegebene Position (abgesehen von allen übrigen) auch mit Rücksicht auf seine Häufigkeit gewählt. (Vom subjektiven Standpunkt des Autors aus trägt der Prozeß möglicherweise einen entgegengesetzten Charakter: ein für eine bestimmte Schlüsselposition erfolgreich gefundenes Wort wiederholt sich im folgenden Text nicht, wodurch seine Häufigkeit beschränkt wird; jedoch kann eine solche Interpretation bei weiteren Überlegungen nicht berücksichtigt werden.)

Aber wenn man nicht von im voraus festgelegten Wahrscheinlichkeiten jedes Wortes im Text und von einer unabhängigen Auswahl der Wörter für diese oder jene Position sprechen kann, dann muß man, um die beobachtete Übereinstimmung der Häufigkeitsstruktur eines Textes mit seiner gesamten Länge zu erklären, die Existenz von Kontroll-Mechanismen anerkennen, die das Anwachsen des Wortschatzes nach einer bestimmten Kurve und (letzten Endes) die Verwirklichung der kanonischen Form des Zipf-Mandelbrot-Gesetzes im gesamten Umfang des Textes sicherstellen. Ein allgemeines Merkmal solcher Mechanismen, unabhängig von ihren konkreten Strukturen, scheinen periodische oder fast periodische Schwankungen um ein bestimmtes grundlegendes Niveau zu sein. Das heißt, auch die Kurve des Anwachsens des Wortschatzes muß solche periodischen Fluktuationen zeigen.

Aus einer Reihe von Gründen ist eine direkte Untersuchung solcher Mikrofluktuationen unmittelbar an der Kurve des Anwachsens des Wortschatzes schwierig. Der Hauptgrund liegt darin, daß die Funktion des Wortschatzwachstums in einem statistisch homogenen Text nicht abnehmend verläuft; aber die aufeinander-

folgenden Zunahmen des Wortschatzes nehmen in gleichen Abschnitten eines Textes systematisch ab. Der Prozeß wurde jedoch nur an einem einzigen Fall untersucht, und es ist unmöglich, irgendwelche Mittelwerte zu der Klasse der Prozesse zu bilden. Deshalb wurde eine Untersuchung der Verteilung der Positionen der einmal vorkommenden Wörter auf die Textlänge durchgeführt. Die Anzahl dieser Wörter ist sehr hoch (in den Auszügen aus den Texten übersteigt sie die Hälfte des Wortschatzes), daher müssen die allgemeinen Änderungen der Zunahme neuer Wörter unbedingt auch für die Menge der einmal vorkommenden Wörter gelten. Die Verteilung der Positionen der einmal vorkommenden Wörter (in der gegebenen Stichprobe) auf die Textlänge ist für einen statistisch homogenen Text gleichmäßig, was die Analyse außerordentlich vereinfacht.

Es wurden folgende Texte untersucht: "Pique Dame" von A.S. Puškin (Gesamtumfang N = 6856 Wortverwendungen), eine Nachschrift der Heldensage "Il'ja Muromec und Zar Kalin" (N = 3380) und vier Auszüge aus Texten L.N. Tolstojs, von denen jeder die ersten 10000 Wortverwendungen vom Anfang des Textes an beinhaltet - "Die Kosaken", "Krieg und Frieden", "Auferstehung" und "Kreutzersonate". Jede Stichprobe wurde in aufeinanderfolgende, sich nicht überschneidende Unterstichproben zu je 200 Wortverwendungen unterteilt, und in jeder dieser Stichproben wurde die Anzahl der Wörter berechnet, von denen ein jedes genau einmal in der ganzen Stichprobe vorkam. Dann wurde die mittlere Abweichung und die mittlere quadratische Abweichung für jede Stichprobe berechnet.

Es zeigte sich, daß die mittlere quadratische Abweichung in allen Stichproben 1.5 - 2 mal größer als die erwartete war, die man ausgehend von der Hypothese der Homogenität eines Textes berechnet hatte, und in allen Fällen überschritt sie deutlich das "Drei-Sigma"-Konfidenzintervall für diese Größe. So konnte eine statistische Homogenität eines Textes auch für die Verteilung der einmal vorkommenden Wörter nicht gezeigt werden.

Dann wurden die aufeinanderfolgenden Abweichungen der Anzahl der einmal vorkommenden Wörter in den Unterstichproben

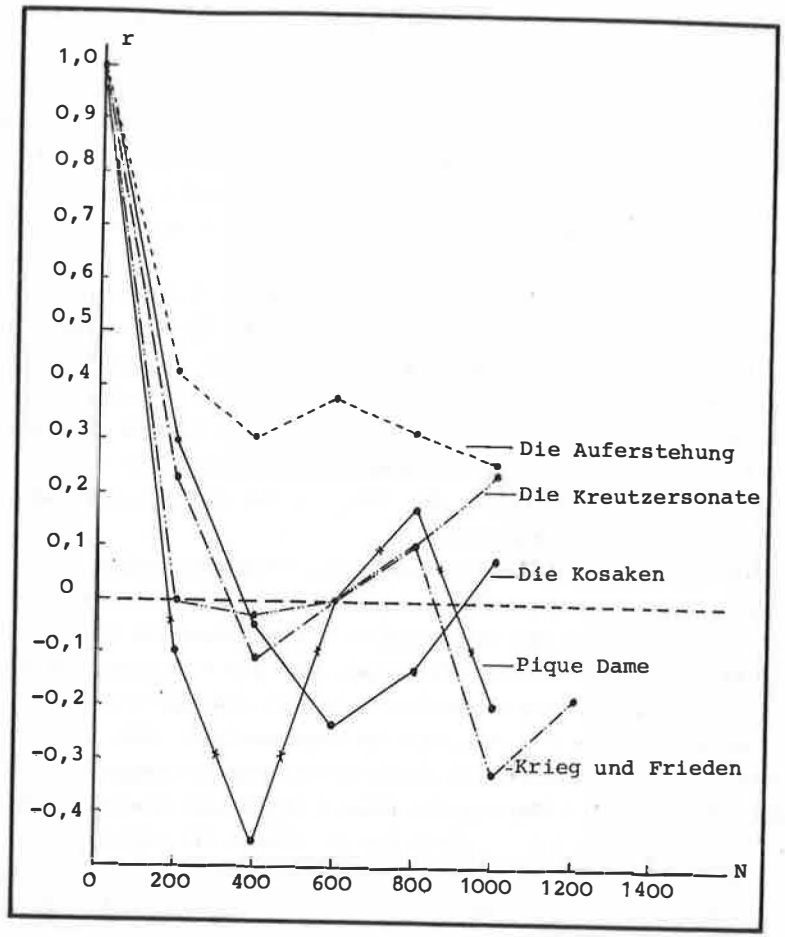


Abb. 1. Autokorrelationen der Vorzeichen der Abweichung vom Mittelwert

hinsichtlich des Mittelwertes der gegebenen Stichprobe untersucht. In der Abbildung 1 werden die Autokorrelationsfunktionen für die aufeinanderfolgenden Werte des Vorzeichens der Abweichung vom Mittelwert gezeigt. (Alle positiven Abweichungen wur-

den bei der Berechnung der Autokorrelationsfunktion mit +1, die negativen mit -1 bezeichnet). Es zeigte sich, daß die Perioden der positiven Abweichungen (d.h. der Bereiche eines Textes, in denen die Dichte der einmal vorkommenden Wörter über dem Mittelwert liegt) und die Perioden negativer Abweichungen abwechselnd vorkommen, und daß dieser Prozeß ziemlich regelmäßig ist. Zwei Texte ("Pique Dame" und "Krieg und Frieden") zeigten sogar auf zwei Halbwellen eine negative Korrelation, was von einer deutlich ausgeprägten Periodizität zeugt.

Der einzige Text, der keine ausgeprägte Periodizität bei einer stark positiven Korrelation zeigt, ist "Auferstehung" von L.N. Tolstoj. Das hängt damit zusammen, daß die Mehrzahl der negativen Abweichungen in der ersten Hälfte der Stichprobe gruppiert ist, und die Mehrzahl der positiven in der zweiten Hälfte. Wenn es also in dem gegebenen Fall doch eine Periodizität gibt, dann übersteigt die Länge ihrer Welle den Umfang der untersuchten Stichprobe.

Für die restlichen Texte liegt die Länge der Welle (die mittlere Länge der Zone der Konzentration + der Zone der Dezentration) wie aus der Abbildung 1 ersichtlich, in den Grenzen von 500 - 1000 Wortverwendungen, was 2 - 4 Seiten des gewöhnlichen Buchformats ausmacht. Wenn man die Hypothese des Vorhandenseins von Kontroll-Mechanismen annimmt, die das Anwachsen des Wortschatzes im generierten Text bewerten und korrigieren (die erste Konsequenz dieser Hypothese bestätigt die analysierte Abbildung 1), dann ist es dieser Umfang eines Textes, in dem ein Mensch die tatsächliche Richtung der Kurve des Anwachsens des Wortschatzes bewertet und notwendige Korrekturen anbringt.

Von der inhalts- und literaturwissenschaftlichen Seite kann man jedoch bemerken, daß 2 - 4 Seiten der typische Umfang kleiner Kapitel (gewöhnlich durch Ziffern gekennzeichnet) oder entfalteter Episoden in einer Erzählung sind. Es ist klar, daß am Anfang einer jeden solchen Episode im Zusammenhang mit einem neuen Handlungsort, neuen Personen usw., der Zuwachs an neuen Wörtern größer sein muß, zum Ende hin aber etwas geringer wer-

den muß. Die Reihenfolge solcher Kapitel oder Episoden kann die beobachtete Korrelation völlig erklären. Aber andererseits kann selbst das Erscheinen eines neuen Kapitels von der Notwendigkeit hervorgerufen sein, den Zuwachs neuer Wörter zu vergrößern - kann doch ein gesteigertes Anwachsen der Lexik nicht an einer im erwähnten Sinne "leeren" Stelle entstehen.

Texte mit großem Umfang, die einen erhöhten relativen Vokabularreichtum haben, weisen in der Tat eine größere Anzahl handelnder Personen, die recht kompliziert zusammenwirken und eine große Vielfalt des sozialen Milieus und Hintergrundes, in dem die Handlung abläuft, auf. Oft wird dazu noch ein Wechsel der "Standpunkte" hinzugefügt, von denen aus der Leser auf die Handlung sieht, und andere Verfahren. So kann man zu dem Schluß kommen, daß die kanonische Form des Zipf-Mandelbrotschen Gesetzes, die im gesamten Umfang eines Werkes realisiert wird, eine Art formbildender Invariante für literarische Texte darstellt, die, wenn man sich so ausdrücken kann, einige quantitative Charakteristika der Komposition einer Erzählung diktiert.

Aber welcher Art die inhaltliche Interpretation der beobachteten Erscheinungen auch sei, die statistische Analyse eines Textes enthüllt sowohl die objektiv vorhandenen Gesetzmäßigkeiten seiner Organisation und Struktur, als auch die ihn generierenden Mechanismen.

Psychologische Aspekte der quantitativen Organisation von künstlerischen Texten

M.G. Boroda, Ju.K. Orlov

Die Untersuchung der strukturellen Gesetzmäßigkeiten von Kunstwerken wie Texten, die Analyse ihres "sprachlichen" Aufbaus, wird zu einem zunehmend benutzten methodischen Weg zur Erforschung des künstlerischen Denkens, der Gesetzmäßigkeiten der Rezeption und der Produktion von künstlerischen Äußerungen. Wegen der Vielzahl der die Wahrnehmung beeinflussenden und in Wechselwirkung stehenden Faktoren ist es im künstlerischen Bereich oft sehr schwierig, ein psychologisches Experiment anzusetzen. Unter diesen Umständen ist die Strukturanalyse von künstlerischen Texten eine sehr wichtige und manchmal die einzige Quelle der Information über das künstlerische Denken: Bei diesem Vorgehen können sehr exakte Fragen nach dem Aufbau und den Beziehungen der verschiedenen Ebenen eines künstlerischen Werks formuliert werden, hier wurde ein umfangreiches quantitatives Instrumentarium der Analyse ausgearbeitet. Selbstverständlich bedürfen die auf diesem Weg zu gewinnenden Resultate noch einer vertieften **psychologischen** Interpretation.

Das Ziel der vorliegenden Arbeit ist es, Psychologen auf einige generelle quantitative Gesetzmäßigkeiten, die an der Organisation künstlerischer (literarischer, musikalischer, bildnerischer) Werke verschiedener Stilrichtungen beobachtbar sind, aufmerksam zu machen. Diese Gesetzmäßigkeiten konnten bis heute weitgehend präzisiert werden und erfordern nun eine psychologische Deutung und Interpretation. Da diese Gesetzmäßigkeiten zuerst an literarischem Material aufgedeckt wurden und dort mit einer so geläufigen Einheit wie dem Wort verbunden sind, empfiehlt es sich, sie auch vor allem an diesem Material zu erörtern.

Seit dem Ende des vorigen Jahrhunderts haben Linguisten sogenannte **Häufigkeitswörterbücher** (sowie Konkordanzen, Indizes,

Symphonien[1]) usw.) zusammengestellt.

Ein derartiges Wörterbuch ist ein Verzeichnis aller in einem Text (oder in einem Textabschnitt oder in einer Menge von Texten) auftretenden Wörter, das die Vorkommenshäufigkeit jedes Wortes (oder seine Lokation im Text) angibt. Solche Untersuchungen gründeten sich bis in jüngste Zeit auf die sogenannte "Herdansche Konzeption" (Herdan 1966), nach der jede linguistische Einheit in der Rede eine konstante Verwendungswahrscheinlichkeit hat und die beobachteten Häufigkeiten diese Wahrscheinlichkeit mehr oder weniger genau widerspiegeln. Die statistische Beschreibung der Sprache oder des Sprechflusses diene dazu, diese Wahrscheinlichkeit für jede untersuchte Einheit oder Klasse von Einheiten mit maximaler Genauigkeit zu bestimmen.

Da die Wahrscheinlichkeit traditionell als Resultat der "Überschneidung" zahlreicher nur schwach verbundener Kausalfaktoren betrachtet wird, wird die Schätzung der Wahrscheinlichkeit meist für die Schlußphase einer rein statistischen Untersuchung gehalten. Untersuchungen dieser Art haben auch eine Reihe interessanter Tatsachen aufgedeckt, und auf ihrer Grundlage konnte man kombinatorisch-probabilistische Modelle von befriedigender Realitätsanpassung konstruieren. Die Resultate finden Anwendung in der Stilanalyse (insbesondere bei der Ermittlung der Autorschaft), bei der Berechnung des Deckungsgrades eines Textes mit einem Wortschatz bestimmten Umfangs, bei der maschinellen Übersetzung, bei der Analyse des Informationsflusses in automatischen Regelungssystemen, in Informationssuchsystemen und vielen anderen Bereichen. Doch mit dem Anwachsen der Datenbasis und der Vervollkommnung der Methoden zu ihrer Bearbeitung haben sich Befunde angesammelt, die nicht in Herdans Konzeption passen. Sie widersprechen ihr nicht direkt, sondern sie bilden **eine andere "Realitätsschicht"**, in der die Wahrscheinlichkeit **als Grenzwert einer beobachtb**aren Größe (der relativen Häufigkeit) keine Rolle mehr spielt.

Wir wollen dies am Beispiel einiger anschaulicher numerischer Charakteristika eines literarischen Textes untersuchen.

"Odnaždy igrali v karty u konnogvardjejca Narumova. Dolgaja

zimnaja noč' prošla nezametno; seli užinat' v pjatom času utra." [Eines Tages spielte man beim Gardekavalleristen Narumov Karten. Die lange Winternacht verging, ehe man es bemerkte; um fünf Uhr morgens setzte man sich zum Abendessen.]

Diese beiden ersten Sätze aus "Pique Dame" enthalten 18 Wörter. Nur ein Wort tritt zweimal auf, die Präposition "v" [in]. Alle übrigen Wörter sind verschieden. D.h. der Wortschatz dieses Abschnitts ist mit v = 17 um Eins kleiner als der Umfang N = 18. Die Anzahl der genau einmal auftretenden (nicht wiederholten) Wörter ist mit v_1 = 16 noch um Eins kleiner. Der Quotient aus der Anzahl der einmaligen (nicht wiederholten) Wörter zum Wortschatz ist $v_1/v = 16/17 = 0.94$.

Was geschieht mit diesen Zahlen bei Fortsetzung der Lektüre des Textes? Der Wortschatz v wird anwachsen, da immer neue Wörter auftreten werden. Auch die Zahl der nur einmal vorkommenden Wörter wird zunehmen, aber nicht so schnell wie der Wortschatz, denn ein Teil der schon einmal aufgetretenen Wörter beginnt sich zu wiederholen und zu zweifach, dreifach usw. vorkommenden Wörtern zu werden. Folglich wird der Quotient v_1/v sinken. Beispielsweise ist in den ersten 100 Wortverwendungen von "Pique Dame" $v_1/v = 70/82 = 0.854$, in den ersten 150 Wortverwendungen $v_1/v = 84/1o7 = 0.785$. Bei N = 1256 wird $v_1/v = 374/544 = 0.688$. Bezogen auf den vollständigen Text von "Pique Dame" (N = 6861) ist $v_1/v = 1146/1928 = 0.594$.

Da der Wortbestand einer Sprache offensichtlich endlich ist, muß der Wortschatz v einer lexikalischen Stichprobe bei unbegrenztem Wachstum ihres Umfangs N gegen einen endlichen Grenzwert V, der Anzahl der unterscheidbaren Wörter der Sprache, streben. Dabei muß die Zahl der einmaligen Wörter allmählich verschwinden, denn bei unbegrenzt wachsendem Stichprobenumfang muß jedes in der Stichprobe vorkommende Wort früher oder später erneut vorkommen. Mit N gegen ∞ geht der Quotient v_1/v gegen Null.

Der Quotient v_1/v kann also in lexikalischen Stichproben jeden Zahlenwert zwischen Null und Eins annehmen, und mit wachsendem Stichprobenumfang fällt er ab. Diese Schlußfolgerung wird durch experimentelle Daten gut gestützt (siehe Tabelle 1; die

Zahlen sind aus den Arbeiten Orlov (1978a,b) entnommen, wo ihre Quellen genannt sind), auch wenn keine Stichproben bekannt sind, bei denen der Quotient einen sehr kleinen Wert annehmen würde. Nach dem Eindruck scheint es keine bevorzugten Werte für den Quotienten im Intervall [0, 1] zu geben; er zeigt eine deutliche (wenn auch nicht monotone) Tendenz, mit wachsender Stichprobengröße abzunehmen.

In der Linguostatistik ist jedoch die Ansicht verbreitet, daß das Verhältnis v_1/v nahe bei 0.5. liege. In dieser Allgemeinheit ist die Ansicht offensichtlich falsch, wie aus Tabelle 1 folgt; allerdings gibt es eine Klasse von lexikalischen Stichproben, in denen das Verhältnis tatsächlich nahe bei 0.5 liegt.

Umfang und Zusammensetzung der Stichproben in Tabelle 1 hängen vom Belieben des Untersuchers oder vom Zufall ab (vom Zufall z.B. die vollständige Ausgabe der Werke A.S. Puškins, die durch den Tod des Autors "abgebrochen" ist). Wenn wir aber die vollständigen Texte einzelner Werke untersuchen, so ergibt sich ein wesentlich anderes Bild (Tabelle 2; Daten aus den zitierten Quellen). Schon ohne eine spezielle statistische Aufbereitung kann man sehen, daß das Verhältnis v_1/v bei den einzelnen Texten tatsächlich nahe bei 1/2 liegt. Dabei fällt auf, daß der Quotient über Texte sehr verschiedener Länge annähernd konstant bleibt, während doch aus allgemeinen Erwägungen sowie aus den Daten der Tabelle 1 folgt, daß er mit wachsender Stichprobengröße fallen soll (und tatsächlich fällt er ja auch).

Es gibt also zwei gegenläufige Tendenzen. Die eine wirkt in beliebigen lexikalischen Aggregaten, die andere in hochorganisierten, im Hinblick auf die menschliche Wahrnehmung "optimierten" Nachrichten. Es ist sehr wichtig zu sehen, daß die Erfüllung der Gleichung $v_1/v \approx 0.5$ in Texten unterschiedlichen Umfangs[2] unterschiedliche Häufigkeitsorganisation dieser Texte erfordert. Bei der mathematischen Analyse der Situation (Orlov 1976), bei der rechnerischen Überprüfung des Zusammenhangs des Wortschatzwachstums mit der Zunahme des Stichprobenumfangs an Texten verschiedenen Gesamtumfangs (Nadarejš-

Tabelle 1. Analyse unvollständiger Texte

Nr.	Stichprobe	Z	v	v_1	v_1/v	F_{max}	v(Z)	$\frac{v-v(Z)}{v(Z)}$ %
1	Puškin, "Kapitänstochter" (Teil)	5000	1671	1133	0.678	202	943	+77.2
2	Puškin, Sämtliche Werke	544777	21197	6388	0.301	25026	54800	-61.3
3	Puškin, "Eugen Onegin" (Teil)	1000	590	470	0.796	46	261	+126.3
4	Russische Texte über Elektronik	200388	6826	2009	0.293	6686	22700	-69.9
5	Häufigkeitswörterbuch der russischen Sprache	1056382	39268	13379	0.341	42854	99058	-60.4
6	Häufigkeitswörterbuch militärischer Texte	689214	3000	712	0.234	20300	69500	-95.7
7	Mickiewicz, "Herr Tadeus" (Buch 1)	6587	2257	1460	0.647	247	1195	-47.5

Tabelle 2. Analyse vollständiger Texte

Nr.	Werk	Z	v	v_1	v_1/v	F_{max}	v(Z)	$\frac{v-v(Z)}{v(Z)}$ %
1	Puškin, "Kapitänstochter"	29345	4783	2384	0.498	1160	4160	+15.00
2	Rustaveli, "Held im Tigerfell"	42120	5965	2995	0.502	880	6200	- 3.79
3	Byron, "Don Juan"	130745	14411	7250	0.503	6002	15000	- 3.93
4	Byline, "Wolga und Mikula"	930	254	116	0.457	37	257	- 0.12
5	Shakespeare, "König Lear"	25471	3391	1933	0.570	984	3710	- 9.40
6	Mickiewicz, "Herr Tadeus"	64510	9250	4360	0.471	2058	8480	+ 8.30
7	Joyce, "Ulysses"	260430	29899	16432	0.549	14877	27100	+10.33

vili & Orlov 1971) und der Betrachtung einer Reihe weiterer Befunde (Orlov 1978a) hat sich folgendes herausgestellt: Damit die Gleichung $v_1/v \approx 0.5$ in Texten unterschiedlichen Umfangs gelten kann, muß bei größeren Texten der Anteil an häufigen Wörtern kleiner und entsprechend das Vokabular je Längeneinheit des Textes größer sein als es bei kleineren Texten der Fall ist, wo es mehr relativ häufige Wörter und ein kleineres Vokabular je Längeneinheit geben muß. Folglich ist die Wachstumskurve des Vokabulars in künstlerischen Texten (für wissenschaftliche, technische und philosophische gilt das nicht) eine Funktion ihrer Gesamtlänge, und zwar muß sie desto steiler sein, je größer die Gesamtlänge ist. Mit anderen Worten, der Autor muß schon zu Beginn die lexikalische Repetitionsstrategie entsprechend der zu erwartenden Länge seines noch gar nicht geschriebenen Werkes festlegen (je mehr er zu schreiben vorhat, desto weniger darf er sich wiederholen, und umgekehrt). Man kann sogar den Genauigkeitsgrad einer solchen Übereinstimmung bestimmen.

Orlov (1976, 1978b) hat gezeigt, daß die ideale Beziehung zwischen dem Umfang Z (der Anzahl der Wortverwendungen) eines vollständigen literarischen Werkes und seinem Vokabular v(Z) durch die Formel

$$v(Z) \approx Z/\ln F_{max}$$

auszudrücken ist, wobei F_{max} die absolute Häufigkeit (Wiederholungszahl) des häufigsten Wortes im Text bedeutet. Beim Vergleich mit empirischen Daten zeigt sich, daß bei großen Texten in den weitaus meisten Fällen die relative Abweichung zwischen dem tatsächlichen Vokabular und dem nach der Formel berechneten nicht größer als $\pm$ 20% ist (Beispiele in den beiden letzten Spalten von Tab. 2), wohingegen bei willkürlichen lexikalischen Stichproben die Abweichungen erheblich größer sind (die letzten beiden Spalten von Tab. 1).

Ein ganz analoger Zusammenhang hat sich in der melodischen Organisation von musikalischen Texten aus der Gruppe der Stilrichtungen des 18. bis 20. Jahrhunderts gezeigt. Die elementare

melodische Analyseeinheit war bei diesen Texten das "F-Motiv" - eine Einheit vom Typus eines Motivs (Boroda 1973, 1977)[3]. Wie aus Tabelle 3 hervorgeht, bleiben auch in diesem Fall die relativen Abweichungen in der Hauptsache innerhalb des Intervalls ± 20%. Es ist nicht zu übersehen, daß diese Größe gerade die Unterschiedsschwelle der menschlichen Wahrnehmung physikalischer Größen charakterisiert. Man hat also den Eindruck, daß der Autor eines künstlerischen Textes das Vokabular und die Häufigkeitsstruktur des Textes gerade mit der Genauigkeit kontrolliert, die seiner perzeptuellen "Auflösungskraft" entspricht (zur Kontrolle der Häufigkeitsstruktur ausführlicher Orlov 1970).

Damit aber nun eine solche Kontrolle möglich ist - und sei es nur in Form einer Abschätzung des Gesamtverhältnisses v_1/v der wiederholten und nicht wiederholten Wörter - muß jedes Vorkommen jedes Wortes registriert werden. Ein solcher Prozeß kann natürlich nur außerhalb des Bewußtseins des Autors oder Komponisten ablaufen; denn auf der Bewußtseinsebene sind alle beschriebenen Gesetzmäßigkeiten völlig "unsichtbar".

In diesem Zusammenhang erhebt sich eine Reihe von Fragen. Worin besteht der objektive psychophysiologische Sinn der beobachteten Gesetzmäßigkeiten? Was ist ihr subjektiver Sinn - m.a.W. durch welche Empfindungen (oder Vorstellungen) wird bei einem konkreten Text die Erfüllung (oder Nichterfüllung) dieser Gesetzmäßigkeiten im Bewußtsein vermittelt? Welches ist der Mechanismus der Kontrolle und Überwachung? Welche Rolle spielt schließlich die angezielte künstlerische Form (Drama, Roman, Sonate, Symphonie usw.) bei der Aufprägung einer Wiederholungsstruktur auf den künstlerischen Text?

Orlov (1974) formulierte Hypothesen über die Verbindung der beobachteten Erscheinungen mit Mechanismen der probabilistischen Prognose, insbesondere die Hypothese, daß die Produktion von Kunst eine Art "Trainingstest" für diese Mechanismen sei, der sie maximal belaste. Es wird dort auch erwogen, daß die Erfüllung der beschriebenen Gesetzmäßigkeiten vielleicht mit Empfindungen der "Vollständigkeit", der "Abgeschlossenheit" usw. des künstlerischen Textes zusammenhängt. Boroda (1979) äußerte

Tabelle 3. Analyse von musikalischem Material

Nr.	Material	Z	v	v_1	$\frac{v_1}{v}$	F_{max}	v(Z)	$\frac{v-v(Z)}{v(Z)}$ %
1	Scarlatti, Sonate 1 (Peters)	250	72	37	0,514	28	72,4	- 0,55
2	Haydn, Symphonie 45 ("Abschied")	1304	340	164	0,482	57	317	+ 7,25
3	Beethoven, Rondo op. 51 Nr. 1	624	162	83	0,512	60	150	+ 8,00
4	Chopin, Sonate Nr. 3	2364	504	237	0,470	126	485	+ 3,90
5	Mjaskovskij, Fuge e-Moll	214	60	35	0,573	35	58,5	+ 2,57
6	Kabalevskij, Rondo op. 59	625	171	84	0,491	59	150	+14,00
7	Šostakovič, Präludium und Fuge op. 87,9	677	146	74	0,506	146	135	+ 8,46
8	Levitin, Sonatine für Soloflöte	760	159	76	0,479	111	160	- 0,63
9	Haydn, Symphonie 45, Teil IV (Adagio)	235	93	54	0,580	34	65	+44,60
10	Chopin, Sonate Nr. 2, Teil III	333	39	3	0,077	41	88	-55,60
11	Prokofjev, Sonate für Geigensolo, Teil II	237	136	105	0,771	18	77,5	+75,50
12	Šostakovič, Präludium und Fuge op. 87 Nr. 9, Fragment (Fuge)	528	70	23	0,328	146	105	-33,40
13	Levitin, Sonatine D/Soloflöte, Finale	341	39	7	0,180	111	72	-44,30
14	Bach, Auswahl aus "Wohltemperiertes Klavier"	2852	603	383	0,643	279	506	+21,00
15	Scarlatti, Auswahl aus Klaviersonaten	1263	425	210	0,494	56	308	+38,00

Nr. 1 bis 8 sind vollständige Musikstücke, Nr. 9 bis 13 sind relativ abgeschlossene Bruchstücke aus den Texten, Nr. 14 und 15 sind Auswahlen aus mehreren Texten. Die Daten stammen aus Boroda (1977) und seinen unveröffentlichten Untersuchungen.

die Hypothese, daß die Hörer eines Musikstückes (oder der Leser eines literarischen Werkes), in dem das Zipf-Mandelbrotsche Gesetz erfüllt ist, eine Prognose über die Struktur der Wiederholungen im ganzen Werk aufstellt, nachdem er einen Ausschnitt gehört hat. Aufgrund solcher Voraussagen können "Spiele" des Autors mit dem Hörer entstehen usw.

Eine Sonderstellung hat die Frage, ob die beobachteten Gesetzmäßigkeiten auch zur "Diagnose des Talents" verwendet werden können. Ein plausibler Gedanke ist, daß hochkarätige Künstler diese Gesetzmäßigkeiten in ihren Werken besonders exakt realisieren. Ein Vergleich von Notentexten Chopins, Fields und Czernys (Orlov 1970) scheint diese Vermutung zu stützen, aber die Frage ist sehr kompliziert und jedenfalls von einer endgültigen Lösung noch weit entfernt.

Die Tatsache, daß Kunstwerke eine "Mikroorganisation" haben, die nicht bewußt erfaßt und analysiert wird, verdient die volle Aufmerksamkeit nicht nur von Linguisten und Mathematikern, sondern auch von Psychologen, Untersuchern der künstlerischen Kreativität, Kunstwissenschaftlern und Kybernetikern. Derartige Untersuchungen erlauben es offensichtlich, nicht nur die eigentümliche Organisation des Kunstwerks, die mit den Empfindungen der "Wohlgeformtheit", "Ausgeglichenheit", "Vollständigkeit", "Abgeschlossenheit" des künstlerischen Ganzen zusammenhängen, sondern auch neue, bisher unbekannte Erscheinungen und Gesetzmäßigkeiten der Informationsverarbeitung im menschlichen Gehirn zu entdecken.

ANMERKUNGEN

1. Konkordanz oder Index ist das Wörterbuch des Textes (Menge der Texte u.ä.), in dem die Position (Zeile, Seite) jedes Wortes im Text angegeben ist. Diese Angaben lehnen sich an eine konkrete Edition des Textes. Symphonie ist das Wörterbuch des Textes (Abschnitts usw.), in dem für jedes Wort alle Sätze, in denen das Wort vorkommt, zitiert werden.

2. Die Tabellen 1 und 2 sollen natürlich nur zur Illustration von Schlußfolgerungen dienen, die bei der Analyse eines wesentlich größeren Materials gewonnen wurden. Auch die Gleichung $v_1/v \approx 0{,}5$ ist nur ein Spezialfall eines komplizierteren Typs von Abhängigkeiten, wie sie durch das sogenannte Zipf-Mandelbrotsche Gesetz beschrieben werden und in den vollständigen Textausgaben einzelner literarischer Werke beobachtbar sind (siehe Nadarejšvili & Orlov 1971; Orlov 1978a, b, 1976, 1970, 1974).

3. Ein melodischer "Schnitt" des musikalischen Textes (bei homophonem Text die "Hauptstimme", bei polyphonem Text die Gesamtheit der Stimmen) wurde in F-Motive zerlegt (bei einem polyphonen Text wurde jede Stimme einzeln in F-Motive zerlegt, und dann wurden die erhaltenen Anzahlen zusammengefaßt). Danach wurden die folgenden Textcharakteristika bestimmt: Länge N - das ist die Anzahl aller im Text vorkommenden F-Motive; F-Motiv-Inventar oder (Intonationsbestand); v - das ist die Anzahl unterschiedlicher F-Motive. (Zwei F-Motive gelten als identisch, wenn das eine aus dem anderen durch eine Parallelverschiebung nach der Höhe hervorgeht.) Es wurden auch die Verwendungshäufigkeiten aller F-Motive berechnet und auf diese Weise ein musikalisches Analogon eines Häufigkeitswörterbuches erstellt. Vgl. Kap. 8 in diesem Band.

Nachwort zur deutschen Ausgabe

Die kollektive Monographie, die wir dem westlichen Leser vorgelegt haben, gibt die Ergebnisse unserer mehr als zehnjährigen Bemühungen im Bereich der quantitativen Textanalyse wieder. Obwohl die einzelnen Aufsätze von verschiedenen Autoren zu verschiedenen Zeiten geschrieben wurden, hoffen wir, daß der Leser nicht nur die Einheit der "Ideenplattform" bemerkt, sondern auch die Entfaltung des gemeinsamen Gedankens, die Suche und das Aufspüren des am Anfang unseres Weges nicht ganz klaren Ziels wahrgenommen hat.

Im Nachwort möchte ich eine Art "Wegweiser" durch das Labyrinth der Ideen, in denen wir umherirrten, liefern und die Geschichte dieser Wanderung schildern. Vielleicht ist der "historische Aspekt" etwas persönlich ausgefallen. Aber unsere Arbeit war für uns nie etwas Äußeres: Sie gab uns die Möglichkeit, uns selbst besser kennenzulernen.

* * *

Es fing Ende 1966, Anfang 1967 an. Ich lernte zwei junge Menschen kennen und diese Bekanntschaft veränderte sowohl mein persönliches als auch mein wissenschaftliches Schicksal.

Moisei Boroda war zu jener Zeit Student im sechsten Semester an dem Konservatorium in Tbilisi. Wir trafen uns beim Abhören der Aufnahmen moderner Musik, das vom Verband georgischer Komponisten veranstaltet wurde. Beim Hören von Honegger und Messiaen, Orff und Stockhausen haben wir uns allmählich einander angenähert. Die Gedankengänge des jungen Mannes haben mich in Erstaunen versetzt. In seinen Aussagen über Musik lag etwas, was mich bewegte zu sagen: "Wissen Sie, Mischa, Sie sind ein Mathematiker."

Der junge Mann war sehr überrascht. Sein ganzes Leben lang hat er Geige gelernt. Er hatte absolutes Gehör und seine Zukunft schien für ihn entschieden zu sein. Jedoch ein Jahr nach unserem Treffen legte er die Geige beiseite und wechselte zur musikwissen-

schaftlichen Fakultät des Konservatoriums. Ich habe ihm vorgeschlagen, sich ein Lehrbuch der Mathematik für Hochschulen anzusehen. Er hat sich in das Buch wortwörtlich eingegraben und trennte sich von ihm nicht einmal in der U-Bahn. Vielleicht wäre aus ihm ein guter Mathematiker geworden; wie mir schien, war Mathematik das für ihn von der Natur ursprünglich bestimmte Element. Die Vereinigung einer derartigen Naturgabe mit einer soliden musikalischen Ausbildung brachte ganz einmalige Resultate. Einige von ihnen kann der Leser dieses Bandes selbst beurteilen.

Ungefähr zu derselben Zeit lernte ich Isabella Nadarejšvili kennen, die gerade ihr Studium an der Universität in Tbilisi abgeschlossen hatte. Sie hat ihren Weg auch suchen müssen: Zunächst immatrikulierte sie sich an der Fakultät für westeuropäische Sprachen, später wechselte sie zu der Fakultät für Kybernetik, wo erst vor kurzem die Abteilung für strukturelle und mathematische Linguistik eingerichtet wurde. Nach Abschluß des Studiums begann sie, im Institut für Kybernetik der Akademie der Wissenschaften der GSSR zu arbeiten. Ihr erster Schritt im Institut rief Bestürzung hervor: Sie weigerte sich irgendwelche Striche auf Millimeterpapier zu zählen und erklärte, daß sie nicht bereit sei, mechanische Arbeiten durchzuführen. Die widerspenstige Mitarbeiterin wurde aus einem Laboratorium ins andere geschoben, bis sie in meine Gruppe kam, die sich zu der Zeit mit der Ausarbeitung eines Translators aus ALGOL-60 beschäftigte. Sie wurde bald darauf meine Frau.

Die Arbeit über den Translator endete durch von uns unabhängige Umstände in einer Sackgasse, und ich befand mich in einem Stillstand. Zu dieser Zeit entschied ich mich, die Antwort auf die Frage, die mich seit meiner Studienzeit quälte, zu suchen, nämlich wodurch sich gute Musik von der schlechten objektiv unterscheidet. Kunstwissenschaftliche Arbeiten haben mich nicht ganz befriedigt: In dem mich interessierenden Bereich wurde hier alles auf ein geheimnisvolles "Etwas" zurückgeführt, wodurch sich gute Kunst von der mittelmäßigen und von der schlechten irgendwie unterscheidet. Könnte man vielleicht dieses "Etwas" dennoch quantitativ messen? Wäre es zum Beispiel möglich, daß es bei besseren Autoren mehr un-

terschiedliche Tonkombinationen in gleichlangen Abschnitten des Notentextes gibt?

Dies war schon eine Art Arbeitshypothese. Moisei fing an Notentexte zu kodieren, ich bereitete das Zählprogramm vor und bald konnten wir den Computer mit den Melodien von Chopin (ein Genie!), von Field (ein mittelmäßiger Komponist) und Czerny (ein Pädagoge der Pianotechnik) füttern. Die Arbeitshypothese wurde nicht sehr überzeugend bestätigt: In 11000 Tönen aus verschiedenen Melodien von Chopin ergaben sich 2187 unterschiedliche "musikalische Trigramme", bei Field waren es 1931, bei Czerny 1729. Viel mehr als dieser Unterschied überraschte uns eine andere Erscheinung: In jeder der untersuchten Stichproben wurden erstaunlich genaue Häufigkeitsproportionen von "Trigrammen" gefunden. Die einmal vorkommenden "Trigramme" bildeten ungefähr die Hälfte aller unterschiedlichen "Trigramme", die Zahl der zwei- und dreimaligen "Trigramme" (gesamt) die Hälfte dieser Hälfte, die Zahl derjenigen, die 4 bis 7 mal vorkamen, war noch um die Hälfte kleiner usw. Auf einem Histogramm in doppellogarithmischem Maßstab bildete diese Abnahme eine geradlinige Treppe, die zu den Achsen im 45° Winkel stand (vgl. Anhang zu Orlov 1976, Nr. 3 in diesem Band).

Der professionelle Leser ahnt schon, daß wir im musikalischen Material dieselbe Abhängigkeit gefunden haben, die in der Linguistik den Namen von Zipf, in der Ökonomie den Namen von Pareto, in der Bibliographie den Namen von Bradford, in der Scientometrie den Namen von Lotka usw. trägt. Wir haben davon damals nichts geahnt und sahen die entdeckte Gesetzmäßigkeit mit dem ganzen Enthusiasmus der Erstentdecker an. Etwas vorgreifend muß ich bemerken, daß sich unser damaliger Dilettantismus sogar als nützlich erwies: Unsere Formulierung der entdeckten Gesetzmäßigkeit als die "Funktion der strukturellen Dichte" f(p) (S. 87 f. in diesem Band) erwies sich als geeigneter für die mathematische Analyse als die traditionelle Formulierung bei den oben erwähnten Forschern. Dies hat später das Verständnis für die Situation stark erleichtert. Hätten wir von Anfang an die Literatur über "Zipfsche" Gesetzmäßigkeiten gelesen, so wären wir von den traditionellen Formulierungen, mit deren Hilfe man keine guten Arbeitsformeln erhal-

ten kann, beeinflußt worden. Etwas anderes war wahrscheinlich noch wertvoller: Da wir diese Gesetzmäßigkeit selbstständig entdeckt hatten, sind wir mit ihr auch später so frei umgegangen als ob sie unser Eigentum wäre, ohne Rücksicht auf Autoritäten - wir haben einfach angefangen darüber nachzudenken, was das alles bedeutet.

In dieser Zeit hat es uns das Schicksal gegönnt, über die entdeckten Abhängigkeiten mit dem Naturwissenschaftler Dr. Sc. A.L. Brudno zu sprechen. Er sagte:

„Wissen Sie, in der Musik kenne ich mich nicht sehr gut aus. Aber Sie, Georgier" - er wendete sich an meine Frau - „haben doch das Poem 'Held im Tigerfell'. Schauen Sie, ob es vielleicht dort etwas ähnliches gibt."

Das war ein sehr wertvoller Rat, denn aus diesem Grunde hat sich die Linguistin Isabella Nadarejšvili der ursprünglich musikologischen Arbeit angeschlossen. Sie erstellte das Häufigkeitswörterbuch des Poems von Rustaveli. Als wir ein Histogramm desselben Typs wie früher in der Musik erhalten haben, waren wir konsterniert - im doppellogarithmischen Maßstab ergab sich dieselbe Treppe mit 45° Neigung zu den Achsen!

Zuerst dachten wir (ebenso wie unsere Vorgänger), daß wir es mit einer allgemeinen sprachlichen Gesetzmäßigkeit zu tun hätten. Aber dann häuften sich Tatsachen, die mit dieser Ansicht nicht übereinstimmten: Die Treppe wurde ideal geradlinig erst, wenn wir Häufigkeitsangaben aus dem gesamten Text eines großen literarischen Werkes in Betracht zogen. Wenn wir nur Teile von Texten oder viele Texte vereinigt in eine Stichprobe nahmen, so änderte sich das Bild, die Treppe krümmte sich: Im ersten Fall auf eine Seite, im zweiten Fall auf die andere. Es wurde klar, daß die entdeckten Abhängigkeiten sich nicht auf die Sprache als ganze, sondern auf den Einzeltext beziehen, und zwar auf den hochorganisierten, der aktiven menschlichen Wahrnehmung angepaßten, auf die Mechanismen dieser Wahrnehmung eingestimmten Text. Diese vorläufigen Beobachtungen wurden in zwei Aufsätzen erläutert (Nadarejšvili, Orlov 1969, Nr. 10 in diesem Band; Orlov 1970a, hier nicht vorhanden; das Material dieser Arbeit wurde vollständig in Orlov 1978a, Nr. 4 in diesem Band, übernommen).

Die angehäuften Tatsachen brauchten dringend nicht nur "inhaltliche Hypothesen", sondern auch eine seriöse theoretische Analyse. Die vorhandene Literatur über das Zipfsche Gesetz (unser Dilettantismus war zu der Zeit natürlich im Schwinden) hat uns wenig gegeben. Das Problem der Deformation der Häufigkeitsgraphiken, der Abweichungen vom Zipfschen Gesetz wurde hier überhaupt nicht näher besprochen; es gab nur diverse empirische Korrekturen zu den Formeln von Zipf und Mandelbrot.

Es zeigte sich, daß das Zipf-Mandelbrotsche Gesetz eine merkwürdige Eigenschaft hat: Wenn es für eine Stichprobe mit einem bestimmten Umfang gilt (dieser Umfang wurde in diesem Band als Z - nach Zipf - bezeichnet), dann gilt es nicht mehr für Stichproben mit einem beliebigen anderen Umfang (es wird angenommen, daß der Text statistisch homogen ist), was notwendigerweise zu der Entstehung einer Erscheinung führt, die man im professionellen Jargon als Abweichung (Krümmung) des "Schweifes" bezeichnet. Es ist gelungen, diese Abweichungen theoretisch zu erfassen und die Theorie wurde experimentell befriedigend bestätigt (vgl. Orlov 1976, 1978a, Nr. 3,4 in diesem Band).

Gleichzeitig hat sich die etwas unerwartete theoretische Folgerung aus den erfahrenen Beobachtungen über die Erfüllung des Zipfschen Gesetzes (in unserer Formulierung) in vollständigen literarischen Texten geklärt. Aus den erhaltenen Formeln folgte, daß in gleichgroßen Abschnitten aus Texten unterschiedlicher Länge derjenige Abschnitt den größeren Wortschatz hat, der aus einem längeren Text stammt. Mit anderen Worten, je länger ein literarisches Werk, desto schneller wächst in ihm der Wortschatz, desto größer ist in ihm der relative (nicht nur der absolute) Vokabularreichtum.

Der bekannte Mathematiker A.N. Kolmogorov verfolgte aufmerksam unsere Arbeit und schlug vor, die Bestätigung dieses theoretisch vorausgesagten Effekts zuerst in verschiedenen Werken eines Autors zu suchen. Denn bis jetzt wurde als selbstverständlich angenommen, daß jeder Autor seinen eigenen, individuellen relativen Vokabularreichtum hat, und in gleichgroßen Stichproben aus seinen Texten ungefähr gleichgroße Wortschätze vorkommen sollten (diese

Annahme wird bei der Entscheidung über die strittige Autorschaft verwendet). Wenn es sich also zeigen würde, daß in unterschiedlich langen Werken eines Autors unterschiedliches Vokabularzuwachstempo besteht (mit anderen Worten, wenn der Autor dieses Tempo in Abhängigkeit von der künftigen Länge des gerade geschriebenen Werkes ändert), dann hätten wir ein starkes Argument sowohl zugunsten unserer Interpretation des "Zipfschen Gesetzes" als auch zugunsten der Hypothese, daß sich dieses Gesetz am vollständigen Text des literarischen Werkes infolge der zielgerichteten (obwohl ganz unterbewußten) Bemühungen des Autors realisiert.

Zur Überprüfung haben wir vier Texte von Leo Tolstoj gewählt, "Kreutzersonate", "Die Kosaken", "Krieg und Frieden" und "Die Auferstehung". Die ersten zwei Texte bestätigen eindeutig unsere Annahme. Die Kurve des Vokabularzuwachses in "Die Kosaken" (dieser Text ist doppelt so lang wie die "Kreutzersonate") stieg viel steiler als in der "Kreutzersonate", wobei der Unterschied zwischen den Werten der beobachteten Wortschätze mit der theoretischen Voraussage übereinstimmte. Obwohl in "Krieg und Frieden" und in "Die Auferstehung" die theoretische Prognose des Wortschatzes sehr hoch war, ergab sich ein beobachteter Wortschatz von derselben Größenordnung wie in den beiden ersten Texten. Es zeigte sich, daß sich das Zipf-Mandelbrotsche Gesetz in diesen Werken nicht bei den vollen Textlängen, sondern bei Teilen, die der Verfasser selbst kenntlich gemacht hat, realisiert (17 Teile in "Krieg und Frieden" und 3 Teile in "Die Auferstehung"). Dabei stimmt die theoretische Kurve des Vokabularwachsens mit den theoretischen Kurven für mittlere Umfänge dieser Teile überein. Und da die Länge eines Teiles in "Die Auferstehung" doppelt so groß ist wie in "Krieg und Frieden", ist das Vokabular von "Die Auferstehung" reicher!

Dieses Resultat haben wir in Nadarejšvili, Orlov (1971) veröffentlicht. Dieser Aufsatz ist hier nicht vorhanden, weil sein Material in Orlov (1978a) übernommen wurde (vgl. Nr. 4 dieses Bandes).

Man hätte die Abhängigkeit des relativen Vokabularreichtums von dem vollen Textumfang wahrscheinlich längst ohne uns entdeckt, wenn sie von einer ganzen Reihe von begleitenden Erschei-

nungen nicht verdeckt wäre: Erstens, die oben erwähnte Realisierung des Zipf-Mandelbrotschen Gesetzes in einzelnen Teilen von großen, "zyklischen" Werken; zweitens, die Nichterfüllung des Gesetzes in kurzen zeitgenössischen Texten (Gedichte, Novellen); drittens, der Zuwachs des relativen Vokabularreichtums mit der Zeit (ausführlicher vgl. Nr. 4 und Nr. 1 in diesem Band).

Während ich mich mit der theoretischen Analyse beschäftigt habe, überprüfte I. Nadarejšvili ihre Konsequenzen an literarischen Texten. M. Boroda fing in den 70er Jahren mit der quantitativen Analyse musikalischer Texte an. Obwohl wir schon in der ersten Arbeit die Erfüllung des Zipf-Mandelbrotschen Gesetzes bei melodischen Intervallen und Intervallgruppen beobachtet haben, hatten wir keine Lust in dieser Richtung weiterzuarbeiten. Die Zerlegung der Melodie in unsere "Verbindungen" klang für das Gehör sehr unnatürlich. Auf der anderen Seite weiß jeder, der sich mit Musik beschäftigt, daß es in ihr keine Einheiten wie Silbe, Wort, Satz gibt. An die traditionellen musikwissenschaftlichen Begriffe wie Motiv, Teilmotiv, Phrase kann man sich aber nicht anlehnen, weil sie sehr vage sind, ihre Segmentierung ist subjektiv, und daher können unterschiedliche Forscher bei ihrer Zählung unterschiedliche Zahlen bekommen. Das heißt, bevor man mit einer quantitativen Musikanalyse beginnt, muß man die Vorstellung von der Mikrostruktur des musikalischen Textes präzisieren, um eine für das Gehör natürliche Zähleinheit eindeutig segmentieren zu können.

M. Boroda hat dieses einzigartige Problem erfolgreich gelöst. Wie man heute feststellen kann, ist die von ihm definierte melodische Einheit "Formalmotiv" (F-Motiv) nicht nur für eine statistische Musikanalyse geeignet. Die Zerlegung der Melodie in F-Motive wird nicht als etwas der Melodie künstlich aufgezwungenes verspürt; im Gegenteil, diese Zerlegung erläutert die logische und die rhythmische Struktur der Melodie (vgl. die Beispiele in Borodas Arbeiten in diesem Band). Die Häufigkeiten der F-Motive in musikalischen Texten liegen genauso gut auf den Zipf-Mandelbrotschen Kurven, wie die Worthäufigkeiten in literarischen Texten. Und genauso ergaben sich auch Abweichungen bei Abschnitten

oder Teilen musikalischer Werke; und genauso ergab sich in längeren Kompositionen eine steilere Zuwachskurve des "musikalischen Vokabulars" als in kürzeren (vollständigen) Kompositionen (Boroda 1977, 1979).

Die Zunahme neuer Elemente im Text und die allgemeinen Proportionen von häufigen und seltenen Elementen sind also von den allgemeinsten Eigenschaften künstlerischer Kompositionen abhängig: Von der Länge (in Anzahl der Zähleinheiten) und von der Häufigkeit des häufigsten Elements. Diese Abhängigkeit kann nicht durch irgendwelche allgemeine statistische Überlegungen geklärt werden; sie bezeugt den individuellen Zugang des Autors zur Konstruktion des künstlerischen Textes. Mit anderen Worten, es muß eine dem Autor nicht bewußte Kontrolle der Häufigkeitsstruktur geben; grob gesagt, der Autor muß eine Prozedur durchführen, die der Erstellung eines (laufenden!) Häufigkeitswörterbuchs des sich gerade im Entstehen befindenden Textes und der Überprüfung der Übereinstimmung der Häufigkeitsgraphik mit dem Zipf-Mandelbrotschen Gesetz äquivalent ist. Eine mögliche Variante dieser Prozedur (in Form einer Kontrolle der differentiellen Geschwindigkeit des Vokabularzuwachses) findet man in Orlov (1978a, Nr. 4).

In dieser Richtung sind wir nicht weiter vorwärtsgekommen (obwohl sie uns am vielverheißendsten scheint). Ebenso blieb die Frage des objektiven Messens des Grades der künstlerischen Qualität eines Textes ungelöst, obwohl sie uns bei unseren ersten Schritten begeistert hat. Die Untersuchung nichtkünstlerischer Texte (wissenschaftliche, technische, philosophische) zeigte, daß diese im Schnitt dem Zipf-Mandelbrotschen Gesetz viel schlechter folgen als künstlerische (Orlov 1978a). Kann es also sein, daß irgendwelche ästhetischen Empfindungen von dem Grad der Präzision abhängen, mit dem im Text die vom Zipf-Mandelbrotschen Gesetz diktierten Proportionen erscheinen? Oder gibt es einen Bereich der zugelassenen Abweichungen von diesem Gesetz, innerhalb dessen die Genauigkeitsverfeinerungen irrelevant sind? Um dieses Problem zu lösen, müßte man die "reine" Statistik einer immensen Menge von Texten von verschiedenen Genres und künstlerischen Qualitäten ermitteln, die Methode der Qualitätsschätzung durch Experten anwenden usw. usw. Es übersteigt einfach unsere Kräfte.

Außer künstlerischen Texten gibt es aber noch verschiedene andere: "Fachsprachen", "Informationsflüsse" usw., die mit Rangverteilungen des Zipfschen Typs zusammenhängen. Es zeigte sich, daß die Klasse der durch theoretische Analyse erhaltenen Kurven, die vom Zipf-Mandelbrotschen Gesetz abweichen (man könnte sie als "quasizipfsche" bezeichnen), allgemein genug diversen "Flüssen", "Teilsprachen" usw. entspricht (Orlov 1978a,b; 1980). Das aufgestellte Modell war imstande sogar biologische Stichproben zu beschreiben und vorauszusagen (vgl. Orlov, Ju.K., Statističeskoe modelirovanie sootnošenij častot vidov v ekologičeskich vyborkach (linguistiko - ekologičeskie paralleli) / Statistisches Modellieren von Häufigkeitsverhältnissen der Spezies in ökologischen Stichproben (linguistisch-ökologische Parallelen)/. In: Pesenko, Ju.A. (Hrsg.), Količestvennye metody v ekologii životnych. Leningrad, Izdatel'stvo Zoologičeskogo instituta AN SSSR 1980, 99-101). In diesem Falle hat aber der "Zipfsche Umfang" keine inhaltliche Interpretation und die Zipf-Mandelbrotschen Kurven werden zu einem Spezialfall einer breiteren Klasse der quasizipfschen Strukturen.

Eine tiefere theoretische Analyse der Situation führte zu dem Schluß, daß wenn in der Stichprobe viele einmal vorkommende Elemente vorhanden sind (was in lexikalischen Stichproben immer der Fall ist), dann schöpft die Stichprobe noch von weitem das ganze Inventar (Vokabular) der Grundgesamtheit nicht aus. Unter solchen Beobachtungsbedingungen sind die Stichprobendaten in Bezug auf die Struktur der Grundgesamtheit stark verschoben. Mit anderen Worten, wenn die Stichproben relativ klein sind (obwohl sie absolut genommen sehr groß sein können), dann sind die bei ihnen beobachteten "Rangverteilungen" (des Zipfschen Typs) eine Art Trugbilder, Fiktionen, die mit den Beobachtungsbedingungen zusammenhängen (vgl. ausführlicher Orlov, Nr. 1 in diesem Band). In solchen Situationen entstehen spezifische Probleme der Schätzung der Verteilung der Grundgesamtheit aufgrund der Stichprobenkenngrößen. Man kann beispielsweise den Versuch, das "potentielle Vokabular" von A.S. Puškin aufgrund der Häufigkeitsdaten aus seinem Gesamtwerk zu schätzen, zeigen. In seinem Gesamtwerk gibt es

etwa 21000 unterschiedliche Wörter; die Schätzung ergibt etwa 60000 Wörter, die in einem bestimmten Sinne im Kopf des Klassikers der russischen Literatur "bereit zur Verwendung" vorhanden waren (Orlov, Ju.K., Čitasvili, R.Ja., Dvuchparametričeskaja model' častotnoj struktury leksiki / Ein zweiparametrisches Modell der Häufigkeitsstruktur des Lexikons/. In: Škola - seminar po prikladnoj i inženernoj lingvistike 3-14, ijulja 1978g (tezisy dokladov i soobščenij). Machačkala 1978: 30-31). R.Ja. Čitašvili bereitet mit mir zur Zeit eine Reihe von Publikationen über mathematische Probleme einer derartigen Schätzung vor.

Bis zum heutigen Tag sieht die Situation folgendermaßen aus: Es gibt eine einheitliche Verteilung, die die beobachtete Vielfalt von Häufigkeitsstrukturen ("Rangverteilungen") erzeugt. Diese Verteilung entsteht sozusagen von alleine als Ergebnis des Funktionierens solcher komplexer Systeme wie Sprache (genauer, Systeme der Redekommunikationen), Biotopen usw. Wenn wir aber einen konkreten künstlerischen Text betrachten, so zeigt es sich, daß sich unter allen möglichen "Trugbildformen" der Häufigkeitsgraphiken gerade die "kanonische", Zipf-Mandelbrotsche Form realisiert. In willkürlichen Stichproben findet man diese Form nur selten.

Die Existenz solcher scheinbar unterschiedlichen, jedoch zweieinheitlichen Gesetze für Sprache und Text zwingt notgedrungen zu besonderer Vorsicht bei den methodologischen Problemen der Linguostatistik. Es ist insbesondere unzulässig aus mehreren Texten eine gemischte Stichprobe zu erheben, ohne die Daten aus jedem Einzeltext zu fixieren, da die Zahlen sonst keinen bestimmten Sinn ergeben. Diesen methodologischen Problemen widmet sich die Arbeit von Nadarejšvili, Orlov (1978, Nr. 2) und der erste Aufsatz (Nr. 1), der speziell für diesen Band geschrieben wurde.

Es ist zu bemerken, daß alle theoretischen Konstrukte nur für statistisch homogene Texte zuverlässig gelten, d.h. für solche, die man als eine festgelegte Folge zufälliger Ziehungen aus einer lexikalischen Grundgesamtheit betrachten kann. Reale zusammenhängende Texte und lexikalische Stichproben unterscheiden sich natürlich von dieser Idealisierung. Aber wie und in welchem Ausmaß?

Der Klärung dieses Problems sind die Arbeiten von Boroda, Nadarejšvili, Orlov, Čitašvili (1977, Nr. 12 in diesem Band) und Nadarejšvili, Nadarejšvili, Orlov (1977, Nr. 13 in diesem Band) gewidmet. Es zeigte sich, daß sogar seltene Wörter eine Tendenz zur Häufung in relativ kurzen Textabschnitten aufweisen; A.N. Kolmogorov schlug uns vor, diese Erscheinung als "Ballung" zu bezeichnen. Diese Erscheinung verursacht in Anfangsteilen des Textes ein etwas verlangsamtes Anwachsen des Vokabulars im Vergleich mit dem Anwachsen in einem statistisch homogenen Text; danach beschleunigt sich aber der Vokabularzuwachs. Im großen und ganzen läuft die Kurve des Vokaburlarzuwachses im zusammenhängenden Text fast immer etwas tiefer als die in einem ideal homogenen Text (s. Anhang zu Nadarejšvili, Orlov 1978, Nr. 2 hier); auch andere Forscher haben diese Erscheinung beobachtet.

Ein anderer Unterschied zwischen zusammenhängenden hochorganisierten Texten und zufälligen Elementenmengen besteht (vom statistischen Gesichtspunkt) in einer Art "Neigung" der Elemente unterschiedlicher Häufigkeit in bestimmten Positionen im Text aufzutreten. Wenn der Text deutlich strukturiert ist, wie z.B. der poetische Text, dann konzentrieren sich die seltenen Wörter am Versende; Wörter mit großer Häufigkeit "meiden" diese Position (Nadarejšvili, Orlov 1969, Nr. 10 in diesem Band). Eine ähnliche Erscheinung wurde auch im musikalischen Material beobachtet (Boroda, Orlov 1970).

Zum Schluß möchte ich kurz drei Arbeiten erwähnen, die auf den ersten Blick etwas abseits von dem grundlegenden "Ideengang" dieses Sammelbandes stehen. Es ist vor allen Dingen M. Borodas Arbeit "Zur Bestimmung einer phrasenähnlichen melodischen Informationseinheit in der Musik" (Nr. 7 in diesem Band). Zusammen mit dem F-Motiv gründet die R-Phrase (rhythmische Phrase) die Familie der formal definierten musikalischen Einheiten, die für quantitative Analysen hinreichend natürlich und effektiv sind. Sie ist bedeutend länger und "bedeutungsvoller" als das F-Motiv und obwohl der Begriff des F-Motivs in ihre Definition nicht eingeht, so besteht sie immer, wie Boroda zeigte, aus einer ganzen Anzahl von F-Motiven. Dies ist das <u>erste</u> streng bewiesene <u>Theorem</u> in der Musikwissenschaft.

Die Aufstellung eines hierarchischen Systems musikalischer Einheiten - Boroda beschäftigt sich zur Zeit eben mit dieser Aufgabe - scheint mir für die Untersuchung der Musik äußerst wichtig. Ich bin überzeugt, daß erst nach der Erstellung eines solchen Systems und nach der detaillierten Analyse der Musik mit Hilfe einzelner Einheiten und ihrer gegenseitigen Zusammenhänge es möglich sein wird, die Arbeiten über maschinelle Musikkompositionen aus dem Zustand der "Computeralchymie", in dem sie sich bis jetzt befinden, hinauszuführen. Auch die traditionelle Musikwissenschaft gewinnt viel, wenn sie formalisierte Begriffe verwendet. Schon jetzt gibt es eine Reihe sehr interessanter Resultate über den Zusammenhang des F-Motivs mit dem Wort in Vokaltexten, über "physikalische" Eigenschaften des F-Motivs (seine mittlere Länge in realen musikalischen Werken ist praktisch gleich der mittleren Länge des Wortes in zusammenhängender Rede) usw. Die Untersuchungen auf der Ebene der R-Phrasen haben erst begonnen, aber es ist bereits gelungen, interessante Resultate über den Unterschied zwischen den Eigenschaften der R-Phrasen in Vokal- und Instrumentalmusik zu bekommen. Ich kann hier leider nicht ausführlicher berichten, aber ich hoffe, daß in der Reihe "Quantitative Linguistics" ein Band über "Quantitative Musikanalyse (sowjetische Beiträge)" herausgegeben von M. Boroda erscheinen wird, der diese Arbeiten enthalten wird.

I. Nadarejšvili konzentrierte sich in den letzten Jahren auf die statistische Analyse georgischer Texte. In einer Arbeit (1970) entdeckte sie die Wandlungen der Häufigkeit der Konjunktion "und" in georgischen Texten je nach Genre und Zeit. In der älteren Prosa (bis zum XVIII-XIX Jh.) war ihre Häufigkeit sehr groß, 0.08-0.14; in der gereimten Poesie derselben Periode sehr niedrig, etwa 0.02; denselben Unterschied findet man auch im Folklore. In der heutigen Zeit hat sich die Häufigkeit auf 0.04-0.06 stabilisiert und ist typisch für moderne Texte auch in anderen Sprachen, wobei es keinen Unterschied zwischen Prosa und Poesie gibt. Aus technischen Gründen konnten wir diese Arbeit hier leider nicht bringen.

In Nadarejšvili (1978, Nr. 5 in diesem Band) findet man eine ausführliche Analyse von Ähnlichkeiten und Unterschieden (stati-

stisch gesehen) in den Texten des georgischen Schriftstellers Konstantin Gamsachurdia. I. Nadarejšvilis Arbeiten stehen an der Grenze zwischen der "traditionellen" quantitativen Linguistik und der Disziplin, die man als "vergleichende statistische Literaturwissenschaft" bezeichnen kann. Hier möchte ich mein Bedauern darüber aussprechen, daß quantitative Methoden bisher wenig Interesse derjenigen Forscher erweckt haben, die sich nicht mit "Sprache" und "Rede", sondern mit Texten als solchen beschäftigen. Es ist eine merkwürdige Ironie des Schicksals, daß gerade diejenigen Texte, über die ein beträchtliches Zahlenmaterial zur Verfügung steht, der komparativen Analyse völlig entgangen sind. Eine der Ursachen besteht darin, daß man Texte unterschiedlicher Länge wegen der "Verschiebung I. Art" (vgl. Orlov Nr. 1 in diesem Band) schwer miteinander vergleichen kann. Die Neutralisierung dieser Erscheinung war bisher nur durch Ausgleich der Stichprobenumfänge (oder durch Wahl der Texte gleicher Länge) möglich, was die Möglichkeiten der Untersuchung stark beschränkte. Wir hoffen, daß die hier erläuterten Methoden, die es erlauben, Kenngrößen von Texten unterschiedlicher Länge zu vergleichen, der Entwicklung der Forschung in diesem Bereich behilflich sein werden.

Und schließlich noch ein paar Worte zu der Untersuchung über die Verteilung der Farbflächen in der Malerei (Vološin, Orlov 1972, Nr. 9 in diesem Band). Dies war die Dissertation des (damaligen) Studenten der Fakultät für Kybernetik Boris Vološin. Ich habe ihm dieses Thema zu einer Zeit vorgeschlagen, als wir noch über das Zipfsche Gesetz relativ wenig wußten. Die kurze Frist und die unvollkommenen Methoden, die Vološin zur Verfügung standen, veranlaßten uns, die Resultate als vorläufig zu betrachten; sie waren aber so interessant, daß wir uns entschlossen, diese bisher nicht erschienene Arbeit in diesen Band aufzunehmen. Vološin konnte leider seine Untersuchung nicht fortsetzen; wir wären sehr froh, wenn diese Arbeit das Interesse für dieses bisher unerforschte Problem erwecken würde.

Ich bedanke mich bei meinen Freunden und Mitarbeitern, Mitgliedern unserer kleinen Gruppe, für das Glück des gemeinsamen wissenschaftlichen Schaffens. Ich bin Dr. A.N. Kolmogorov und Dr. A.L. Brudno sehr für ihr Interesse, ihre Kritik und Unterstützung unserer Forschung verpflichtet. Ein besonderer Dank gehört Dr. G. Altmann, der uns vorschlug, unsere Arbeiten in dieser Reihe zu veröffentlichen. Wir sind alle sehr gerührt von seinem Vorwort, in dem er unsere Arbeiten so hoch einschätzt. Nur in einem Punkt möchten wir mit ihm polemisieren. Er schreibt über das Ende der "romantischen Epoche" individueller Forschung, aber wir sind überzeugt, daß Romantik auch bei der kollektiven Forschung möglich ist. Denn, gerade in solchen Mikroteams entsteht manchmal die besondere Atmosphäre der erhöhten Kreativität und der gegenseitigen "intellektuellen Stimulanz", die es erlaubt, Fragen zu stellen und zu lösen, die einen einzelnen Verstand überfordern.

Tbilisi, Institut für Kybernetik
der Akademie der Wissenschaften
der Georgischen SSR

Jurij K. Orlov

März 1981

LITERATUR

ABULADZE, L.B., Über eine statistische Wortschatzanalyse der künstlerischen Prosa von Vāža Pšavela (in georgischer Sprache). In: Šaradzenidze, T.S. (Hrsg.), Voprosy sovremennogo obščego i matematičeskogo jazykoznanija. Tbilisi, Mecniereba 1966, 94-148

A Concordance to Byron's Don Juan. Ithaca, Cornell University Press 1967

ALEKSEEV, P.M., Statističeskaja leksikografija v anglistike [Die statistische Lexikographie in der Anglistik]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki, Moskva, MGPIIJA 1969, 3-7

ALEKSEEV, P.M., Častotnye slovari anglijskogo jazyka i ich praktičeskoe primenenie [Häufigkeitswörterbücher des Englischen und ihre praktische Anwendung]. In: Piotrowski, R.G. (Hrsg.) Statistika reči i avtomatičeskij analiz teksta. Leningrad, Nauka 1971, 160-178

ALEKSEEV, P.M., Statičeskaja leksikografija (tipologija, sostavlenie i primenenie častotnych slovarej) [Statistische Lexikographie (Typologie, Aufbau und Anwendung von Häufigkeitswörterbüchern)]. Leningrad, LGPI 1975

ALEKSEEV, P.M., Kvantitativnaja tipologija teksta [Quantitative Texttypologie]. Leningrad, Diss. 1977

ALINEI, M. (ed.) Spogli ellettronici dell' Italiano delle origini e del duecento. Il Forme 5: Dante Alighiere. La Commedia. Bologna, Il Mulino 1971

ALINEI, M. (ed.), Spogli ellettronici dell' Italiano delle origini e del duecento, Il Forme 8: Dante Alighiere. „La Vita Nuova". Bologna, Il Mulino 1971

ANDREEV, N.D., Raspredelitel'nyj slovar' i semantičeskie polja [Verteilungswörterbuch und semantische Felder]. In: Andreev, N.D. (Hrsg.), Statistiko-kombinatornoe modelirovanie jazykov. Leningrad, Nauka 1965, 490-496

APOLLINAIRE, G., Calligrammes, Concordances, Index verborum et Relèves statistiques. Paris, Larousse s.d.

ARAPOV, M.V., EFIMOVA, E.N., ŠREJDER, Ju.A., O smysle rangovych raspredelenij [Über den Sinn von Rangverteilungen]. Naučno-techničeskaja informacija Ser. 2, 1975, (Nr. 1), 9-20

BEKTAEV, K.B., Častotnyj slovar' jazyka Abaja [Häufigkeitswörterbuch der Sprache von Abaj]. In: Mežvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacija lingvostatičeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 36-37

BEKTAEV, K.B., BELOCERKOVSKAJA, L.I., PIOTROWSKI, R.G., Norma - situacija - tekst i lingvostatističeskie priemy issledovanija [Norm - Situation - Text und die sprach-statistischen Untersuchungsverfahren]. In: Piotrowski, R.G., Graudina, L.K., Ickovič, V.A. (Hrsg.), Jazykovaja norma i statistika. Moskva, Nauka 1977, 5-42

BEKTAEV, K.B., LUK'JANENKOV, K.F., O zakonach raspredelenija edinic pis'mennoj reči [Über die Verteilungsgesetzmäßigkeiten der Einheiten der geschriebenen Sprache]. In: Piotrowski, R.G. (Hrsg.), Statiskika reči i avtomatičeskij analiz teksta. Leningrad, Nauka 1971, 47-112

BOL'ŠEV, L.N., SMIRNOV, N.V., Tablicy matematičeskoj statistiki [Tabellen der mathematischen Statistik]. Moskva, Nauka 1965

BOOTH, A.D., A "law" of occurrence for words of low frequency. Information and Control 10, 1967, 409-418

BORODA, M.G., K voprosu o metroritmičeski elementarnoj edinice v muzyke [Zur Frage der metrorhythmisch elementaren Einheit in der Musik]. Soobščenija AN GSSR 71, 1973 (Nr. 3), 745-748

BORODA, M.G., O častotnoj strukture muzykal'nych soobščenij [Über die Häufigkeitsstruktur musikalischer Nachrichten]. Soobščenija AN GSSR 76, 1974 (Nr. 2), 178-202

*BORODA, M.G., O melodičeskoj elementarnoj edinice [Zur melodischen Elementareinheit]. In: Gošovskij, V.L. (Hrsg.), Materialy pervogo vsesojuznogo seminara po mašinnym aspektam algoritmičeskogo analiza muzykal'nych tekstov. Erevan, Izdatel'stvo AN Arm. SSR 1975, 112-120

*BORODA, M.G., Častotnye struktury muzykal'nych tekstov [Häufigkeitsstrukturen musikalischer Texte]. In: Saversašvili, A.V. (Hrsg.), Sbornik statej Tbiliskoj gosudarstvennoj konservatorii. Tbilisi, Mecniereba 1977, 178-202

BORODA, M.G., Principy organizacii povtorov na mikrourovne muzykal'nogo teksta [Organisationsprinzipien der Wiederholungen auf der Mikroebene eines musikalischen Textes]. Habilitationsschrift, Tbilisi 1979

*BORODA, M.G., NADAREJŠVILI, I.Š., ORLOV, Ju.K., ČITAŠVILI, R.Ja., O charaktere raspredelenija informacionnych edinic maloj častoty v chudožestvennych tekstach [Über den Charakter der Verteilung von Informationseinheiten geringer Häufigkeiten in künstlerischen Texten]. Semiotika i informatika 9, 1977, 23-24

*BORODA, M.G., ORLOV, Ju.K., O nekotorych statističeskich osobennostjach muzykal'nych soobščenij [Über einige statistische Besonderheiten musikalischer Nachrichten]. Soobščenija AN GSSR 57, 1970 (Nr. 2), 301-303

*BORODA, M.G., ORLOV, Ju.K., O nekotorych psichologičeskich aspektach količestvennoj organizacii chudožestvennych tekstov [Über einige psychologische Aspekte der quantitativen Organisation künstlerischer Texte]. In: Prangišvili, A.S., Serozia, A.E., Bassin, F.V. (Hrsg.), Bessoznatel'noe 3. Tbilisi, Mecniereba 1978, 302-309

BORODIN, V.V., MATER, E.A., Sostavlenie slovnikov na EVM [Die Erstellung von Wörterbüchern mit Hilfe der EDV]. In: Avtomatizacija informacionnych rabot i voprosy matematičeskoj lingvistiki. Seminar. Vypusk I. Kiev 1967, 54-71

BUCKOJ, A., Struktura muzykal'nych proizvedenij [Die Struktur musikalischer Werke]. Leningrad-Moskva, Gosudarstvennoe muzykal'noe izdatel'stvo 1948

BUDMAN, M.M., Statistika anglijskich tekstov po avtomobilestroeniju [Die Statistik englischer Texte über den Automobilbau]. In: Piotrowski, R.G. (Hrsg.), Avtomatičeskaja pererabotka teksta metodami prikladnoj lingvistiki. Kišinev, Politechničeskij Institut 1971, 274-277

BUKOVIČ, V.A., Častotnyj slovar' anglijskogo pod'jazyka elektronnovyčislitel'noj techniki [Häufigkeitswörterbuch der englischen Fachsprache der Computertechnik]. In: Piotrowski, R.G. (Hrsg.), Statistika teksta, Bd. 1, Minsk Izdatel'stvo BGU 1969, 414-426

BULACHOV, M.G., Materialy dlja častotnogo slovarja russkogo jazyka (Pisarev, "Realisty"). In: Leksikologija i gramatyka, Minsk 1969

CARROLL, J.B., Word-frequency studies and the lognormal distribution. In: Zale, E.M. (Hrsg.), Proceedings of the conference on language and language behavior. New York, Appleton-Century-Crofts 1968, 213-235

CARROLL, J.B., A rationale for an asymptotic lognormal form of word-frequency distribution. Princeton, Princeton University Press 1969

DARČUK, N.P., Individual'noe i obščee v leksičeskoj sisteme avtorskogo stilja (na materiale sovremennoj ukrainskoj chudožestvennoj prozy) [Das Individuelle und das Allgemeine im lexikalischen System des Autorenstils]. Habilitationsschrift, Kiev 1975

Dictionaire des fréquences . Vocabulaire litteraire des XIX et XXe siècles. Paris, Didier 1971

FINKENSTAEDT, Th., WOLF, D., Statistische Untersuchungen des englischen Wortschatzes mit Hilfe eines Computers. Beiträge zur Linguistik und Informationsverarbeitung 16, 1969, 7-34

FRUMKINA, R.M., Statističeskie metody izučenija leksiki [Statistische Methoden der Untersuchung der Lexik]. Moskva, Nauka 1964

FRUMKINA, R.M., K voprosu o tak nazyvaemom zakone Cipfa [Zur Frage des sogenannten Zipfschen Gesetzes]. Voprosy jazykoznanija 1961, Nr. 2, 117-122

FRUMKINA, R.M. (Hrsg.), Materialy k častotnomu slovarju jazyka Puškina (prospekt) [Materialien zum Häufigkeitswörterbuch der Sprache Puškins (Prospekt)]. Moskva, Institut jazykoznanija AN SSSR 1963

FUKS, V., Matematičeskaja teorija slovoobrazovanija [Fucks, W., Mathematical theory of word formation]. In: Siforov, V.I. (Hrsg.), Teorija peredači soobščenij. Moskva, Izdatel'stvo inostrannoj literatury 1957, 221-247

GAČEČILADZE, T.G., CILOSANI, T.V., Ob odnom metode izučenija statističeskoj struktury teksta [Über eine Methode zur Untersuchung der statistischen Struktur des Textes]. In: Piotrowski, R.G. (Hrsg.), Statistika reči i avtomaticeskij analiz teksta. Leningrad, Nauka 1971, 113-133

GAČEČILADZE, T.G., ELIAŠVILI, A.I., Statistika bukv sovremennogo literaturnogo gruzinskogo jazyka [Buchstabenstatistik der modernen georgischen Literatursprache]. Soobščenija AN GSSR 20, 1958 (Nr. 5), 565-567

GENKEL', M.A., Častotnyj slovar' romana D.N. Mamina-Sibirjaka "Privalovskie milliony" [Häufigkeitswörterbuch des Romans "Privalovskie milliony" von D.N. Mamin-Sibirjak]. In: Meždvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacii lingvo-statističeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 34-36

GLEZER, V.D., CUKKERMAN, I.I., Informacija i zrenie [Information und Sehen]. Moskva-Leningrad, Izdatel'stvo AN SSSR 1961

GOMEZ, C.F., Vocabulario de Cervantes. Madrid 1962

GUIRAUD, P., Les caractères statistiques du vocabulaire. Paris, Presses universitaires de France 1954

GUIRAUD, P., Problèmes et méthodes de la statistique linguistique. Paris, Presses universitaires de France 1960

HART, C., A concordance to "Finnegans Wake". Minneapolis, University of Minnesota Press 1963

HERDAN, G., The advanced theory of language as choice and chance. Berlin, Springer 1966

HILLER, L., ISAACSON, L., Experimental music. New York, McGraw-Hill 1963

IMNAJŠVILI, I., Konkordanz zum georgischen Evangelium (in georgischer Sprache). Tbilisi, AN GSSR 1948

JARBUS, A.L., Rol'dviženija glaz v procese zrenija [Die Rolle der Augenbewegung beim Prozess des Sehens]. Moskva, Nauka 1966

JELÍNEK, J., BEČKA, J.V., TEŠITELOVÁ, M., Frekvence slov, slovních druhů a tvarů v českém jazyce [Häufigkeit der Wörter, Wortarten und Wortformen im Čechischen]. Prag, Státni Pedagogické Nakladatelství 1961

JOOS, M., Rezension von Zipf (1935). Language 12, 1936, 196-210

JOSSELSON, H.H., The Russian word count and frequency analysis of grammar categories of standard literary Russian. Detroit, University Press 1953

KAC, B., O nekotorych čertach struktury variacionnogo cikla [Einige Merkmale der Struktur des Variationszyklus]. In: Voprosy teorii i estetiki muzyki, Bd. 2, Leningrad 1972

KAEDING, F.W., Häufigkeitswörterbuch der deutschen Sprache. Berlin, Selbstverlag des Herausgebers 1897-1898

KALININ, V.M., Nekotorye statističeskie zakony matematičeskoj lingvistiki [Einige statistische Gesetze der mathematischen Linguistik]. Problemy kibernetiki 11, 1964, 245-255

KALININ, V.M., Funkcionaly, svjazannye s raspredeleniem Puassona, i statičeskaja struktura teksta [Die mit der Poisson-Verteilung zusammenhängenden Funktionale und die statistische Struktur des Textes]. Trudy matematičeskogo instituta imeni V.A. Steklova 29, 1965, 182-197

KALININA, E.A., Častotnyj slovar' russkogo pod'jazyka elektroniki [Häufigkeitswörterbuch der russischen Teilsprache der Elektronik]. In: Piotrowski, R.G. (Hrsg.), Statistika reči. Leningrad, Nauka 1968, 144-150

KAŠIRINA, M.E., O tipach raspredelenija leksičeskich edinic v tekste [Über die Verteilungstypen lexikalischer Einheiten im Text]. In: Piotrowski, R.G. (Hrsg.), Statistika reči i avtomaticeskij analiz teksta. Leningrad, Nauka 1974, 144-150

KATUAR, G.L., Muzykal'naja forma. Bd. 1. [Die musikalische Form]. Moskva, Muzykal'noe gosudarstvennoe izdatel'stvo 1936

KOLGUŠKIN, A.N., Lingvistika v voennom dele [Die Linguistik im Militärwesen]. Moskva, Voenizdat 1970

KUNICKIJ, V.N., Jazyk i slog komedii A.S. Griboedova "Gore ot uma" [Sprache und Stil von Griboedovs Komödie "Verstand schafft Leiden"]. Kiev 1896

KVARACCHELIJA, G.S., Opyt sravmitel'no-statističeskogo analiza leksiki statističeskimi metodami [Versuch einer vergleichend-statistischen Analyse des Lexikons mit statistischen Methoden]. Voprosy sovremennogo obščego i matematičeskogo jazykoznanije, Vol. 2. Tbilisi, Mecniereba 1966, 152-182

LEBEDEV, D.S., GARMAŠ, V.A., Statističeskij analiz trechbukvennych sočetanij russkogo teksta. Problemy peredači informacii 2, 1958, 78-80

LENCMAN, Ja.A., Proizchoždenie christianstva [Der Ursprung des Christentums]. Moskva, AN SSSR 1960

LEVITSKIJ, V.V., Častotnyj slovar' učebnych posobij med-instituta [Häufigkeitswörterbuch der Lehrwerke des medizinischen Instituts]. Moskva, Izdatel'stvo Moskovskogo Universiteta 1966

LJATINA, A.M., Opyt statističeskogo analiza jazyka pisatelja (po materialam "Podnjatoj celiny" Šolochova) [Versuch einer statistischen Analyse der Schriftstellersprache (am Beispiel von Šolochovs "Neuland unterm Pflug")]. Dissertation, Leningrad 1968

LUDIN, D.M., Opyt opisanija statističeskimi metodami sovremennogo afganskogo jazyka (puštu) [Versuch einer Beschreibung des modernen Afghanischen (Puschtu) mit Hilfe statistischer Methoden]. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1971

LUK'JANENKOV, K.F., Leksiko-statističeskoe opisanie anglijskogo naučno-techničeskogo teksta s pomoščju elektronnoj vyčislitel'noj mašiny (pod'jazyk sudovych mechanizmov) [Lexikostatistische Beschreibung eines englischen wissenschaftlich-technischen Textes mit Hilfe der EDV (Teilsprache der Schiffsantriebswerke)]. Habilitationsschrift, Minsk 1969

MAJAKOVSKIJ, V.V., Kak delat'stichi [Wie macht man Gedichte]. Polnoe sobranie sočinenij. Bd. 12. Moskva, Chudožestvennaja Literatura 1959, 81-117

MANDELBROT, B., An information theory of the statistical structure of language. In: Jackson, W. (Hrsg.), Communication Theory, New York, Academic Press 1953, 503-512

MANDELBROT, B., O rekurrentnom kodirovanii, ograničivajuščem vlijanie pomech [On recurrent noise limiting coding]. In: Siforov, V.I. (Hrsg.), Teorija peredači soobščenij. Moskva, Izdatel'stvo inostrannoj literatury 1957, 139-157

MAZEL', L.A., CUKKERMAN, V.A., Analiz muzykal'nych proizvedenij [Analyse musikalischer Werke]. Moskva, Muzyka 1967

MELIK-GUSSEJNOVA, R.S., Častotnyj slovar' anglijskich tekstov po fizike tverdogo tela [Häufigkeitswörterbuch englischer Texte über Festkörperphysik]. In: Piotrowski, R.G. (Hrsg.), Statistika reči i avtomatičeskij analiz teksta. Leningrad, Nauka 1971, 191-196

MORGENTHALER, R., Statistik des Neutestamentlichen Wortschatzes. Zürich-Frankfurt, Gotthelf 1958

MULLER, Ch., Initiation à la statistique linguistique. Paris, Librairie Larousse 1968

MUSSO, N., Le vocabulaire de Figaro dans „Le mariage". Etudes de linguistique appliquée 6, 1972, 89-98

*NADAREJŠVILI, G.Š., NADAREJŠVILI, I.Š., ORLOV, Ju.K., Nestacionarnye javlenija v procese porožđenija teksta [Nichtstationäre Erscheinungen im Prozess der Texterzeugung]. Trudy Instituta kibernetiki AN GSSR 1, 1977, 275-285

NADAREJŠVILI, I.Š., Über die Häufigkeit der Präposition "und" im Evolutionsprozeß der georgischen Literatursprache (in georgischer Sprache). Soobscenija AN GSSR 57, 1970 (Nr.2), 505-507

*NADAREJŠVILI, I.Š., Sravnitel'nyj statističeskij analiz leksiki kak metod izučenija tvorčestva pisatelja [Die vergleichende statistische Analyse der Lexik als Methode zur Untersuchung des schriftstellerischen Schaffens]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 45-52

*NADAREJŠVILI, I.Š., ORLOV, Ju.K., Ob upotreblenii slov različnoj častnosti v poeme Rustaveli "Vitjaz' v tigrovoj škure" [Über die Verwendung von Wörtern unterschiedlicher Häufigkeit in Rustavelis Poem "Der Held im Tigerfell"]. Soobščenija AN GSSR 55, 1969 (Nr. 2), 505-508

NADAREJŠVILI, I.Š., ORLOV, Ju.K., Rost leksiki kak funkcija dliny teksta (na primere proizvedenij L.N. Tolstogo i Dž. Džojsa) [Das Anwachsen der Lexik als Funktion der Textlänge (am Beispiel der Werke von L.N. Tolstoj und J.Joyce)]. Soobščenija AN GSSR 64, 1971 (Nr. 3), 549-552

NADAREJŠVILI, I.Š., ORLOV, Ju.K., Opisanie rosta slovarnogo zapasa s pomoščju modeli častotnoj struktury leksiki [Die Beschreibung des Vokabularzuwachses mit Hilfe des Modells der Häufigkeitsstruktur der Lexik]. In: Cocua, T. (Hrsg.), Tezisy naucnoj sessii "Voprosy vzaimosvjazi isskustva i nauki". Tbilisi 1974, 8-9

*NADAREJŠVILI, I.Š., ORLOV, Ju.K., Metod polnoj fiksacii teksta pri lingvostatističeskom analize [Die Methode der vollständigen Textfixierung in der linguostatistischen Analyse]. Linguistica (Tartu) 10, 1978, 65-84

NALIMOV, V.V., Teoretičeskaja biologija? Ee vse ešče net... [Theoretische Biologie? Es gibt sie immer noch nicht...]. Znanie-sila 7, 1979, 9-11

NEŠITOJ, V.V., Ocenka leksičeskoj blizosti tekstov [Die Schätzung der lexikalischen Textähnlichkeit]. In: Deduškin, A.H., Nešitoj, V.V., Cvimetidze, S.V., (Hrsgs.), Razrabotka avtomatizirovannych sistem naučno-techničeskoj informacii. Minsk, Bel. NIINTI 1976, 13-35

NIKOLAEVA, A.U., Fortepiannyj stil' Skrjabina. Moskva, Diss. 1974

NOVAK, D.A., Častotnyj slovar' sovremennogo moldavskogo i rumynskogo jazykov [Häufigkeitswörterbuch der moldauischen und der rumänischen Gegenwartssprache]. Dissertation, Leningrad 1962

ORLOV, Ju.K., Obobščenie zakona Cipfa [Eine Verallgemeinerung des Zipfschen Gesetzes]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969a, 255-261

ORLOV, Ju.K., Leksičeskij spektr literaturnych tekstov i dinamika rosta slovarja [Das lexikalische Spektrum literarischer Texte und die Dynamik des Vokabularzuwachses]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969b, 250-254

ORLOV, Ju.K., O statističeskoj strukture soobščenij, optimal'nych dlja čelovečeskogo vosprijatija [Zur statistischen Strukturen der für die menschliche Perzeption optimalen Nachrichten]. Naučno-techničeskaja informacija Serija 2, 1970a (Nr. 8), 11-16

ORLOV, Ju.K., Obobščenie zakona Cipfa-Mandel'brota [Eine Verallgemeinerung des Zipf-Mandelbrotschen Gesetzes]. Soobščenija AN GSSR 57, 1970b (Nr. 1), 37-40

ORLOV, Ju.K., Častotnye struktury konečnych soobščenij v nekotorych estestvennych informacionnych sistemach [Die Häufigkeitsstrukturen endlicher Nachrichten in einigen natürlichen Informationssystemen]. Habilitationsschrift, Tbilisi 1974

*ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i častotnye struktury informacionnych edinic različnych urovnej [Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Häufigkeitsstrukturen der Informationseinheiten verschiedener Ebenen]. In: Guseva, E.K. (Hrsg.), Vyčislitel'naja lingvistika. Moskva, Nauka 1976, 179-202

ORLOV, Ju.K., Nekotorye aspekty organizacii informacii čelovekom [Einige Aspekte der Informationsorganisation durch den Menschen]. Trudy Instituta kibernetiki AN GSSR 1, 1977, 267-275

*ORLOV, Ju.K., Model' častotnoj struktury leksiki [Das Modell der Häufigkeitsstruktur der Lexik]. In: Andrjuščenko, V.M. (Hrsg.), Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva, Moskovskij Gosudarstvennyj Universitet 1978a, 59-118

ORLOV, Ju.K., Statističeskoe modelirovanie rečevych potokov [Die statistische Modellierung des Redeflusses]. In: Piotrowski, R.G. (Hrsg.), Voprosy kibernetiki 41: Statistika reči i avtomatičeskij analiz teksta 1978b, 66-106

ORLOV, Ju.K., Informacionnye potoki: statističeskij analiz i prognozirovanie [Informationsflüsse: Statistische Analyse und Prognostizierung]. Naučno-techničeskaja informacija, Serija 2, 1980 (Nr. 2), 23-30

OSMANOV, M.N.O., Častotnyj slovar' Unsuri [Häufigkeitswörterbuch von Unsuri]. Moskva, Nauka 1970

OVSIENKO, Ju.G., Slovar' russkoj razgovornoj reči [Wörterbuch der russischen Umgangssprache]. In: Mežvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacija lingvostatističeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 23-25

PEREBEJNOS, V.I. (Hrsg.), Častotnyj slovar' sovremennoj ukrainskoj chudožestvennoj prozy [Häufigkeitswörterbuch der modernen ukrainischen künstlerischen Prosa]. Kiev, Institut jazykovedenija AN USSR 1969

PIOTROWSKI, R.G., Tekst, mašina, celovek [Text, Computer, Mensch]. Leningrad, Nauka 1975

PIRS, Dz., Simvoly, signaly, šumy [Pierce, J.R., Symbols, signals and noises: the nature and process of communication. London 1962]. Moskva, Mir 1967

POLIKAROV, A.A. Elementy teoretičeskoj sociolingvistiki [Elemente der theoretischen Soziolinguistik]. Moskva, Izdatel'stvo Moskovskogo Universiteta 1979

ROBERTSON, A., Proizchoždenie christianstva [Robertson, A., The origins of Christianity, London, Lawrence & Wishert 1953]. Moskva, Izdatel'stvo inostrannoj literatury 1959

RUSTAVELI, S., Vephis-ṭqaosani. Tbilisi, Izdatel'stvo AN GSSR 1956

SAMBOR, J., Badania statystyczne nad slownictwem (na materiale „Pana Tadeusza"). Wroclaw - Warszawa - Krakow, 1969

ŠANIDZE, A. (Hrsg.), Konkordanz zu S. Rustavelis Poem "Der Held im Tigerfell" (in georgischer Sprache). Tbilisi, Izdatel'stvo Tbiliskogo Gosudarstvennogo Universiteta 1956

ŠAJKEVIČ, A.Ja., Differenciacija statističeskich klassifikacij tekstov [Die Differenzierung statistischer Textklassifikationen]. Linguistika 11, 1979, 100-106

SEEGER, H. Musiklexikon in zwei Bänden. Leipzig, VEB Deutscher Verlag für Musik 1966

SÉNÈQUE, De Clementia. Index verborum. Relèves statistiques. La Haye, Mouton 1968

SÉNÈQUE, De Brevitate Vitae. Index verborum. Relèves statistiques. La Haye, Mouton 1968

SIMON, H.A., Some further notes on a class of skew distribution functions. Information and Control 3, 1960, 90-98

SINENKO, G.D., Častotnyj slovar' "Povesti o detstve" F.V. Gladkova i voprosy literaturnogo stilja [Häufigkeitswörterbuch von "Erzählungen von der Kindheit" von F.V. Gladkov und Probleme des literarischen Stils]. Vesnik Belaruskaga dzjaržajnaga universiteta imja V.I. Lenina. Seryja IV/1. Minsk 1973, 40-46

SOSSJUR, F. de, Kurs obščej lingvistiki [Saussure, F. de, Cours de linguistique générale]. Moskva, Progress 1977

SPEVACK, M., A complete and systematic concordance to the works of Shakespeare. Vol. I,II. Hildesheim, Olms 1968

SPOSOBIN, I.V., Muzykal'naja forma [Die musikalische Form]. Moskva, Muzyka 1972

ŠTEJNFELD, E.A., Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka [Häufigkeitswörterbuch der modernen russischen Schriftsprache]. Tallin, NII Pedagogiki ESSR 1963

SUDAVIČENE, L., Iz opyta sostavlenija častotnogo slovarja jazyka K. Paustovskogo [Aus dem Versuch der Erstellung des Häufigkeitswörterbuchs der Sprache von K. Paustovskij]. Vil'njus, Učenye zapiski vysšich učebnych zavedenij Litovskoj SSR. Jazykoznanije 22, 1971

TEŠITELOVÁ, M., On the so-called vocabulary richness. Prague Studies in Mathematical Linguistics 3, 1972, 103-120

TJULIN, Ju.N. (Hrsg.), Muzykal'naja forma [Die musikalische Form]. Moskva, Muzyka 1974

TOKAREV, V.P., JAKUBAJTIS, T.A., Matematiko-statističeskaja model' raspredelenija leksem [Ein mathematisch-statistisches Modell der Lexemverteilung]. In: Tjurina, L. (Hrsg.), Matematičeskie metody v jazykoznanii. Riga, Zinatne 1969, 7-46

TOMBEUR, P., R. de Saint-Trond: Gesta Abbatum trudonensium I-VIII. Index verborum. Relêves statistiques. La Haye, Mouton 1965

TULDAVA, Ju.A., Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov [Eine statistische Methode für den Vergleich des Wortschatzes von zwei Texten]. Linguistica 4, 1971, 199-200

TVOROGOV, O.V., O primenenii častotnych slovarej v istoričeskoj leksikologii russkogo jazyka [Über die Anwendung von Häufigkeitswörterbüchern in der historischen Lexikologie des Russischen]. Voprosy jazykoznanija 1967/2, 109-117

VJALKINA, L.V., LUKINA, G.N., Opyt primenenija nekotorych metodov matematičeskoj statistiki k izučeniju drevnerusskich tekstov [Versuch der Anwendung einiger mathematisch-statistischer Methoden zur Untersuchung altrussischer Texte]. In: Avanesov, R.I. (Hrsg.), Issledovanija po istoričeskoj leksikologii drevnerusskogo jazyka. Moskva, Nauka 1964, 298-307

*VOLOŠIN, B.A., ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i raspredelenie dolej cvetovych ploščadej v proizvedenijach živopisi [Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Verteilung der Farbflächen in Gemälden]. Tbilisi, Institut kibernetiki AN GSSR 1972

VREDE, V., Religija i cerkov' v cvete naučnoj mysli i svobodnoj kritiki. Kniga 3-ja: Proizchoždenie knig Novogo Zaveta [Der Ursprung der Bücher des Neuen Testaments]. Moskva 1908

YULE, G.U., The statistical study of literary vocabulary. London, Cambridge University Press 1944

ŽANTIEVA, D.G., Džejms Džojs [James Joyce]. Moskva, Vysšaja skola 1967

ZASORINA, L.N., Avtomatizacija i statistika v leksikografi [Automatisierung und Statistik in der Lexikographie]. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966

ZASORINA, L.N. (Hrsg.), Častotnyj slovar' russkogo jazyka [Häufigkeitswörterbuch des Russischen]. Moskva, Russkij jazyk 1977

ZIPF, G.K., The psycho-biology of language. Boston, Houghton Mifflin Company 1935

ZIPF, G.K., Human behavior and the principle of least effort. Cambridge (Mass.), Addison-Wesley Press 1949

REGISTER