QUANTITATIVE LINGUISTICS

Vol. 45

Glottometrika 12

edited

by

Rolf Hammerl

Universitätsverlag Dr. N. Brockmeyer
Bochum 1990

QUANTITATIVE LINGUISTICS

CIP-Titelaufnahme der Deutschen Bibliothek

Glottometrika 12 / by edited Rolf Hammerl – Bochum: Universitätsverl. Brockmeyer.
ISSN 0932-7991

12 (1990)
(Quantitative linguistics ; Vol. 45)
ISBN 3-88339-843-8

NE: GT

ISBN 3-88339-843-8

Querenburger Höhe 239, 4630 Bochum 1
Gesamtherstellung: Druck Thiebes GmbH & Co. KG Hagen

Die Aufsätze dieses Sammelbandes (mit Ausnahme der Artikel von J.B.M. Guy, L. Hrebicek, D. Rumpel und M. Trauth) entstanden im Rahmen des von der *Stiftung Volkswagenwerk* geförderten Projekts "Sprachliche Synergetik".

Als Herausgeber dieses Sammelbandes möchte ich der *Stiftung Volkswagenwerk* und der *Alexander von Humboldt-Stiftung* danken für die Unterstützung der Forschungen zur sprachlichen Synergetik.

R.H.

Inhaltsverzeichnis

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 1–24

Länge–Frequenz, Länge–Rangnummer: Überprüfung von zwei lexikalischen Modellen

Rolf Hammerl
Bochum

1. Ziel der Untersuchungen

Bei der Überprüfung des von Köhler (1986) erarbeiteten Basismodells zur Beschreibung der Abhängigkeiten zwischen der Länge und Häufigkeit von Lexemen, der Polylexie und Länge, der Polytextie und Polylexie und der Frequenz und Polytextie an deutschem Sprachmaterial wurde festgestellt, daß besonders für die Länge L in Abhängigkeit von der Frequenz F (Länge L als abhängige Größe, Frequenz F als unabhängige Größe) eine relativ starke Streuung der empirischen Werte auftrat. Es wurde vermutet, daß es sich hier um eine "Oszillation der Lexik" (Köhler 1986, 137ff.) handelt, die in Form einer "um die theoretische hyperbolische Funktion schwingende Kurve" an der Frequenzachse (nicht in der Zeit-Dimension) in Erscheinung tritt.

Das Problem der Oszillation der Lexik wurde von uns schon an anderer Stelle diskutiert (vgl. Hammerl 1989): Solange eine Oszillation (d.h. periodische Schwingungen) nicht nachgewiesen werden konnte (durch Anwendung spezieller Tests), solange sollte man von einer zufälligen Streuung der Werte für die Lexemlänge in Abhängigkeit von der Frequenz sprechen.

Als Ursache dieser Streuung sehen wir mehrdimensionale Veränderungen solcher Eigenschaften wie der Frequenz, Polylexie und Länge von konkreten Lexemen in der Zeit an (vgl. Hammerl 1989).

Ziel dieses Aufsatzes ist es, erstens das von Köhler abgeleitete Modell

$$L = g \cdot F^k, \tag{1}$$

wo g und k Konstanten sind, an polnischem Sprachmaterial zu überprüfen und auch das Problem der "Oszillation" der Lexik zu untersuchen.

Bei der Überprüfung der "Oszillation" der Lexik auf Zufälligkeit werden zwei spezielle Tests angewandt.

Der Zusammenhang zwischen der Anwendungshäufigkeit von Lexemen und der Lexemlänge wurde von Arapov (1988, 112ff.) als

$$L = a + b \ln r \tag{2}$$

beschrieben, wo r die Rangnummer der nach abnehmender Häufigkeit geordneten Lexeme in einem Häufigkeitswörterbuch ist; a und b stellen wiederum Konstanten dar. Arapov verwendet also als Ausdruck der Anwendungshäufigkeit ("Gebräuchlichkeit", russ. "upotrebitel'nost'") nicht die Frequenz F wie Köhler, sondern die Rangnummer r. Dies ist auch prinzipiell möglich, da ja zwischen der Frequenz F und der Rangnummer r in vielen Untersuchungen ein Zusammenhang nachgewiesen worden ist, der als Zipfsches Gesetz bekannt ist.

Wir wollen in diesem Aufsatz beide Modelle – nach einer kurzen Diskussion deren Herleitung – in empirischen Untersuchungen überprüfen und auch bestimmte Deduktionen aus diesen Modellen.

2. Zipfsches Gesetz

Der Zusammenhang zwischen der Rangnummer r und der Frequenz F wird als Zipfsches Gesetz bezeichnet und wurde von Zipf (1949) in der Form

$$F = c \cdot r^d \tag{3}$$

dargestellt, wo c und d wiederum Konstanten sind (d<0). Dieses Gesetz wurde mehrmals modifiziert und auf recht verschiedene Weisen abgeleitet (vgl. z.B. Mandelbrot 1953, 1954a,b; Orlov 1982; Arapov 1988). Trotz der vielen Untersuchungen, die diesem Gesetz gewidmet waren, sind viele Probleme bisher noch nicht befriedigend gelöst worden (vgl. Altmann 1988, 70ff.):

a) Der Rang ist eine Hilfsvariable, welche eine spezielle Art der Ordnung der Lexeme (nach abnehmender Häufigkeit) beschreibt; die Frequenz ist eine Zufallsvariable (vgl. Kapitel 3).

b) Gilt das Zipfsche Gesetz nur für das Vokabular bestimmter (z.B. "vollendeter") Texte oder auch für das Gesamtvokabular mehrerer Texte?
c) Gilt das Gesetz nur für Lexeme oder auch für andere sprachliche und außersprachliche Einheiten?
d) Wie können die Konstanten des Zipfsches Gesetzes interpretiert werden?

Wir wollen nun nur auf zwei Probleme kurz eingehen. Orlov (1982) hat festgestellt, daß das Zipfsche Gesetz nur für einzelne Texte mit einer bestimmten Länge gilt, die als "vollendete" Texte bezeichnet werden; der Begriff des "vollendeten" Textes wird aber nicht hinreichend präzisiert (vgl. auch Arapov 1988, 31). Orlov begründet, daß das Vokabular von Häufigkeitswörterbüchern, die auf der Grundlage mehrerer Textproben erarbeitet wurden, einen "fiktiven" Text betrifft, für welchen das Zipfsche Gesetz nicht gelten muß. Diese Bedenken beruhen vor allem darauf, daß in empirischen Untersuchungen für das Vokabular solcher Häufigkeitswörterbücher in der Regel im Bereich hoher und niedriger Häufigkeiten Abweichungen vom Modell (Zipfsches Gesetz) aufgetreten sind. Außerdem gibt es immer relativ viele Wörter mit kleinen Häufigkeiten (z.B. F=1,2,3), die aber alphabetisch geordnet werden und deshalb unterschiedliche Rangnummern erhalten. Bisher wurde die Frage noch nicht untersucht, inwieweit eine solche Ordnung gerechtfertigt ist.

Auch wenn man das Vokabular einzelner vollendeter Texte untersucht, tritt das genannte Problem auf, daß die Zahl der lexikalischen Einheiten mit einer geringen Häufigkeit relativ hoch ist und daß Einheiten mit derselben Häufigkeit unterschiedliche Rangnummern erhalten. Also wird es auch hier im Bereich kleiner Häufigkeiten Abweichungen vom Modell geben.

Einen neuen und sehr wertvollen Beitrag führt Arapov (1988, 26ff.) in die Diskussion um die Gültigkeit des Zipfschen Gesetzes ein. Er lenkt die Aufmerksamkeit darauf, daß bisher in der Regel Modifikationen des Zipfschen Gesetzes als Ergebnis empirischer Verallgemeinerungen vorgenommen wurden. Bei der Ableitung des Zipfschen Gesetzes wird aber immer von relativ einfachen Textmodellen ausgegangen, welche die quantitative Struktur realer Texte eben nur zu einem bestimmten Teil erklären können. Somit sind auch die Abweichungen der

empirischen Werte von den theoretischen Werten, welche das Zipfsche Gesetz voraussagt, verständlich. Erst die Berücksichtigung weiterer Eigenschaften der quantitativen Struktur realer Texte kann zur Ableitung von Modellen führen, welche die oben genannten Probleme lösen können.

Wir leiten für unsere Untersuchungen in diesem Aufsatz folgende Schlußfolgerung ab:

Das Zipfsche Gesetz (z.B. in der Form, wie es in Gleichung (3) gegeben ist) beschreibt die quantitative Struktur für das Vokabular von realen Texten immer mit einer bestimmten Ungenauigkeit (so gilt es z.B. für sprachliche Einheiten, die alle in der sprachlichen Realität eine unterschiedliche Häufigkeit besitzen, denn jedem Wert für den Rang r wird eine andere Häufigkeit zugeordnet; außerdem werden die spezifischen Eigenschaften grammatischer Einheiten, die in der Regel die ersten Plätze in der Rangordnung einnehmen, nicht berücksichtigt). Es gilt somit approximativ auch für das Vokabular in Häufigkeitswörterbüchern (vgl. Kapitel 4.4).

3. Rangnummer als Hilfsvariable

3.1. Das Zipfsche Gesetz stellt eine Relation zwischen einer Zufallsvariablen (der Frequenz) und einer "Hilfsvariablen" (der Rangnummer) dar. Was verbirgt sich aber hinter der Rangnummer eigentlich?

Sicherlich ist die Rangnummer eine besondere "Hilfsvariable", denn in vielen Untersuchungen aus verschiedenen wissenschaftlichen Disziplinen konnte ein Zusammenhang zwischen der Rangnummer und der Frequenz festgestellt werden, der durch das Zipfsche Gesetz approximativ beschrieben werden konnte.

In Klassifikationsexperimenten (vgl. Arapov 1988, 26ff.) konnte z.B. gezeigt werden, daß die Respondenten immer solche Klassifizierungen vornehmen, wo wenige relativ große Klassen und mehrere entsprechend kleinere Klassen unterschieden werden. Auch die Zahl der unterschiedenen Klassen wird nicht beliebig gewählt. Die Ergebnisse dieser Experimente lassen die Vermutung zu, daß sich hinter dieser Art der Klassifizierung ein bestimmtes Ordnungsprinzip verbirgt, daß unbewußt eingehalten wird.

Bei der Widerspiegelung der Realität durch den Menschen laufen ständig Klassifikationsverfahren ab, welche die Vielfalt der Realität in einer für den Menschen adäquaten Form festhalten, d.h. in einer solchen Form, die den Prinzipien der menschlichen Informationsverarbeitung und -speicherung gerecht wird. Die Grenzen dieser Klassen sind nicht eindeutig angebbar, da stets kontinuierliche Veränderungen in der gebildeten Struktur auftreten. Größe (Umfang) dieser Klassen und auch deren Zahl ist somit nicht beliebig frei wählbar. Die Klassen können z.B. nach deren Größe strukturiert werden; es könnte in diesen Fällen auch eine bestimmte Reihenfolge solcher Klassen angegeben werden, die durch natürliche Zahlen (Rangnummer) beschrieben wird.

Sicher unterliegen nicht alle Erscheinungen einer Klassifizierung nach derselben Struktur, da sich z.B. die vielfältigen Klassifikationsprozeduren gegenseitig beeinflussen können, was zur Bildung neuer Strukturen führen kann. Wir vermuten aber, daß bestimmte grundlegende strukturelle Eigenschaften einem allgemeinen Ordnungsprinzip gehorchen, welches im menschlichen Verhalten (auch in der Sprache) beobachtbar ist. Zu einer solchen Eigenschaft könnte auch die Häufigkeit von Lexemen gezählt werden, genauer: die Ordnung der Lexeme nach abnehmender "subjektiver" Häufigkeit. Diese "unscharfe, subjektive" Struktur spiegelt sich in den Texten einer Sprache wieder und kann aus diesen Texten abgelesen und approximiert werden. Die Rangnummer r stellt dabei immer die Zahl der Klassen dar, deren Umfänge größer oder gleich dem Umfang der Klasse mit der Rangnummer r ist.

3.2. Will man nun diese Struktur des Vokabulars beschreiben, so muß sowohl die Häufigkeit als auch die entsprechende Rangnummer der lexikalischen Einheiten bekannt sein. Die Häufigkeiten resultieren direkt aus den empirischen Untersuchungen, die Rangnummern werden dagegen erst im Nachhinein festgelegt.

In den meisten Fällen werden die Einheiten nach abnehmender Häufigkeit geordnet und innerhalb der Gruppen mit derselben Häufigkeit nach dem Alphabet (anders z.B. bei Hoffmann 1980).

Bedenkt man aber, daß viele Lexeme andere Rangnummern erhalten, obwohl sich deren Häufigkeiten nicht unterscheiden, und daß eine solche Bedingung vom Zipfschen Modell nicht berücksichtigt wird, so muß daraus folgende Schlußfolgerung gezogen werden: Alle m Le-

xeme mit derselben Häufigkeit sollten durch die mittlere Rangnummer $\bar{r}$ (arithmetischer Mittelwert) dieser Lexemgruppe repräsentiert werden, wobei

$$\bar{r} = \frac{r_1 - r_2}{2} \tag{4}$$

gilt (r_1 stellt die maximale Rangnummer und r_2 die minimale Rangnummer dieser Lexeme dar), bzw. durch das gewogene Mittel

$$\bar{r} = \frac{\sum_{i=1}^{m} F_i r_i}{\sum_{i=1}^{m} F_i}. \tag{5}$$

Auch im Bereich großer Häufigkeiten sollte man ähnlich verfahren, da jede Häufigkeit dort in der Regel durch eine einzige (grammatische) Einheit repräsentiert wird und der spezielle Charakter dieser Einheiten im Zipfschen Gesetz nicht speziell berücksichtigt wurde. Wir haben deshalb in unseren Untersuchungen im Kapitel 4 jeweils m der häufigsten Lexeme zusammengefaßt ($m > 9$) und nach Gleichung (4) deren mittleren Rang und nach Gleichung (6) deren mittlere Frequenz berechnet. Es wurde aber stets darauf geachtet, daß Lexeme mit derselben Häufigkeit nicht zu verschiedenen Klassen gehören.

$$\bar{F} = \frac{\sum_{i}^{m} F_i}{m} \tag{6}$$

Arapov (1988, 125) wählt ein teilweise anderes Vorgehen: Er läßt Wörter mit geringen Häufigkeiten ganz außer acht. Wörter mit größeren Häufigkeiten werden zu Gruppen mit je 25 Wörtern zusammengefaßt und deren mittlere Häufigkeit und Frequenz berechnet. Nicht erwähnt wird aber, ob Lexeme mit derselben Häufigkeit (die ja alphabetisch geordnet wurden) zu verschiedenen Klassen gehören können oder nicht. Vermutlich wurden sie in bestimmten Fällen unterschiedlichen Klassen zugeordnet, was besser vermieden werden sollte, da ja die alphabetische Ordnung dieser Lexeme keine inhaltliche Begründung hat.

Hoffmann (1980) wendet den Begriff Rangnummer in einer etwas anderen Bedeutung an, denn alle Lexeme mit derselben Häufigkeit erhalten dieselbe Rangnummer. Somit ist die Zahl der verwendeten unterschiedlichen Rangnummern relativ klein. Hoffmann umgeht dadurch das Problem, Lexemen mit derselben Häufigkeit unterschiedliche Rangnummern zuordnen zu müssen. Kann aber bei einer solchen Verwendung des Begriffes *Rangnummer* das Zipfsche Gesetz genauso gut beschrieben werden wie bei der obigen Verwendung des Begriffes Rangnummer? Dieses Problem wird im Kapitel 4 an einem Beispiel untersucht.

4. Ableitung der Modelle und empirische Untersuchungen

Zur Überprüfung der Abhängigkeiten der Länge von der Frequenz bzw. der Länge von der Rangnummer werden wir das Vokabular aus zwei polnischen Häufigkeitslisten untersuchen. Die erste Häufigkeitsliste (Probe I) betrifft 50 Gedichte von M. Sęp Szarzyński (vgl. Fleischer 1988, 87ff) mit einem Gesamtumfang von 5653 Tokens und 1703 Types (Lexeme) und die zweite Häufigkeitsliste (Probe II) das polnische Häufigkeitswörterbuch (Słownictwo 1974-1977), welches nach der Eliminierung von Eigennamen, Fremdwörtern und fremdsprachigen Zitaten einen Umfang von 30041 Lexemen hat. Dieses Häufigkeitswörterbuch entstand auf der Grundlage von 500 Textproben (aus fünf Funktionalstilen) mit einer Länge von je 1000 Wortstellen. In beiden Proben wurde die Lexemlänge als Zahl der Silben gemessen. Für Probe I wurden zwei Fälle unterschieden:

a) Nur Lexeme mit derselben Häufigkeit wurden in Klassen zusammengefaßt (d.h. die häufigsten Lexeme wurden nicht in Klassen zusammengefaßt, so wie es im Kapitel 3.1 angegeben wurde).
b) Auch häufige Lexeme wurden zusammengefaßt (vgl. Kapitel 3.1).

Die empirischen Daten für die Abhängigkeit der Länge von der Rangnummer bzw. der Frequenz (Probe Ia, Ib und II) werden in Tabelle 1, 2 und 3 angeführt.

In allen Untersuchungen wurde die mittlere Rangnummer $\bar{r}$ für die Lexeme dieser Klassen – falls diese keine ganze Zahl war – stets auf die nächsthöhere ganze Zahl aufgerundet.
Für die Überprüfung der Güte der Anpassung der nach den (in den folgenden Kapiteln beschriebenen) Modellen errechneten Daten an die

xeme mit derselben Häufigkeit sollten durch die mittlere Rangnummer $\bar{r}$ (arithmetischer Mittelwert) dieser Lexemgruppe repräsentiert werden, wobei

$$\bar{r} = \frac{r_1 - r_2}{2} \tag{4}$$

gilt (r_1 stellt die maximale Rangnummer und r_2 die minimale Rangnummer dieser Lexeme dar), bzw. durch das gewogene Mittel

$$\bar{r} = \frac{\sum_{i=1}^{m} F_i r_i}{\sum_{i=1}^{m} F_i}. \tag{5}$$

Auch im Bereich großer Häufigkeiten sollte man ähnlich verfahren, da jede Häufigkeit dort in der Regel durch eine einzige (grammatische) Einheit repräsentiert wird und der spezielle Charakter dieser Einheiten im Zipfschen Gesetz nicht speziell berücksichtigt wurde. Wir haben deshalb in unseren Untersuchungen im Kapitel 4 jeweils m der häufigsten Lexeme zusammengefaßt ($m > 9$) und nach Gleichung (4) deren mittleren Rang und nach Gleichung (6) deren mittlere Frequenz berechnet. Es wurde aber stets darauf geachtet, daß Lexeme mit derselben Häufigkeit nicht zu verschiedenen Klassen gehören.

$$\bar{F} = \frac{\sum_{i}^{m} F_i}{m} \tag{6}$$

Arapov (1988, 125) wählt ein teilweise anderes Vorgehen: Er läßt Wörter mit geringen Häufigkeiten ganz außer acht. Wörter mit größeren Häufigkeiten werden zu Gruppen mit je 25 Wörtern zusammengefaßt und deren mittlere Häufigkeit und Frequenz berechnet. Nicht erwähnt wird aber, ob Lexeme mit derselben Häufigkeit (die ja alphabetisch geordnet wurden) zu verschiedenen Klassen gehören können oder nicht. Vermutlich wurden sie in bestimmten Fällen unterschiedlichen Klassen zugeordnet, was besser vermieden werden sollte, da ja die alphabetische Ordnung dieser Lexeme keine inhaltliche Begründung hat.

Hoffmann (1980) wendet den Begriff Rangnummer in einer etwas anderen Bedeutung an, denn alle Lexeme mit derselben Häufigkeit erhalten dieselbe Rangnummer. Somit ist die Zahl der verwendeten unterschiedlichen Rangnummern relativ klein. Hoffmann umgeht dadurch das Problem, Lexemen mit derselben Häufigkeit unterschiedliche Rangnummern zuordnen zu müssen. Kann aber bei einer solchen Verwendung des Begriffes *Rangnummer* das Zipfsche Gesetz genauso gut beschrieben werden wie bei der obigen Verwendung des Begriffes Rangnummer? Dieses Problem wird im Kapitel 4 an einem Beispiel untersucht.

4. Ableitung der Modelle und empirische Untersuchungen

Zur Überprüfung der Abhängigkeiten der Länge von der Frequenz bzw. der Länge von der Rangnummer werden wir das Vokabular aus zwei polnischen Häufigkeitslisten untersuchen. Die erste Häufigkeitsliste (Probe I) betrifft 50 Gedichte von M. Sęp Szarzyński (vgl. Fleischer 1988, 87ff) mit einem Gesamtumfang von 5653 Tokens und 1703 Types (Lexeme) und die zweite Häufigkeitsliste (Probe II) das polnische Häufigkeitswörterbuch (Słownictwo 1974-1977), welches nach der Eliminierung von Eigennamen, Fremdwörtern und fremdsprachigen Zitaten einen Umfang von 30041 Lexemen hat. Dieses Häufigkeitswörterbuch entstand auf der Grundlage von 500 Textproben (aus fünf Funktionalstilen) mit einer Länge von je 1000 Wortstellen. In beiden Proben wurde die Lexemlänge als Zahl der Silben gemessen. Für Probe I wurden zwei Fälle unterschieden:

a) Nur Lexeme mit derselben Häufigkeit wurden in Klassen zusammengefaßt (d.h. die häufigsten Lexeme wurden nicht in Klassen zusammengefaßt, so wie es im Kapitel 3.1 angegeben wurde).
b) Auch häufige Lexeme wurden zusammengefaßt (vgl. Kapitel 3.1).

Die empirischen Daten für die Abhängigkeit der Länge von der Rangnummer bzw. der Frequenz (Probe Ia, Ib und II) werden in Tabelle 1, 2 und 3 angeführt.

In allen Untersuchungen wurde die mittlere Rangnummer $\bar{r}$ für die Lexeme dieser Klassen – falls diese keine ganze Zahl war – stets auf die nächsthöhere ganze Zahl aufgerundet.
Für die Überprüfung der Güte der Anpassung der nach den (in den folgenden Kapiteln beschriebenen) Modellen errechneten Daten an die

Tabelle 1
Abhängigkeit der Länge $\bar{L}$ von der Rangnummer $\bar{r}$ bzw. der Frequenz $\bar{F}$ (Probe Ia) (r_H gibt die Rangzählung von Hoffmann (1980) wider)

$\bar{r}$	r_H	$\bar{F}$	$\bar{L}$	$\bar{r}$	r_H	$\bar{F}$	$\bar{L}$
1	1	186	1,0	32	23	23	3,50
2	2	125	3,0	33	24	22	3,00
3	3	117	1,0	35	25	21	2,66
4	4	97	3,0	37	26	20	6,00
5	5	78	1,0	39	27	19	4,00
6	6	72	4,0	41	28	18	4,00
7	7	69	2,0	43	29	17	4,00
8	8	62	5,0	45	30	16	6,67
9	9	61	2,0	48	31	14	6,33
11	10	60	2,5	54	32	13	5,00
12	11	49	1,0	63	33	12	4,60
14	12	45	3,5	69	34	11	7,00
15	13	44	2,0	77	35	10	4,38
16	14	41	2,0	89	36	9	4,55
18	15	39	3,5	104	37	8	5,61
20	16	38	3,0	124	38	7	5,09
22	17	36	3,0	156	39	6	5,35
24	18	34	2,0	199	40	5	5,64
25	19	33	2,0	264	41	4	6,27
26	20	30	2,0	380	42	3	5,96
28	21	29	2,33	599	43	2	6,49
30	22	24	4,0	1225	44	1	6,97

entsprechenden empirischen Daten haben wir nicht den F-Test verwendet. Erstens gilt dieser Test nur für lineare Modelle (das Modell von Arapov ist linear; das Modell von Köhler kann linearisiert werden); nichtlineare Modelle müssen – soweit es überhaupt möglich ist – linearisiert werden, wobei der F-Test aber für die linearisierten Daten gilt und nur mit einem bestimmten (aber unbekannten) Fehler für das entsprechende nichtlineare Modell. Zweitens setzt dieser Test voraus, daß

Tabelle 2
Abhängigkeit der Länge $\bar{L}$ von der Rangnummer $\bar{r}$ bzw. der Frequenz $\bar{F}$ (Probe Ib)

$\bar{r}$	$\bar{F}$	$\bar{L}$	$\bar{r}$	$\bar{F}$	$\bar{L}$
6	89,73	2,45	104	8	5,61
17	41,60	2,80	124	7	5,09
27	29,46	2,64	156	6	5,35
38	19,70	3,70	199	5	5,64
50	14,07	5,53	264	4	6,27
64	11,77	5,15	380	3	5,96
77	10,00	4,38	599	2	6,49
89	9,00	4,55	1225	1	9,97

die Daten normalverteilten Grundgesamtheiten mit derselben Varianz entnommen wurden. Diese Bedingungen werden aber in vielen Fällen nicht erfüllt.

Wir haben in unseren Untersuchungen, wo wir es ausschließlich mit Modellen mit nur 2 Parametern zu tun haben, die beide auf dieselben empirischen Daten angewandt werden, den multiplen Determinationskoeffizienten D berechnet:

$$D = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{7}$$

wo $\hat{y}_i$ die theoretischen Werte der jeweiligen abhängigen Größe (bei uns: der Länge L) und y_i die entsprechenden empirischen Werte darstellen, deren Mittelwert $\bar{y}$ ist (vgl. auch: Rogalinska, Hammerl 1990).

In allen empirischen Untersuchungen, die in den folgenden Kapiteln besprochen werden, hätte die Anwendung des F-Testes erbracht, daß die entsprechenden Modelle die empirischen Daten signifikant beschreiben (wir haben das in allen Fällen überprüft). Wie die berechneten D-Werte jedoch aussagen, ist der Anteil der durch das Modell erklärten Varianz der empirischen Werte noch relativ gering. Der F-Test könnte somit die Vermutung erbringen, daß im jeweiligen Modell schon die wesentlichsten unabhängigen Variablen für die Erklärung der Varianz der abhängigen Variablen berücksichtigt wurden. Dies wäre natürlich

Tabelle 3
Abhängigkeit der Länge $\bar{L}$ von der Rangnummer $\bar{r}$ bzw. Frequenz $\bar{F}$ (Probe II)

$\bar{r}$	$\bar{F}$	$\bar{L}$	$\bar{r}$	$\bar{F}$	$\bar{L}$	$\bar{r}$	$\bar{F}$	$\bar{L}$	$\bar{r}$	$\bar{F}$	$\bar{L}$
8	7272,7	1,00	378	150,8	2,42	883	73	2,67	1828	35	2,78
21	2079,7	1,20	390	147,1	2,69	901	71,6	2,75	1881	34	2,95
31	1343,3	1,40	403	144,3	3,08	923	69,6	2,92	1924	33	2,55
41	1021,7	1,60	416	142,2	2,53	946	68	2,60	1972	32	2,62
51	867,9	1,50	430	138,1	2,33	967	67	2,23	2035	31	2,68
61	744,9	1,50	441	135,3	2,73	990	66	2,54	2092	30	2,91
71	651,4	1,70	452	132,7	2,30	1009	65	2,53	2152	29	2,68
81	569,6	1,70	463	129,5	2,15	1026	64	2,53	2220	26	2,99
91	490,5	1,70	476	125,2	2,33	1041	63	2,91	2291	27	2,76
101	450,4	1,80	489	122,4	2,21	1056	62	2,26	2371	26	2,81
111	416,9	2,09	504	119,3	2,75	1072	61	2,77	2451	25	2,84
122	393,7	1,70	520	116,4	2,56	1091	59,3	3,16	2523	24	2,87
132	372,2	1,90	537	113,4	2,58	1116	58	2,54	2607	23	2,99
142	350,6	1,90	555	110,5	2,29	1138	57	2,57	2708	22	2,89
152	331,2	2,00	573	108,6	2,74	1158	56	2,94	2816	21	2,95
162	312,4	1,80	588	105,4	2,09	1177	55	2,60	2939	20	2,78
172	295,4	2,20	603	103,3	2,06	1198	54	2,77	3064	19	2,91
182	279,5	2,00	618	101,6	2,69	1216	53	2,67	3189	18	2,90
192	265,9	2,20	631	99,4	2,67	1233	52	2,67	3329	17	3,01
202	257,6	2,50	643	97,8	2,15	1257	51	2,45	3475	16	3,02
212	248,0	2,36	659	95,6	2,17	1288	50	2,71	3649	15	3,00
222	240,8	2,30	675	93,4	2,40	1313	49	3,11	3857	14	2,93
232	233,6	2,30	693	91,4	2,75	1339	48	2,58	4088	13	3,10
242	225,7	1,80	709	89,3	2,46	1368	47	2,84	4344	12	3,03
254	219,2	2,31	724	87,6	2,18	1398	46	2,62	4626	11	2,99
266	209,2	1,90	742	85,7	2,74	1429	45	2,28	4955	10	3,06
276	198,3	2,60	757	84,0	2,60	1459	44	2,57	5353	9	3,10
287	193,5	2,42	769	83,0	2,50	1489	43	2,61	5846	8	3,10
298	189,3	3,27	786	81,4	2,29	1524	42	2,67	6456	7	3,13
310	183,0	2,31	807	79,6	2,43	1556	41	2,93	7202	6	3,15
323	177,7	1,77	825	78,0	2,63	1592	40	2,55	8166	5	3,26
335	170,9	2,27	839	77,0	2,20	1633	39	2,73	9526	4	3,25
346	165,0	2,10	849	76,0	2,50	1673	38	2,71	11561	3	3,29
356	161,3	2,20	859	75,0	2,82	1722	37	2,93	15139	2	3,46
366	156,3	2,45	870	74,0	2,18	1778	36	2,90	23765	1	3,72

ein Fehlschluß, der bei der Berechnung des Determinationskoeffizienten kaum möglich ist.

4.1. Abhängigkeit der Lexemlänge L von der Rangnummer r

Guiraud (1959) hat aus rein kombinatorischen Überlegungen, wo alle möglichen Buchstabenkombinationen aus den Buchstaben der jeweiligen Sprache bei der Bildung von Wörtern zugelassen wurden (die Buchstaben wurden als unabhängige Ereignisse angesehen, deren Auftrittswahrscheinlichkeit in jeder Buchstabenposition gleich sein sollte), folgende Abhängigkeit zwischen der Länge L der Wörter und deren Wahrscheinlichkeit p bzw. zwischen der Länge und deren Rang r abgeleitet:

$$L \sim \lg\ r \tag{8}$$

und

$$L \sim \lg\ p. \tag{9}$$

Die Modellannahmen von Guiraud (alle Buchstabenkombinationen sind möglich; die Buchstaben sind unabhängige Ereignisse mit derselben Häufigkeit in jeder Buchstabenposition) sind natürlich in realen Texten nicht erfüllt. Erstens sind nicht alle Buchstabenkombinationen möglich, zweitens sind die Buchstaben keine unabhängigen Ereignisse, und auch die Wahrscheinlichkeit der Buchstaben ist nicht gleich in allen Buchstabenpositionen.

Arapov (1988, 123) geht bei der Ableitung der Abhängigkeit zwischen der Länge L und der Rangnummer r von etwas anderen Annahmen aus:

a) Die Wahrscheinlichkeit p_i der Elemente eines Wortes (Silben) ist verschieden (diese p_i-Werte gelten aber für alle Positionen innerhalb des Wortes; auf jeder Position i kann ein Element E entweder mit der Wahrscheinlichkeit p_i auftreten oder mit der Wahrscheinlichkeit $q_i = 1 - p_i$ nicht auftreten; im letzteren Falle würde die entsprechende Position frei bleiben, was natürlich Einfluß auf die Länge des "Wortes" hätte). Es kann für jedes Wort,

welches aus n Elementen besteht, eine solche Ordnung der p_i-Werte (i=1,2,...,n) gefunden werden, so daß gilt:

$$p_i^{(k)} = 1/k \quad \text{mit} \quad \begin{matrix} k = 1, 2, \ldots, n, \\ i = 1, 2, \ldots, n. \end{matrix} \tag{10}$$

D.h., es gibt eine Wahrscheinlichkeit $p_i^{(1)} = 1$, eine Wahrscheinlichkeit $p_i^{(2)} = 0,5, p_i^{(3)} = 0,33$ usw.

b) Die Zahl n der Positionen in einem Wort (unabhängig davon, mit welcher Wahrscheinlichkeit p_i sie von einem Element E besetzt werden bzw. mit welcher Wahrscheinlichkeit q_i sie "frei" sind) hängt auf folgende Weise vom Rang des Wortes ab:

$$n = A \cdot r^b, \tag{11}$$

wo A und b Konstanten sind (nach Arapov gilt: b=1/3).

Aus Annahmne a) kann die Wahrscheinlichkeitsverteilung der Wortlängen (in Silben) als verallgemeinerte Binomialverteilung bestimmt werden. Die mittlere Wortlänge $\bar{L}$ der Wörter mit n Positionen kann dann nach

$$\begin{aligned} \bar{L} &= \sum_{i=1}^{n} p_i = \sum_{k=1}^{n} 1/k = \\ &= (1/n) + \ln\ n - (1/2n) - (1/12n2) + \ldots \\ &\approx \ln\ n \end{aligned} \tag{12}$$

bestimmt werden.

Setzt man nun für n den Ausdruck aus Gleichung (11) ein, so gilt:

$$L \approx \ln\ A + b \ln\ r = a + b \ln\ r. \tag{13}$$

Auch dieses Modell von Arapov gilt nur unter den genannten Annahmen. Annahme a), welche die quantitative Silbenstruktur eines Wortes approximativ beschreibt, ist plausibel, nicht aber Annahme b), weil sie von Arapov nicht begründet wird.

Gleichung (13) soll nicht nur für Silben gelten (für die sie ja abgeleitet wurde), sondern auch für Buchstaben. Man sollte die Konstanten

a und b, welche für die Längenmessung in Silben gelten, ganz einfach mit der mittleren Silbenlänge in Buchstaben multiplizieren, um die entsprechenden Konstanten für die Längenmessung in Buchstaben zu erhalten (Arapov 1988, 128ff.).

Die Überprüfung von Gleichung (13) für Probe Ia und Ib ergab folgende Resultate (die Zahlenwerte für die Konstanten wurden itarativ ermittelt):

Probe Ia: a = 0,6541, b = 0,9266, D = 0,58;
Probe Ib: a = 0,5760, b = 0,9464, D = 0,85;
Probe II: a = 0,4606, b = 0,3100, D = 0,77.

Wie diese Ergebnisse beweisen, führt eine Zusammenfassung der häufigsten und seltensten Lexeme (Probe Ib) zu einer wesentlich besseren Anpassung als bei einer Zusammenfassung nur der seltenen Lexeme. Die beste Anpassung wurde für Probe Ib festgestellt, d.h. für das Vokabular aus einer relativ kleinen Zahl von vollendeten Einzeltexten, eine etwas schlechtere Anpassung für Probe II, d.h. für das Vokabular einer relativ großen Zahl verschiedener Textfragmente. Aber auch für Probe Ib und Probe II ist der Anteil der erklärten Varianz noch relativ gering, so daß das Modell von Arapov durch die Berücksichtigung weiterer unabhängiger Variablen (z.B. der Polylexie, vgl. Hammerl 1989) ergänzt werden sollte.

Es fällt auch auf, daß für Probe II der von Arapov angegebene Wert für den Parameter b von b=0,33 fast erreicht wird. Das gilt aber nicht für die beiden anderen Proben. Dieses Problem muß in weiterführenden Untersuchungen analysiert werden, die aber über den Gegenstand dieser Arbeit hinausgehen.

Wir wollen hier noch kurz untersuchen, in welchem Maße sich die Ergebnisse der empirischen Untersuchungen für Probe Ia verändern würden, wenn nicht die Rangzählung von Arapov, sondern die von Hoffmann (1980) zugrunde gelegt wird (vgl. Kapitel 3.1 und Tabelle 1). Diese Untersuchungen ergaben folgende Resultate:

Probe Ia: a = - 0,0967, b = 1,3737, D = 0,46.

Eine solche Rangzählung führt zu wesentlich schlechteren Ergebnissen als die Rangzählung nach Arapov. Aus diesen Gründen haben wir in

den folgenden Untersuchungen auf die Anwendung dieser Rangzählung verzichtet.

4.2. Abhängigkeit der Lexemlänge L von der Frequenz F

Die Relation zwischen der Länge L und der Frequenz F war schon oft Gegenstand sprachwissenschaftlicher Untersuchungen (vgl. Zipf 1949, Guiter 1974, Herdan 1966). Wir wollen uns hier jedoch ausschließlich auf ein Modell von Köhler (1986, 69ff.) beschränken.

Köhler geht von der Hypothese aus, daß die relativen Veränderungen der Länge dL/L beim Übergang von einer Frequenz zu einer anderen, der relativen Frequenzänderung dF/F proportional ist, d.h.

$$\frac{dL}{L} = k\,\frac{dF}{F}, \tag{14}$$

wo k den Proportionalitätsfaktor darstellt. Die Lösung dieser Differentialgleichung ergibt

$$L = g \cdot F^k, \tag{15}$$

wo $g = e^A$ und A eine Integrationskonstante darstellt. Auf eine Interpretation dieser Parameter, die bei Köhler (1986, 53ff., 69ff.) nachgelesen werden kann, soll hier verzichtet werden. Unsere Untersuchungen zur Überprüfung dieses Modells ergaben folgende Ergebnisse:

Probe Ia: g = 8,3939, k = -0,2671, D = 0,58;
Probe Ib: g = 7,6588, k = -0,2160; D = 0,82;
Probe II: g = 4,0346; k = -0,1114; D = 0,72.

Die Werte für den Determinationskoeffizienten sind hier etwas niedriger als die entsprechenden Werte im Modell von Arapov (vgl. Kapitel 4.1.). Auch hier erbringt eine Zusammenfassung der häufigen und seltenen Lexeme bessere Resultate als eine Zusammenfassung nur der seltenen Lexeme. Der Determinationskoeffizient ist wiederum für das Vokabular von vollendeten Einzeltexten höher als für das Vokabular einer großen Zahl verschiedener Textfragmente. Da auch für dieses Modell der Anteil der erklärten Varianz der abhängigen Variablen relativ gering ist, sollte die Berücksichtigung zusätzlicher unabhängiger Variablen überprüft werden (vgl. Hammerl 1989).

Wenn die Modelle von Arapov und Köhler die durch sie beschriebenen Relationen mit hinreichender Genauigkeit beschreiben, so muß das auch für solche Modelle gelten, die aus diesen Modellen deduziert werden können. Dieses Problem wird in den Kapiteln 4.3 und 4.4 untersucht.

4.3. Überprüfung einer Deduktion aus dem Modell von Arapov

Nehmen wir an, es gilt das Zipfsche Gesetz in der Form

$$F = cr^d \tag{16}$$

und die von Arapov abgeleitete Beziehung

$$L = a + b \ln\ r. \tag{17}$$

Aus Gleichung (16) folgt

$$(\ln\ F - \ln\ c)/d = \ln\ r. \tag{18}$$

Setzt man diesen Ausdruck für ln r in Gleichung (17) ein, so erhält man:

$$\begin{aligned} L &= a + b((\ln\ F - \ln\ c)/d) = \\ &\frac{a \cdot d - b \ln\ c}{d} + \frac{b}{d} \ln\ F = \\ &e + f \ln\ F. \end{aligned} \tag{19}$$

Wir haben folgende Untersuchungen durchgeführt:

Zunächst haben wir für die Proben Ia und II die Konstanten c und d aus Gleichung (16) ermittelt (die Konstanten a und b sind ja schon bekannt, vgl. Kapitel 4.1), dann haben wir aus den Proben die "empirischen" Werte für die Konstanten e und f aus Gleichung (18) abgeschätzt, um diese danach mit den "theoretischen" Werten

$$\hat{e} = (ad - b \ln\ c)/d \tag{20}$$

und

$$\hat{f} = b/d \tag{21}$$

vergleichen zu können.

Wenn das Zipfsche Gesetz in der oben angegebenen Form gilt und das Modell von Arapov, so müssen auch die Differenzen zwischen den "empirischen" und "theoretischen" Werten für e und f relativ gering sein. Unsere Untersuchungen ergaben folgende Resultate (die Daten für a und b wurden aus Kapitel 4.1 übernommen):

Probe Ia: a = 0,6541, b = 0,9266;
c = 195,4532; d = -0,5761;
e = 7,6886, f = -1,2480; D = 0,63 ("empirische" Werte);

$\hat{e}$ = 9,1389, $\hat{f}$ = -1,6084 ("theoretische" Werte).

Probe II: a = 0,4606; b = 0,3100;
c = 72852,223; d = -1,1168;
e = 3,8405; f = -0,3061; D = 0,76 ("empirische" Werte);

$\hat{e}$ = 3,5684; $\hat{f}$ = -0,2776 ("theoretische" Werte).

Die durch das Modell (vgl. Gleichung (19) - (21)) vorausgesagten Parameterwerte stimmen relativ gut mit den entsprechenden "empirischen" Parameterwerten überein. Das bedeutet, daß auch Gleichung (19) die Abhängigkeit zwischen der Länge und Frequenz relativ gut beschreibt und in weiteren Untersuchungen zu überprüfen ist. Dieses Ergebnis sagt aber auch aus, daß auch für das Vokabular einer großen Zahl von Textfragmenten das Zipfsche Gesetz in seiner einfachsten Form (vgl. Gleichung (16)) anwendbar ist.

4.4. Überprüfung einer Deduktion aus dem Modell von Köhler

Analog den Ausführungen im vorangegangenen Kapitel gehen wir von der Gültigkeit des Zifschen Gesetzes (vgl. Gleichung (16) aus und von der Gültigkeit des Modells von Köhler:

$$L = gF^k. \tag{22}$$

Setzt man den Ausdruck für F aus Gleichung (16) in Gleichung (22) ein, so gilt:

$$L = g(cr^d)^k = sr^t, \tag{23}$$

wo

$$s = g \cdot c^k \tag{24}$$

und

$$t = d \cdot k. \tag{25}$$

Die Konstanten g und k sind aus Kapitel 4.2 bekannt, die Konstanten c und d haben wir aus den empirischen Daten für Probe Ia und Probe II abgeschätzt. Aus denselben Proben wurden dann die "empirischen" Werte für die Parameter s und t abgeschätzt und mit den nach den Gleichungen (24) und (25) berechneten "theoretischen" Werten verglichen.

Diese Untersuchungen lieferten folgende Ergebnisse:

Probe Ia: g = 8,3939; k = -0,2671;
c = 195,4532; d = -0,5761;
s = 1,7390; t = 0,2170; D = 0,56 ("empirische" Werte);

$\hat{s}$ = 2,0513; $\hat{t}$ = 0,15387 ("theoretische" Werte);

Probe II: g = 4,0346; k = -0,1114;
c = 72852,2230; d = -1,1168;
s = 1,1208; t = 0,1202; D = 0,74 ("empirische" Werte);

$\hat{s}$ = 1,1591; $\hat{t}$ = 0,1244 ("theoretische" Werte).

Auch hier stimmen die durch das Modell (Gleichung (23)–(25)) vorausgesagten "theoretischen" Parameterwerte relativ gut mit den entsprechenden "empirischen" Werten überein. Das bedeutet wiederum, daß Gleichung (23) die Abhängigkeit zwischen der Länge und Rangnummer relativ gut beschreibt und daß auch das Zipfsche Gesetz (Gleichung (16)) für die Beschreibung des Zusammenhanges zwischen der Frequenz und Rangnummer für das Vokabular einer großen Zahl verschiedener Textproben anwendbar ist.

Die Untersuchungen der Kapitel 4.1 bis 4.4 haben gezeigt, daß für die Beschreibung des Zusammenhanges zwischen der Rangnummer und der Länge die Modelle

$$L = a + b \ln r$$

und

$$L = s \cdot r^t$$

anwendbar sind. Für die Beschreibung des Zusammenhanges zwischen der Frequenz und Länge sind die Modelle

$$L = a \cdot F^b$$

und

$$L = e + f \ln F$$

anwendbar.

Diese Modelle gilt es in weiteren empirischen Untersuchungen zu überprüfen; es sind auch weitere theoretische Untersuchungen zur Begründung dieser Modelle notwendig.

4.5. Oszillation der Lexik

Als Oszillation der Lexik beschreibt Köhler (1986) die Streuung der empirischen Zahlenwerte für die mittlere Länge der Lexeme mit einer vorgegebenen Frequenz um die nach Gleichung (22) berechneten theoretischen Werte für die Länge. Köhler vermutet, daß es sich hierbei um "Schwingungen" handelt, die aus der Abhängigkeit der Kürzungsrate lexikalischer Einheiten von der Länge und Frequenz erklärt werden können.

In Hammerl (1989) haben wir ein lexikalisches Modell skizziert, welches mehrdimensionale Abhängigkeiten lexikalischer Einheiten beschreibt. Eine Variante dieses Modells betrifft z.B. die Abhängigkeit der Länge von der Frequenz und Polylexie, wobei eine Rückwirkung der Länge auf die Polylexie und Frequenz und von diesen Größen wieder auf die Länge möglich ist, was aber noch keine plausible Erklärung für die Oszillation der Lexik liefert. Wir vertreten bezüglich der "Oszillation der Lexik" die Auffassung, daß es sich um eine Streuung der empirischen Zahlenwerte um die theoretischen Werte handelt, so wie sie auch bei der Untersuchung anderer Relationen zwischen quantitativen Eigenschaften sprachlicher Einheiten auftritt, ohne diesen Abweichungen den Sonderstatus von "Schwingungen" einzuräumen. Diese Auffassung ist zumindest so lange berechtigt, bis diese "Schwingungen" nicht theoretisch begründet und empirisch als "Schwingungen" nachgewiesen wurden.

Wir wollen für Probe II unter Anwendung spezieller Tests überprüfen, ob die "Oszillation" als periodische Schwingungen nachweisbar ist, und zwar sowohl für die Länge in bezug auf die Frequenz (Modell von Köhler) als auch für die Länge in bezug auf die Rangnummer (Modell von Arapov). Für diese Untersuchungen haben wir den Phasenverteilungstest von Wallis und Moore und den Phasenhäufigkeitstest von Wallis und Moore (vgl. Krause, Metzler 1983, 381ff.) ausgewählt. Beide Test sind in der Lage, Periodizitäten der empirischen Werte einer Größe in bezug auf eine andere Größe nachzuweisen, wenn der Einfluß eines vorliegenden Trends eliminiert wird. Zwischen der Länge und Frequenz bzw. der Länge und Rangnummer besteht natürlich eine Abhängigkeit (vgl. Kapitel 3), die als Trend wirksam wird. Um trendbereinigte Daten L_t für die Länge zu erhalten, haben wir die Differenz aus den entsprechenden empirischen Werten für die Länge und den ent-

sprechenden theoretischen Werten (berechnet nach Gleichung (22) für die Länge in bezug auf die Frequenz bzw. nach Gleichung (17) für die Länge in bezug auf die Rangnummer) gebildet:

$$L_t(F) = L - gF^k, \tag{26}$$

$$L_t(r) = L - a + b \ln\ r. \tag{27}$$

4.5.1. Phasenverteilungstest

Wir erklären die Testmethode anhand der Werte $L_t(F)$ (für die Werte $L_t(r)$ gilt prinzipiell dasselbe Vorgehen).

Ausgangspunkt ist die geordnete Folge der Werte $L_t(F)$ für F=1,2,...,i,...n. Danach werden jeweils zwei benachbarte Werte für $L_t(F)$, d.h. $L_t(F = i)$ und $L_t(F = i + 1)$, hinsichtlich ihrer Größe verglichen und das Ergebnis dieses Vergleichs durch ein Plus- oder Minuszeichen festgehalten:

$$\text{Zeichen "+", falls } L_t(F = i) > L_t(F = i + 1), \tag{28a}$$

$$\text{Zeichen "-", falls } L_t(F = i) < L_t(F = i + 1). \tag{28b}$$

Falls $L_t(F = i) = L_t(F = i + 1)$, so wird $L_t(F = i + 1)$ nicht berücksichtigt und $L_t(F = i)$ mit $L_t(F = i + 2)$ verglichen usw.

Die Folge gleicher Vorzeichen wird als Phase bezeichnet, die Zahl dieser gleichen Vorzeichen als Länge d der Phase. Die erste und letzte Phase wird in den Untersuchungen nicht berücksichtigt, da nicht sicher ist, ob diese Phasen abgeschlossen sind.

Der Test läuft nun darauf hinaus, daß die Zahl n_d der Phasen mit einer Länge von d ermittelt wird. Diese empirischen Werte werden dann mit entsprechenden theoretischen Werten für $\hat{n}_d$ verglichen, die man unter der Annahme erhält, daß keine Periodizitäten der empirischen Werte vorliegen, daß also die Häufigkeit der Phasen zufällig verteilt ist. Diese theoretischen Werte für n_d werden nach

$$\hat{n}_d = \frac{2(d^2 + 3d + 1)(n - d - 2)}{(d + 3)!} \tag{29}$$

berechnet. Die Nullhypothese H_0, daß keine Periodizitäten vorliegen, wird dann abgelehnt, wenn der Wert für

$$\chi^2 = \sum_d \frac{(n_d - \hat{n}_d)^2}{\hat{n}_d} \tag{30}$$

größer als der entsprechende Testwert der Chi-Quadrat-Verteilung für ein vorgegebenes Signifikanzniveau bei 2 Freiheitsgraden ist.

Tabelle 4 führt die aus den empirischen Untersuchungen für die Länge in bezug auf die Frequenz ermittelten Zahlenwerte für Probe II auf und Tabelle 5 die entsprechenden Zahlenwerte für die Länge in bezug auf die Rangnummer. Wir haben hier bewußt nur Probe II berücksichtigt, weil hier relativ viele empirische Zahlenwerte aus einer fast 500 000 Wörter zählenden Stichprobe vorliegen, so daß eine mögliche "periodischen Schwingung" der Zahlenwerte für die Länge besonders deutlich werden sollte.

Tabelle 4
Länge und Häufigkeiten der Phasen (Probe II) - Länge in bezug auf die Frequenz

d	1	2	3	4	5	6	7	$\sum$
n_d	56	22	7	4	0	0	0	89
$\hat{n}_d$	57,08	24,93	7,13	1,54	0,28	0,04	(0,004)	

Tabelle 5
Länge und Häufigkeiten der Phasen (Probe II) - Länge in bezug auf die Rangnummer

d	1	2	3	4	5	6	7	$\sum$
n_d	52	26	5	4	0	0	0	87
$\hat{n}_d$	56,67	24,75	7,07	1,53	0,27	0,04	(0,005)	

Der nach Gleichung (30) berechnete Werte für χ^2 beträgt für die Länge in bezug auf die Frequenz $\chi^2 = 2,81$ und für die Länge in bezug auf die Rangnummer $\chi^2 = 3,57$ (für diesen Fall beträgt n=139, da zwei aufeinanderfolgende Werte für $L_t(r)$ dieselbe Größe hatten: in beiden

Fällen wurden die Werte von n_d für d=4,5,6 und 7 zusammengefaßt). Für ein Signifikanzniveau von 0,05 und für 2 Freiheitsgrade kann aus den Tabellen der Chi-Quadrat-Verteilung ein Wert von 5,99 abgelesen werden. Da χ^2 in beiden Fällen kleiner als dieser Wert ist, können keine Periodizitäten der Werte für die Länge von Lexemen in bezug auf deren Frequenz bzw. in bezug auf deren Rangnummer nachgewiesen werden.

4.5.2. Phasenhäufigkeitstest

Hier interessiert nicht die Verteilung der Häufigkeiten der Phasen, sondern lediglich deren Anzahl. Die aus den empirischen Untersuchungen ermittelte Zahl der Phasen n_s wird mit einer solchen Zahl $\hat{n}_s$ verglichen, die man bei einem zufälligen Auftreten der Phasen erhalten würde. Diese theoretische Zahl der Phasen $\hat{n}_s$ wird nach

$$\hat{n}_s = (2n - 1)/3 \tag{31}$$

ermittelt und die Streuung der Zahl der Phasen als

$$D^2 = (16n - 29)/90. \tag{32}$$

Die Nullhypothese H_0, daß keine Periodizitäten vorliegen, wird abgelehnt, wenn die Testgröße T mit

$$T = \frac{|n_s - \hat{n}_s|}{D^2} \tag{33}$$

größer als der für ein bestimmtes Signifikanzniveau bestimmte Grenzwert der standardisierten normalverteilten Variablen Z ist (für ein Signifikanzniveau von 0,05 beträgt Z=1,96).

Für die Untersuchung der Länge in bezug auf die Frequenz haben wir folgende Ergebnisse erhalten:

$$\hat{n}_s = 93; n_s = 89; D^2 = 24,567 \text{ und } T = 0,1628.$$

Die entsprechenden Ergebnisse für die Länge in bezug auf die Rangnummer (für n=139) lauten:

$$\hat{n}_s = 92,33; n_s = 87; D^2 = 24,389 \text{ und } T = 0,2185.$$

Beide Testgrößen liegen unterhalb des Tabellenwertes von 1,96 (bei einem Signifikanzniveau von α =0,05), so daß auch der Phasenhäufigkeitstest keine Periodizitäten nachweisen kann.

5. Zusammenfassung

Für die Beschreibung der Abhängigkeit zwischen der Länge L und der Frequenz F konnte das Modell von Köhler (1986) bestätigt werden, für die Abhängigkeit der Länge L von der Rangnummer r das Modell von Arapov (1988). Es wurde aber festgestellt, daß in beiden Modellen die zusätzliche Berücksichtigung weiterer Einflußfaktoren (z.B. der Polylexie) überprüft werden sollte.

Unter der Voraussetzung, daß das Zipfsche Gesetz die Abhängigkeit zwischen der Frequenz F und der Rangnummer r mit hinreichender Genauigkeit beschreibt (sowohl für das Vokabular von Einzeltexten als auch für das Vokabular einer größeren Zahl verschiedener Textfragmente), konnten zwei weitere Modelle deduziert werden, die ebenfalls die Abhängigkeit der Länge von der Frequenz und der Länge von der Rangnummer betreffen. Auch diese Modelle wurden in unseren empirischen Untersuchungen bestätigt. Da nun für die Beschreibung der Abhängigkeit der Länge von der Frequenz und der Länge von der Rangnummer je zwei Modelle vorliegen, die sich in unseren empirischen Untersuchungen als gleichwertig erwiesen haben, sollten beide Modelle in zukünftigen Untersuchungen einer tiefgründigen theoretischen Validierung und einer umfangreichen empirischen Validierung unterzogen werden. Solange aber solche Ergebnisse nicht vorliegen, sind prinzipiell beide Modelle für dieselbe Abhängigkeit anwendbar.

Für die Überprüfung dieser Modelle hat es sich als vorteilhaft erwiesen, sowohl die Lexeme mit großen Häufigkeiten zusammenzufassen (in Gruppen zu etwa 10 Lexemen) und auch die Lexeme mit kleineren Häufigkeiten (alle Lexeme mit derselben Häufigkeit wurden prinzipiell zusammengefaßt). Eine solche Zusammenfassung ist deshalb sinnvoll, da das Zipfsche Gesetz jeweils nur einige wenige Aspekte der quantitativen Textstruktur erfaßt und somit die Abhängigkeit zwischen der Frequenz und der Rangnummer nur approximativ (mit einem voraussehbaren Fehler im Bereich großer und kleiner Häufigkeiten) beschreiben kann.

Sowohl für die Länge in bezug auf die Frequenz als auch für die Länge in bezug auf die Rangnummer wurde die "Oszillation" der Lexik durch Anwendung von je zwei Tests überprüft, wobei keine Periodizitäten nachgewiesen werden konnten.

Literatur

Altmann, G. (1988), *Wiederholungen in Texten.* Bochum: Brockmeyer.

Arapov, M.V. (1988), *Kvantitativnaja lingvistika.* Moskva: Nauka.

Fleischer, M. (1988), *Frequenzlisten zur Lyrik von Mikolaj Sęp Szarzyński, Jan Jurkowski und Szymon Szymonowicz und das Problem der statistischen Autorschaftsanalyse.* München: Otto Sagner.

Guiter, H. (1974), "Les relationes frequence – longueur – sens des mots (langues romanes et anglais)". *XIV Congresso Internazionale di linguistica e filologia romanza.* Napoli, 15-20.

Hammerl, R. (1989), "Zum Aufbau eines dynamischen Lexikmodells – Mikro- und Makroprozesse der Lexik". *Glottometrika* 11, 19-40.

Herdan, G. (1966), *The advanced theory of language as choice and chance.* Berlin, Heidelberg, New York: Springer.

Hoffmann, L. (1980) (Hrsg.), *Fachwortschatz Mathematik. Häufigkeitswörterbuch.* Leipzig: Enzyklopädie.

Krause, P., Metzler, B. (1983), *Angewandte Statistik.* Berlin: Deutscher Verlag der Wissenschaften.

Köhler, R. (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Mandelbrot, B. (1953), "An information theory of the statistical structure of language". In: Jackson, W. (Hrsg.), *Communication theory.* New York, Academic Press, 503-512.

Mandelbrot, B. (1954a), "Structure formelle des textes et communication". *Word* 10, 1-27.

Mandelbrot, B. (1954b), "Simple games of strategy occuring in communication trough natural languages". *IRE Transactions, PGIT-3*, 124-137.

Orlov, Ju.K. (1982), "Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache- Rede" in der statistischen Linguistik)". In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.S. (Hrsg.), *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Rogalinska, A., Hammerl, R. (1990), "Über die Untersuchung mehrdimensionaler sprachlicher Relationen (mit Hilfe des Computerprogramms MULTIVAR)". In: Köhler, R. (Hrsg.), *Studies in language synergetics.* (erscheint)

Słownictwo współczesnego jezyka polskiego. Listy frekwencyjne. Tom I-V. Warszawa, Państwowe Wydawnictwo Naukowe, 1974-1977.

Zipf, G.K. (1949), *Human behaviour and the principle of least effort.* Cambridge, Mass.: Addison-Wessley.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 25–40

Differential equation models for the oscillation of the word length as a function of the frequency*)

P. Zörnig, R. Köhler, R. Brinkmöller
Bochum

Introduction

The fact that the length and the frequency of lexical items are interrelated is well known. Several different hypotheses have been formulated about the exact form of their interdependence (e.g. Zipf 1949, Guiter 1974), but it is doubtful whether they have been tested – at least in any satisfying way.

The two variables length and frequency belong to the lexical subsystem under investigation in Köhler's synergetic study (1986), in which a systems theoretical model of some of the properties of lexical items was constructed. The model consists of a system of differential equations, in which each equation represents the relation between a pair of variables. The first and simplest hypothesis concerning length and frequency is a differential equation of first order, the solution of which is the function

$$L(F) = aF^b. \tag{1}$$

The empirical test of this hypothesis by means of data from a German corpus[1)] (cf. Table 1, Fig. 1) yielded a highly positive result: $a = 10.9$, $b = -0.108$, test-criterion = 105 (with 1 and 158 degrees of freedom, $F_{0.01} = 6.8$) i.e. a probability of $2.2 \cdot 10^{-11}$ of getting this result by chance.

*) This study was written as a part of the project "Language Synergetics". The authors are grateful to the STIFTUNG VOLKSWAGENWERK for the kind support.

1) LIMAS-corpus, 500 texts, containing one million words (cf. Schaeder 1976).

A closer look at the deviations from the theoretical curve showed a surprising phenomenon (Köhler 1986:142) – the empirical data points formed an oscillating curve around the theoretical hyperbolic function line. Any attempt to equalize the curve by means of smoothing methods failed (Fig. 2). Therefore the oscillation could not be considered as a random effect. Another possible cause for the shape of the curve could be excluded: If the shortening method of a language were mainly the dropping of syllables, then a step-like discrete functional behavior could result, which would possibly appear in the shape of oscillation-like deviations from the theoretical expected curve. In this case, a monotone function should result as soon as the word length is measured by the number of syllables. Therefore Köhler computed the same data sample once more, this time measuring the length by the number of syllables. The oscillation, however, did not disappear.

It is the purpose of this paper to discuss different approaches to an explanation of the observed phenomenon by means of differential equations.

1. Linear differential equation models

1.1 The Basic Model

The first model we want to propose can be derived from the shortening mechanism (cf. Köhler 1986:144ff): If a lexical item is shortened as a consequence of increasing frequency, the shortening ratio will be the greater the longer the word was before. An expression which is already short will not be abbreviated as much as a long one, even if its frequency becomes very high. Words like *air*, *book*, or *fine* are not likely to be further shortened, whereas words such as *perambulator* and *radiotelephone* can be expected to shrink or be replaced by shorter ones, if their frequency of use increases. These words are in fact abbreviated to *pram* and *RT* respectively. (The German *Katalysator* [catalyst] has been replaced by *Kat* in the last few years, because the word has become highly relevant for all car owners). For this reason the momentary length was introduced into the equation as an additional independent

Tab. 1
The word length $L(F)$ in dependence on the frequency F (cf. Köhler 1986:175f)

	F	L		F	L		F	L
1	1.0	12.9497	31	31.0	9.0000	61	65.0	8.5000
2	2.0	12.4107	32	32.0	6.3333	62	66.0	8.0000
3	3.0	11.9375	33	33.0	8.6667	63	68.0	7.0000
4	4.0	10.0893	34	34.0	7.0000	64	69.0	4.0000
5	5.0	9.8140	35	35.0	7.3333	65	70.0	11.0000
6	6.0	10.3750	36	36.0	6.5000	66	71.0	6.0000
7	7.0	8.8095	37	37.0	5.5000	67	72.0	8.0000
8	8.0	9.4444	38	38.0	7.2500	68	73.0	6.5000
9	9.0	9.3077	39	39.0	4.5000	69	74.0	7.0000
10	10.0	6.6667	40	40.0	7.0000	70	75.0	10.0000
11	11.0	7.5625	41	41.0	7.5000	71	77.0	5.0000
12	12.0	7.0000	42	42.0	9.0000	72	79.0	8.6667
13	13.0	9.2222	43	43.0	8.0000	73	80.0	11.0000
14	14.0	6.1000	44	44.0	11.5000	74	82.0	5.0000
15	15.0	9.3333	45	46.0	8.5000	75	86.0	10.0000
16	16.0	8.2857	46	47.0	5.0000	76	87.0	11.0000
17	17.0	8.2000	47	48.0	7.8333	77	89.0	6.0000
18	18.0	8.0000	48	49.0	4.0000	78	91.0	10.0000
19	19.0	8.1429	49	51.0	5.0000	79	93.0	9.0000
20	20.0	8.5000	50	52.0	5.0000	80	94.0	6.0000
21	21.0	8.0000	51	53.0	7.0000	81	98.0	7.0000
22	22.0	7.0000	52	54.0	9.0000	82	99.0	5.0000
23	23.0	7.1667	53	55.0	5.5000	83	101.0	6.0000
24	24.0	7.7500	54	56.0	4.0000	84	104.0	6.0000
25	25.0	6.0000	55	57.0	4.0000	85	106.0	8.0000
26	26.0	7.6000	56	58.0	7.5000	86	109.0	5.0000
27	27.0	8.3333	57	59.0	5.5000	87	113.0	6.0000
28	28.0	7.3333	58	61.0	6.0000	88	115.0	7.0000
29	29.0	6.0000	59	62.0	8.6000	89	116.0	7.0000
30	30.0	8.6667	60	64.0	6.6667	90	120.0	6.0000

Tab. 1
Fortsetzung

	F	L		F	L		F	L
91	123.0	4.0000	121	205.0	7.0000	151	719.0	6.0000
92	124.0	5.0000	122	210.0	9.0000	152	762.0	5.0000
93	125.0	6.5000	123	219.0	8.0000	153	778.0	4.0000
94	130.0	6.0000	124	220.0	6.0000	154	791.0	4.0000
95	131.0	9.0000	125	221.0	4.0000	155	868.0	3.0000
96	132.0	4.5000	126	225.0	8.0000	156	1014.0	4.0000
97	133.0	6.0000	127	234.0	6.0000	157	1090.0	5.0000
98	134.0	7.0000	128	248.0	3.0000	158	1513.0	6.0000
99	137.0	5.0000	129	265.0	6.0000	159	1781.0	5.0000
100	141.0	7.0000	130	270.0	10.0000	160	1912.0	4.0000
101	142.0	10.0000	131	277.0	5.0000			
102	144.0	5.0000	132	287.0	8.0000			
103	152.0	7.0000	133	290.0	7.0000			
104	155.0	7.0000	134	314.0	12.0000			
105	159.0	10.5000	135	331.0	6.0000			
106	163.0	5.0000	136	342.0	10.0000			
107	167.0	5.0000	137	348.0	5.0000			
108	168.0	10.0000	138	353.0	8.0000			
109	171.0	6.0000	139	367.0	7.0000			
110	173.0	6.0000	140	368.0	5.0000			
111	175.0	7.0000	141	369.0	7.0000			
112	179.0	4.5000	142	390.0	6.0000			
113	182.0	7.0000	143	416.0	6.0000			
114	184.0	9.0000	144	427.0	6.0000			
115	187.0	6.0000	145	483.0	5.0000			
116	190.0	4.0000	146	548.0	6.0000			
117	192.0	11.0000	147	611.0	6.0000			
118	200.0	5.0000	148	657.0	5.0000			
119	201.0	5.0000	149	667.0	6.0000			
120	203.0	8.0000	150	685.0	5.0000			

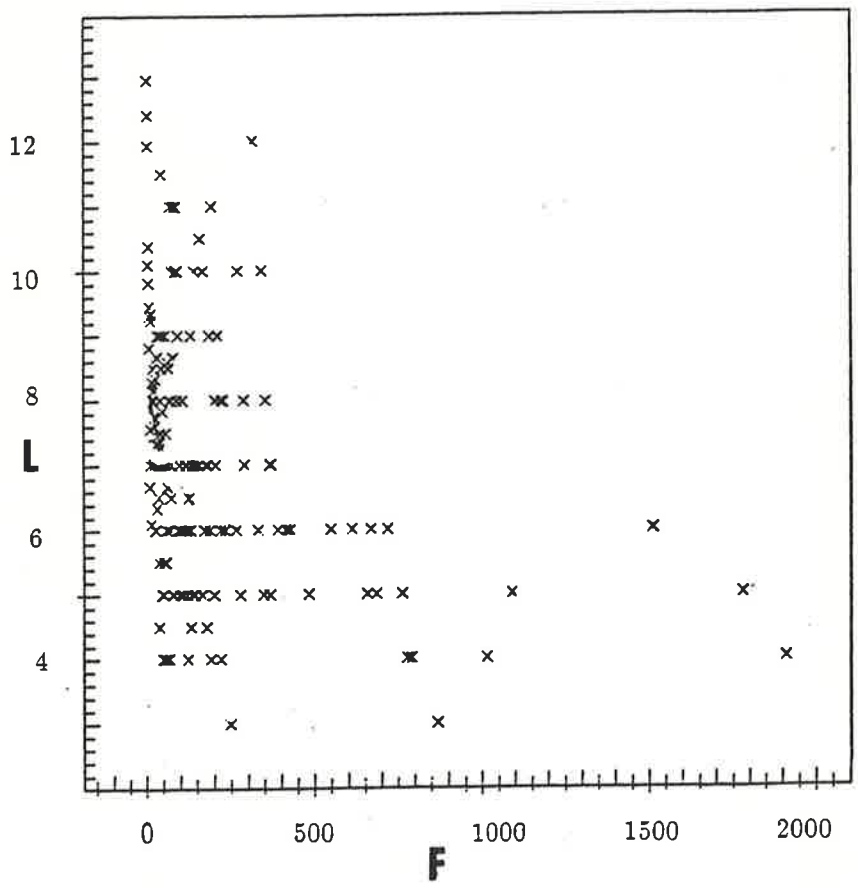


Fig. 1
The word length L as a function of the frequency F
(cf. Tab. 1 and Köhler 1986:138)

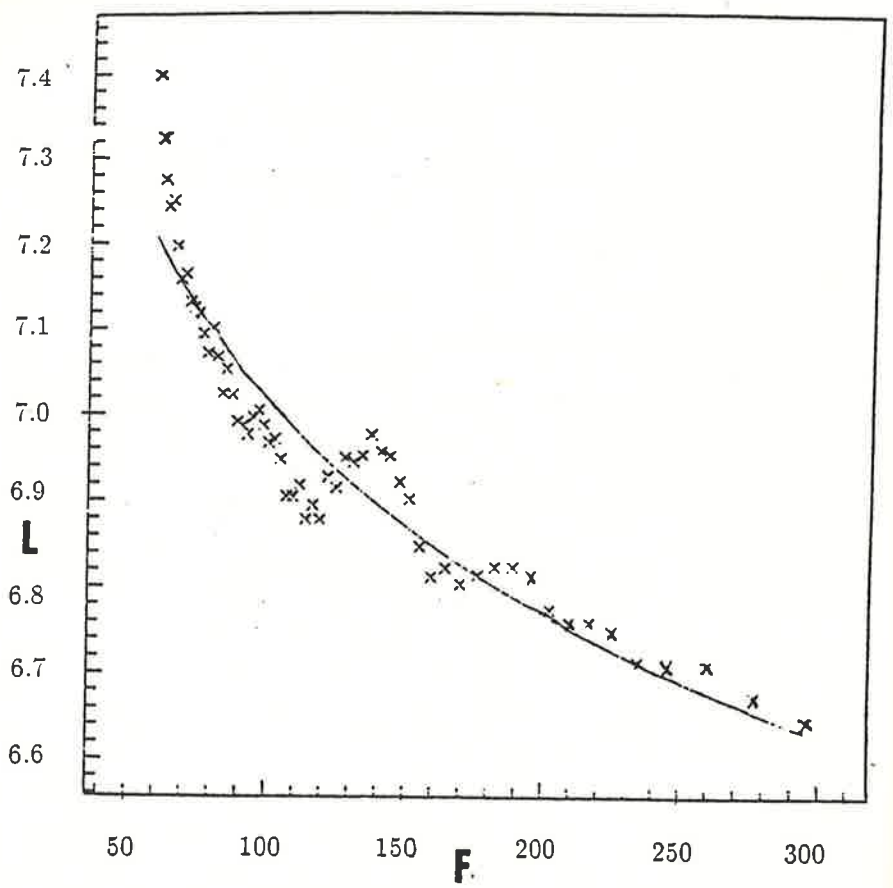


Fig. 2
The word length L as a function of the frequency F with smoothed data (interval length :100, cf. Köhler 1986:143)

variable. Thus, a generalization in form of a differential equation of order 2 was obtained (cf. Köhler 1986:146):

$$L'' + 2AL' + BL = f(F) \tag{2}$$

where A, B are constant with $A^2 < B$ and f is a monotone decreasing function.

1.2 Generalization of the differential equation

The following developments are based on some fundamental considerations of the properties of the solutions of (2). The general solution of (2) is given (cf. e.g. Walter 1972:134;139) by

$$L(F) = L_0(F) + e^{-AF}(\alpha \cos \sqrt{B - A^2}F + \beta \sin \sqrt{B - A^2}F), \tag{3}$$

where $L_0(F)$ denotes any special solution of (2). Using the addition formula of the sine function (cf. e.g. Erwe 1962:209) expression (3) can be transformed into:

$$L(F) = L_0(F) + e^{-AF}\frac{\beta}{\sin y}\sin(\sqrt{B - A^2}F + y) \tag{4}$$
$$(y := \operatorname{arccot}\frac{\alpha}{\beta}, \beta \neq 0).$$

Representation (4) shows that a solution of a differential equation (2) with constant coefficients oscillates with the *constant period length*[2)]

$$\frac{2\Pi}{\sqrt{B - A^2}}.$$

With regard to data sets for further languages it would be desirable if the model would allow variable period lengths. For this purpose we consider the coefficients in (2) as variable, i.e. we choose the generalized equation

[2)] We assume that $L_0(F)$ is not an oscillating function.

$$L'' + A(F)L' + B(F)L = f(F) \tag{5}$$

as a model for the dependence of the word length on the frequency.

1.3 Solving equation (5)

In order to determine a special solution of (5) we assume that the functions A, B and f can be represented in the following form:

$$A(F) = -\frac{g''(F)}{g'(F)} \tag{6}$$

$$B(F) = (g'(F))^2 \tag{7}$$

$$\left.\begin{aligned} f(F) &= ab(b-1)F^{b-2} + abA(F)F^{b-1} + aB(F)F^b \\ &= ab(b-1)F^{b-2} - ab\frac{g''(F)}{g'(F)}F^{b-1} + a(g'(F))^2F^b. \end{aligned}\right\} \tag{8}$$

Now a solution is given as follows:

Proposition 1: The function

$$L(F) = aF^b + c\sin(g(F)) \tag{9}$$

with a monotone, two times differentiable function g is a solution of equation (5) when $A(F)$, $B(F)$, and $f(F)$ are given by (6) - (8).
Proof: The first and second derivative of (9) are given by:

$$\begin{aligned} L(F) &= aF^b + c\sin(g(F))) \\ L'(F) &= abF^{b-1} + cg'(F)\cos(g(F)) \\ L''(F) &= ab(b-1)F^{b-2} + cg''(F)\cos(g(F)) - c(g'(F))^2\sin(g(F)). \end{aligned}$$

This implies using (6) – (8):

$$\begin{aligned} &L'' + A(F)L' + B(F)L \\ &= ab(b-1)F^{b-2} + cg''(F)\cos(g(F)) - c(g'(F))^2\sin(g(F)) \\ &\quad - \frac{g''(F)}{g'(F)}abF^{b-1} - cg''(F)\cos(g(F)) \\ &\quad + (g'(F))^2aF^b + (g'(F))^2c\sin(g(F)) \\ &= ab(b-1)F^{b-2} - \frac{g''(F)}{g'(F)}abF^{b-1} + (g'(F))^2aF^b \\ &= f(F), \end{aligned}$$

i.e. (9) solves the differential equation (5).•

Proposition 1 yields a whole system of solutions of (5), namely all functions of the form (9). All solutions have a variable period length which is determined by the function g in (9).

1.4 Fitting of function (9) to the German data

In order to determine a fitting of function (9) to the smoothed data (cf. fig. 2) we assume for the moment that g has the form

$$g(F) = \alpha F^\beta + \gamma. \tag{10}$$

Using the optimization procedure of Nelder/Mead (1965) the following optimal parameters

$a = 8.9089$	$b = -0.0523$	$c = -0.0610$
$\alpha = 0.0039$	$\beta = 1.5418$	$\gamma = 2.9545$

were found for (9) with (10). The sum of squared deviations is 0.1013. Fig. 3 shows the theoretical curve.

Choosing another function g or using better starting parameters, the fitting could probably be improved. Since the model has to be tested for further languages and with regard to the model in section 2 we will not attempt to find a better fitting for the time being.

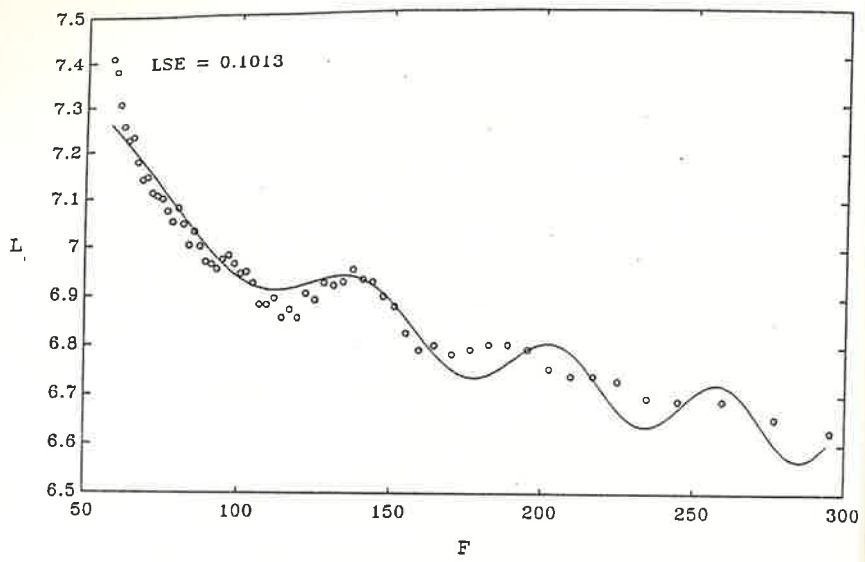


Fig. 3
Fitting of function (9) (with (10)) to the "smoothed" data (interval length: 100, cf. Fig. 2)

1.5 Interpretation of the model

Choosing g as in (10), the functions in (6) – (8) are given as follows:

$$A(F) = -\frac{g''(F)}{g'(F)} = \frac{1-\beta}{F} \tag{11}$$

$$B(F) = (g'(F))^2 = (\alpha\beta)^2 F^{2(\beta-1)} \tag{12}$$

$$\left.\begin{aligned} f(F) \quad &= ab(b-1)F^{b-2} + ab\frac{1-\beta}{F}F^{b-1} \\ &\quad + a(\alpha\beta)^2 F^{2(\beta-1)}F^{b} \\ &= ab(b-\beta)F^{b-2} + a(\alpha\beta)^2 F^{2\beta+b-2}. \end{aligned}\right\} \tag{13}$$

According to proposition 1 the function (9) (with (10)) is a solution of the following differential equation (cf. (5)):

$$\left.\begin{aligned} L'' + \frac{1-\beta}{F}L' + (\alpha\beta)^2 F^{2(\beta-1)}L = \\ = ab(b-\beta)F^{b-2} + a(\alpha\beta)^2 F^{2\beta+b-2}. \end{aligned}\right\} \tag{14}$$

Inserting the optimal parameters in (14) yields

$$\left.\begin{aligned} L'' - 0.5418F^{-1}L' + 0.000036F^{1.0836}L \\ = 0.7427F^{-2.0523} + 0.0003F^{1.0313} \end{aligned}\right\} \tag{15}$$

which can be interpreted as follows:

As is well known, the shortening rate is not constant. Its change L'' depends on the actual length L, the actual shortening rate L' and the frequency F (cf. Köhler 1986:146).

2. A differential-difference equation model

2.1 Construction of an "optimal" function $L(F)$

The following approach is based on the idea of constructing a function $L(F)$ (depending on parameters), which is "optimally suited" for fitting the data points in Fig. 2. We will postpone the consideration of which differential equation is solved by $L(F)$ to section 2.2 below.

Various tests have shown that the data points in Fig. 2 can be better fitted when the approach

$$L(F) = aF^b + a'F^{b'} \tag{16}$$

is used instead of (1).

This corresponds to the assumption that two processes of the same kind but with different parameters are part of the mechanism underlying the function (16). Possibly speaker and hearer differ in their sensitivity with respect to changes in frequency: The fact that a word has increased in frequency of use is in the foreground of a speaker's consciousness, while a hearer is possibly not as much aware of the change.

As optimal parameters for (16) we have found

$$a = 8.3803 \qquad b = -0.0404$$
$$a' = 3.6368 \cdot 10^{12} \qquad b' = -7.4092.$$

The corresponding theoretical curve is presented in Fig. 4.

The sum of squared deviations is 0.0602. Though function (16) does not oscillate, the fitting is already better than in the case of function (9) with (10) (cf. Fig. 3). Fig. 4 shows that the fitting can be improved again, when a sine function with an appropriate frequency-depending damping factor is superimposed to the "basic function" (16). The damping must be minimal for $F = F_0 \approx 140$ and get stronger for increasing distance $|F - F_0|$. Assuming a normally distributed damping we took the following approach:

$$L(F) = aF^b + a'F^{b'} + ce^{d(F-F_0)^2}\sin(\alpha F). \qquad (17)$$

As optimal parameters for (17) we found:

$a = 8.3987$	$b = -0.0409$	$a' = 2.9401 \cdot 10^{12}$
$b' = -7.3585$	$c = -0.0884$	$d = -0.0014$
$F_0 = 136.4733$	$\alpha = 0.1242.$	

The sum of squared deviations (LSE) is now 0.0202, i.e. by taking the oscillation into account, LSE diminishes by a factor of approximately $\frac{1}{3}$ compared with (16). Fig. 5 shows that the "description" of the data set can hardly be improved without introducing further parameters (cf. (17)).

2.2 Derivation of a differential-difference equation

A representation of the derivative $L'(F)$ of the function in (17) by a term with $L(F)$ is possible by use of a differential-difference equation (cf. Bellman/Cooke 1963). The following proposition represents the result.

Proposition 2: The function $L(F)$ in (17) solves the differential-difference equation

$$L'(F) = A(F) + B(F)L(F) + C(F)L(F + \frac{\Pi}{2a}), \qquad (18)$$

where $A(F)$, $B(F)$, $C(F)$ are defined as follows:

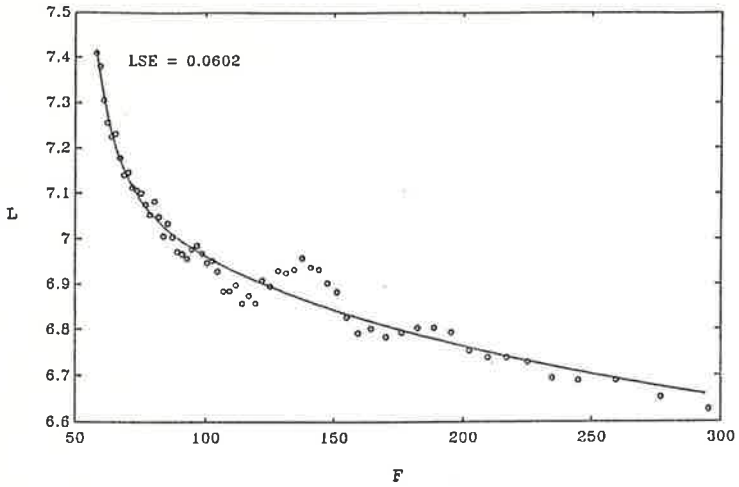


Fig. 4
Fitting of function (16) to the "smoothed" data (interval length: 100)

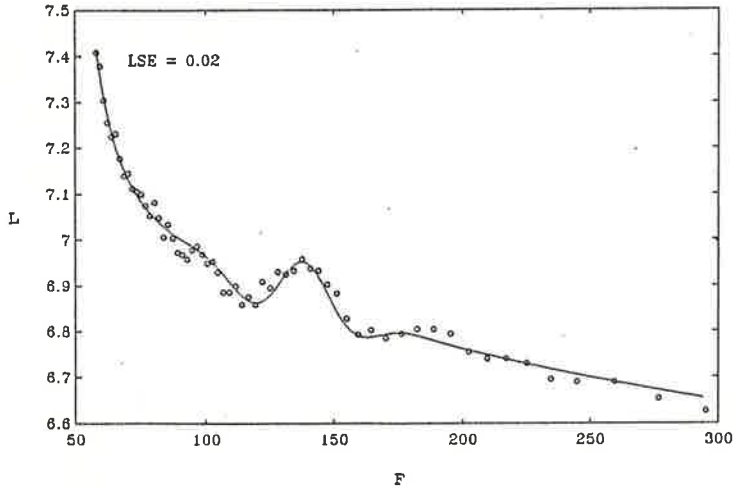


Fig. 5
Fitting of function (17) to the "smoothed" data (cf. Fig. 4)

$$A(F) = L_0'(F) - B(F)L_0(F) - C(F)L_0(F + \frac{\Pi}{2a}), \tag{19}$$

with

$$L_0(F) = aF^b + a'F^{b'}, \tag{20}$$

$$B(F) = 2d(F - F_0), \tag{21}$$

$$C(F) = \alpha e^{d\Pi/\alpha}(F_0 - F - \frac{\Pi}{4\alpha}). \tag{22}$$

Proof: Equation (17) implies

$$\left.\begin{aligned} L'(F) \quad &= abF^{b-1} + a'b'F^{b'-1} \\ &+c \cdot 2d(F - F_0)e^{d(F-F_0)^2}\sin(\alpha F) \\ &+ce^{d(F-F_0)^2}\alpha\cos(\alpha F). \end{aligned}\right\} \tag{23}$$

In the following we represent the second and third term of the sum in (23) by terms of $L(F)$.

Obviously it holds

$$\begin{aligned} &c \cdot 2d(F - F_0)e^{d(F-F_0)^2}\sin(\alpha F) = \\ &2d(F - F_0)(L(F) - L_0(F)) \end{aligned} \tag{24}$$

From the well-known relation

$$\cos x = \sin(x + \frac{\Pi}{2})$$

it follows that

$$\cos(\alpha F) = \sin(\alpha(F + \frac{\Pi}{2\alpha})) \quad (\alpha \neq 0). \tag{25}$$

Moreover it follows from (17)

$$\begin{aligned} &L(F + \frac{\Pi}{2\alpha}) - L_0(F + \frac{\Pi}{2a}) \\ &= ce^{d(F+\frac{\Pi}{2\alpha}-F_0)^2}\sin(\alpha(F + \frac{\Pi}{2\alpha})) \\ &= ce^{d(F-F_0)^2}e^{\frac{d\Pi}{\alpha}(F-F_0+\frac{\Pi}{4\alpha})}\sin(\alpha(F + \frac{\Pi}{2a})). \end{aligned}$$

Using (25), this implies

$$\begin{aligned} &L(F + \frac{\Pi}{2a}) - L_0(F + \frac{\Pi}{2a}) = \\ &ce^{d(F-F_0)^2}e^{\frac{d\Pi}{\alpha}(F-F_0+\frac{\Pi}{4a})}\cos(\alpha F), \end{aligned}$$

which finally leads to

$$\begin{aligned} &ce^{d(F-F_0)^2}\alpha\cos(\alpha F) = \\ &(L(F + \frac{\Pi}{2\alpha}) - L_0(F + \frac{\Pi}{2\alpha}))\alpha e^{\frac{d\Pi}{\alpha}}(F_0 - F - \frac{\Pi}{4\alpha}). \end{aligned} \tag{26}$$

By inserting the right sides of (24) and (26) in (23) we obtain a first representation of $L'(F)$ using by terms of $L(F)$:

$$\begin{aligned} L'(F) &= abF^{b-1} + a'b'F^{b'-1} \\ &+ 2d(F - F_0)(L(F) - L_0(F)) \\ &+ \alpha e^{\frac{d\Pi}{\alpha}}(F_0 - F - \frac{\Pi}{4\alpha})(L(F + \frac{\Pi}{2\alpha}) - L_0(F + \frac{\Pi}{2\alpha})). \end{aligned}$$

Using equations (19) – (22) this implies

$$\begin{aligned} L'(F) &= L_0'(F) + B(F)(L(F) - L_0(F)) \\ &+ C(F)(L(F + \frac{\Pi}{2\alpha}) - L_0(F + \frac{\Pi}{2\alpha})) \\ &= L_0'(F) - B(F)L_0(F) - C(F)L_0(F + \frac{\Pi}{2\alpha}) \\ &+ B(F)L(F) + C(F)L(F + \frac{\Pi}{2\alpha}) \\ &= A(F) + B(F)L(F) + C(F)L(F + \frac{\Pi}{2\alpha}), \end{aligned}$$

i.e. the assertion is proved. •

The question of the linguistic interpretation of (18) will be the subject of further research.

Literature

Bellman, R. & Cooke, K.L. (1963), *Differential-Difference Equations.* London, New York: Academic Press.

Erwe, F. (1962), *Differential- und Integralrechnung, Bd. 1.* Bibliographisches Institut. Mannheim, Wien, Zürich.

Guiter, H. (1974), "Les rélations fréquence – longueur – sens des mots (langues romanes et anglais)" *XIV Congresso Internazionale di linguistica e filologia romanza.* Napoli, 15-20.

Köhler, R. (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1987), "Systems Theoretical Linguistics". In: *Theoretical Linguistics* 14, 213, 241-257.

Nelder, J.A. & Mead, R. (1965), "A simplex method for function minimization". *Computer Journal* 7, 308-313; 8, 27.

Schaeder, B. (1976), "Maschinenlesbare Textkorpora des Deutschen und des Englischen". *Deutsche Sprache* 4, 356-369.

Walter, W. (1976), *Gewöhnliche Differentialgleichungen. Eine Einführung.* 2. Auflage. Berlin, Heidelberg, New York: Springer-Verlag.

Zipf, G.K. (1949), *Human Behavior and the Principle of Least Effort.* Massachusetts: Reading.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 41–60

Zipf-Mandelbrot's Law in a Coherent Text: Towards the Problem of Validity

M.G. Boroda
Tbilisi/Bochum
P. Zörnig
Bochum

Among the problems of study of a coherent text, the problem of quantitative regularities underlying the repetition-organization on its various levels is one of the most urgent (see Altmann 1988, Arapov 1988, Tuldava 1987, a.o.). Being oriented at first towards the study of the *samples*, these investigations, beginning from approximately 1970, have increasingly often been concerned with *complete texts*, sometimes opposing them – at least, hypothetically – to text fragments (see e.g. Arapov 1988, Orlov 1970, 1976, Orlov, Boroda, Nadarejšvili 1982, a.o.).

Beginning in the middle of the 70s, investigations in this direction have taken into consideration not only texts in natural language (in particular, fictional literary texts) but also musical compositions (Boroda 1979, 1982, 1990, Boroda, Polikarpov 1988) and even paintings (Orlov, Vološin 1982).

From its very beginning, even at the level of analysis of lexical samples, this investigational direction has found itself dealing with what is known as Zipf's Law

$$F_i = \frac{K}{i^c}, \quad i = 1, 2, \ldots, V \qquad (1),$$

where V is the *vocabulary size* of the text (sample) – i.e. the number of different units under study in it –, F_i, $i = 1, 2, \ldots, V$, are their *frequencies regulated in a decreasing* (not strictly) *order*, so that $F_i \leq F_{i+1}$, and K and c are certain parameters. As is seen from (1), F_i is a hyperbolic function of i.

Further generalizations of the law – in particular, that suggested by B. Mandelbrot

$$F_i = \frac{K}{(B+i)^c} \qquad (2)$$

(who deduced it from the concept of optimization of the information coding in text), as well as the attempts to construct purely statistical models of the frequency structure of the text/sample – see e.g. Kalinin (1965), and finally – the studies of the "qualitative meaning" of the Zipf-Mandelbrot law and the relations of its parameters with the text length (= the number of all tokens in it), the size of its vocabulary, the frequency of the most frequent word in it, etc. (see Altmann 1988, Arapov 1988, Orlov 1976, 1982a, 1982b, 1988, Krylov 1987, Tuldava 1987) made it possible to reveal certain general principles of the repetition-organization of words in the text.[1)]

In particular, the investigations based on the model of the Zipf-Mandelbrot law suggested by Orlov (1976) where the parameters K and B were hypothesized to be dependent on the text length L and the frequency $F_{\max}$ of the most frequent word in it,

$$K = 1/\ln F_{\max}, \quad B = KL/F_{\max} = L/(F_{\max} \ln F_{\max}) \qquad (3),$$

made it possible to show that in a complete coherent text – especially, in an artistic one – the relations between the size of the groups of frequent

[1)] Let us note here that the structure of repetitions in text can be represented in two different forms – specifically, as a table

Frequency F	The number V_F of units each of which occurs F times in the text

– e.g.

1	V_1
2	V_2
.	.
.	.
.	.
$F_{\max}$	$V_{F_{\max}}$

and a table

and rare words, those between the group of words not repeated in a given text and the whole group of repeated ones, etc. are submitted to regularities very general by nature.

In particular, if the actual frequencies of words in a literary text were organized in diminishing order, it was possible to describe them by the equation (2), with $c = 1$ and K and B defined by (3), with a visible accuracy (see Fig. 1).

It was also possible, within the framework of the model of the Zipf-Mandelbrot law mentioned above, to predict the vocabulary size V of the text (as a function of its length L and the frequency $F_{\max}$ of the most frequent word in it, $V^* = L \cdot (F_{\max} - 1)/(F_{\max} \cdot \ln F_{\max}) = L/\ln F_{\max}$) with an average accuracy of 5-15%. The study also demonstrated general principles in respect to relations between the frequency F and the number V_F of different words in the text – specifically, the hyperbolic relations between F and V_F for relatively small values of F. The relations in the repetition-organization mentioned above revealed themselves with a similar accuracy in various coherent texts (see Orlov 1976, 1982 a.o.).

A number of models of Zipf-Mandelbrot's Law suggested in recent

Frequency F_i	Its number (rank) i in a set of frequencies (regulated in diminishing order) of all different units in the text

– e.g.

$F_{\max}$	1
.	2
.	3
.	.
.	.
1	V

The equations (1), (2) discussed above, as well as similar ones, are used to describe the rank-frequency relations in the text under study; it is this representation of the repetition structure that is traditionally connected with Zipf-Mandelbrot's Law. The relations between the frequency F and the number V_F – i.e. the other representation of the same structure of repetitions – could be described as corollarly of the law. See below in the main text.

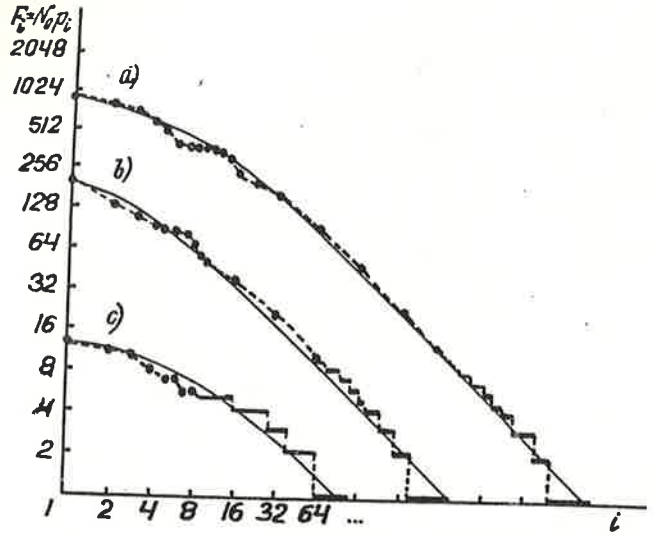


Figure 1.
The frequency structures of literary texts submitted to the Zipf-Mandelbrot law (1) with $c = 1$ and $Z = N_0$.
— theoretical curves (according to (1)).
.–.– actual values of the frequencies
a) Š. Rustaveli, The Knight in the Panther's Skin (in Georgian);
b) A. S. Puškin, Skazka o rybake i rybke;
c) Bylina "Avdotja Rjazanočka" (see Orlov 1975).

Fig. 1

years – e.g., the "*model of the frequency structure of the vocabulary*" (Orlov 1982c), a "*stationary model of the coherent text generation*" (Krylov 1987), a model where the Zipf-like relations in the text and vocabulary are considered to be a result of the "*principle of the maximal dissymmetry of a system*" (Arapov 1988), or that which connects Zipf's Law with the processes of *neuron activation* (Lebedev 1983, 1986) and some others, led to the multilevel, quantitative-systemic study of the repetitions-organization in a coherent text (see e.g. Tuldava 1987). On the other hand, it also made it reasonable to study the relations between the frequency and the length of a unit, or those between the

length of a given unit and that of its neighboring units, etc. which, in its turn, led to revealing autoregulation, synergetic processes in text and language influencing both text generation and the development of language (see Altmann 1988, Köhler 1986, a.o.).

Recent investigations of the structural units in music – specifically, the definition of the basic units of musical language allowing one to derive algorithmically defined segmentation of the musical composition on its melodic level on fragments of different length (Boroda 1973, 1986, 1988, Detlovs 1984, 1988) – made it possible to study the repetition-organization in the whole musical piece, from its beginning to end. Investigations (Boroda 1979, 1982 a.o.) carried out on the material of the musical compositions of the 18th to the 20th centuries on the level of the elementary melodic unit "Formal motif" (F-motif)

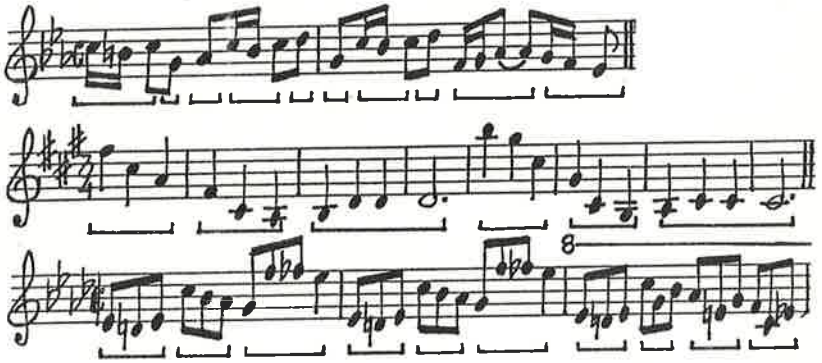
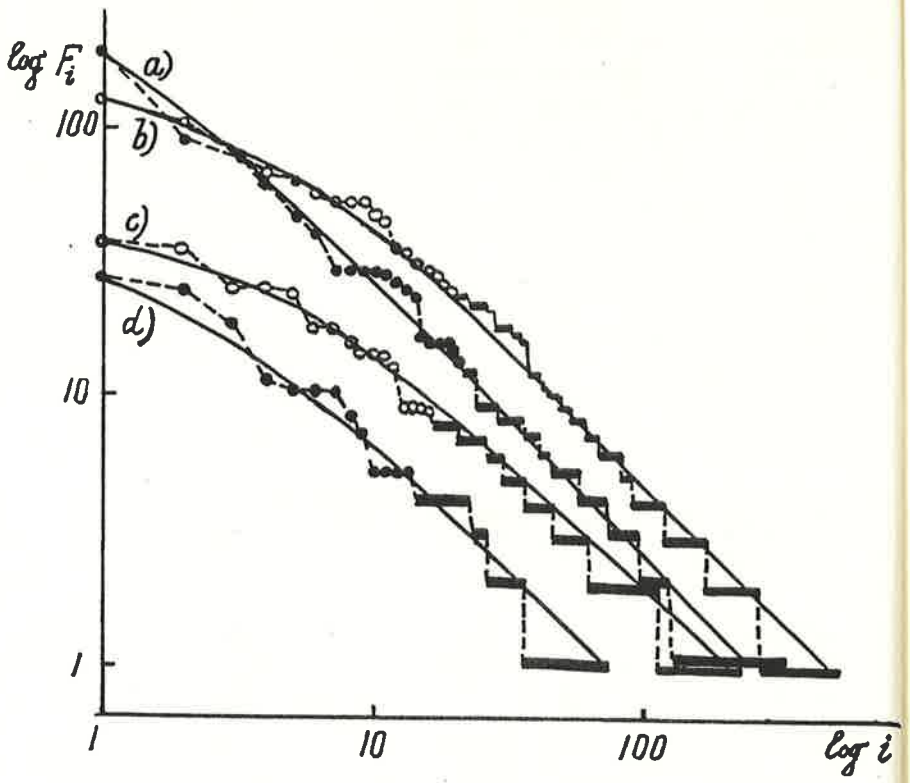


Fig. 2

showed that Zipf-Mandelbrot's Law (2) could be relevant to music, too – in the sense that the frequencies of the F-motifs arranged in diminishing order could be described by the theoretical frequency curve built up according to (2), with the same functional dependencies of the parameters K, B and $c(K = 1/\ln F_{\max}, B = KL/F_{\max}, c = 1)$ as for literary texts (see Fig. 3).

Further investigations showed that the law can be observed not only on the melodic (pitch + rhythmic) level of the musical composition but also in the organization of the *melodic variation of the rhythmical struc-*



a) J.S. Bach. Prelude and Fugue, WTC, I, No 22;
b) F. Chopin. Sonata No 3;
c) L.v. Beethoven. Rondo op. 51 No 1 C-dur;
d) D. Scarlatti. Sonata No 1 (Keller & Weismann, vol. 1). , , etc. denote the actual frequences of the F-motifs; the "theoretical frequency curve" is given by .

Fig. 3

ture of the F-motifs, i.e. the study of Zipf-Mandelbrot's Law revealed that certain general principles of repetition act in musical text on two (or, more correctly, on at least two) levels simultaneously. Similar effects were observable in literary (artistic) texts, too, which made it possible to advance a hypothesis that in the process of text generation

there exists for any given text certain central level of the information-coding which is connected with the genre of the text, or more correctly, with the "*rank of its largeness*" (Boroda, Polikarpov 1988).

Further investigations of Zipf-Mandelbrot's Law in music on the level of motivic units mr-segments (more large than the F-motif) made it possible to reveal some trends in the development of the European musical language in 17-20th centuries (e.g., the *growing role of the "suffixalization-like" rhythmical variation of a motif* where a tone is added to it or subtracted from it, so that it is interpreted as a kernel, or the *slow decrease of the motivic length in time*, etc.) and to observe there autoregulation-processes (in particular, based on conflicting relations between rhythmic and pitch/intervalic/mode) (see Boroda 1990).[2)]

Finally, the first steps made in studying the relations between the "weights" of various colors in a work of painting has shown that Zipf-Mandelbrot's Law, in one of its modifications, could describe these relations, too (Volosin, Orlov 1982, Orlov 1988); this made it possible to link up the revealed effects with the observations on the central and peripheric vision by a human being.

Already these short considerations of Zipf-Mandelbrot's Law show its importance in the study of general regularities of the human communication, as far as it reveals itself in text-generation. Moreover, the studies of the law, the attempts to obtain it from certain general assumptions make it more and more clear that it has much to do with the general laws underlying the organization of any large system where some or other "optimization" is the first-rank task.

In this situation, the question of the validity-test for Zipf-Mandel-

2) In the interest of not overloading the text, we shall not provide any strict definitions of the F-motif or the mr-segment (the latter especially would have to be very extensive). For such definitions cf. Boroda (1982, 1988). Let us only note that

a) the formation of F-motifs in melody is based on the tendency of a tone to be chained with either the previous metrically stronger tone or the following longer in duration tone, and

b) a given melody with a bar structure can be segmentated into F-motifs by the procedure of looking through the melody note-by-note according to the following rule: a note $\mathbf{x}$ is considered the final of the F-motif A iff. $\mathbf{x}$ is: 1) longer in duration than the following note $\mathbf{x}_{+1}$ or 2) equal in duration to $\mathbf{x}_{+1}$ but metrically weaker than it or longer in duration than the previous note $\mathbf{x}$.

brot's Law, a test making it possible to say, on the basis of certain statistical criteria, whether a given text/system is indeed submitted to the law, or, perhaps, which of two texts/systems under study follows the law more accurately, is of primary importance. For coherent texts the validity-test question is specially important: as the investigations show, the complete texts follow it with more visible accuracy than text-fragments (Orlov 1975, 1982, 1988, Arapov 1988, Boroda 1979, 1982). However, it is this question that proves the most difficult in the Zipf-Mandelbrot's Law studies.

Indeed, within the framework of the above-mentioned models of the law where the connection of its parameters K, B and c with the basic quantitative characteristics of text (NB: a *given concrete text*!) is at least postulated, it is hardly possible to apply methods of testing statistical hypotheses to prove whether a given text indeed follows Zipf-Mandelbrot's Law. The point is that the basic postulate for such methods – namely, the idea that a given text is a *subsample of certain general sample* (e.g., that of all texts of a given author) and the frequencies of words in it are *realizations of their probabilities in the general sample* (so that the larger the subsample is taken, the better the frequencies and "probabilities" would coincide with each other) – this idea is, within the framework of such models, not valid. As the investigations show, with an increase of the lexical sample in size the frequencies of words do not approach their hypothetical probabilities in language. The same can be observed in music, even on the level of the melodic intervals, not to speak of the level of motifs.

Thus, to put it succinctly, the models discussed here – as well as any other model where the parameters of the law are functionally connected with the text length, the size of its vocabulary or other quantitative characteristics of a concrete, specific text – face one and the same difficulty: the visual estimations of the coincidence between the actual and the predicted frequencies are insufficient (especially as one often has to decide whether the coincidence in the text A is better than that in B), while the methods of proving statistical hypotheses can not be applied, since their basic ideas contradict those of the respective model. The compromise approach, where the frequencies F_i with small ranks – i.e. those of units frequently occurring in the text – are described by one or other form of the law (2) and for units rarely occurring in the

text, the relations between F and V_F are considered as an object for description (see above Table 1, footnote 1) – an approach, used, as a matter of fact, in a number of papers concerned with Zipf-Mandelbrot's Law in coherent text (see e.g. Orlov 1982, Krylov 1987, a.o.)[3]. does not solve the problem. On the contrary, it may be supposed that the division of the structure of repetitions into two parts and their description from two different points of view, with two equations, renders the validity-question even more complicated.

In such a situation, it seems reasonable to consider an alternative approach to solving the validity-test problem, specifically, the following:

a) to consider the frequencies of the vocabulary-units in the text or the numbers V_F of different units, each of which occurs F times in it, as certain experimental data which should be approximated by the function (2) or some other function connected with it, then – to estimate the values of its parameters K, B and c by minimizing the general divergence between the data and their approximation, and finally – to study the relationships between the parameters with each other, as well as with the text length and other quantitative characteristics of text, and

b) to attempt, within the framework of such an approach, to describe either the relations of the frequency F_i and its rank i for $i = 1, \ldots, V$, or those between the frequency F and the number V_F of the different units in the text by Zipf-Mandelbrot's Law.

The present paper deals with such an approach. As an object of study, the relations between the F and V_F in a complete, coherent text are considered (let us further call the set $\{F, V_F\}$, $F = 1, \ldots, F_{\max}$, *the frequency distribution of text*, and the set of the frequencies arranged in a diminishing order – the *frequency structure of text*). We attempt to describe these relations by the equation (2) and to study, after the values of the parameters are estimated, their interrelations

[3] It is to be noted that any other model postulating the functional connection of the parameters of Zipf-Mandelbrot's Law with the length, vocabulary-size or other quantitative characteristics of a concrete text would meet the same difficulties with the validity test

in a corpus of texts under study.[4] The material of the investigation consists of *musical texts*. The latter have been chosen because the repetition-organization was studied in complete musical compositions of a rather broad group of musical styles and, on the other hand, the law proved to be visually observable there in considerably small texts – on the average, 10-15 times smaller than the literary texts which followed the law with the same values of K, B and c. In other words, certain metastylistic regularities of the text organization revealed themselves in music with special clarity – perhaps due to the fact that various formal structures play an especially important role in music.

For the study, relatively large musical compositions of the 18th to the 20th centuries (sonata, symphony, cycle of the "Prelude and Fugue"-type, etc.) have been chosen. The texts have been analyzed on the level of F-motifs. In a number of them, according to previous investigations, Zipf-Mandelbrot's Law (2) with $K = 1/\ln F_{\max}$, $B = L/(F_{\max} \ln F_{\max})$, $c = 1$, was observable; in others, deviations of the actual frequencies from the values predicted were observable for the frequent F-motifs, whereas the frequencies of rare F-motifs proved to be in better accordance with the predictions.

The optimal parameters K, B and c in (2) have been obtained as follows.

For each of the investigated compositions, an empirical frequency distribution of the motifs was considered (Table 1).

To prove the validity of Zipf-Mandelbrot's Law, an attempt has

[4] To speak more generally, in the suggested approach, Zipf-Mandelbrot's Law is hypothesized to be valid for the frequency distribution. Such an approach can be justified by the fact that the validity of the law for the frequency structure implies its validity for the frequency distribution (see footnote 1 above), though generally speaking, the values of K, B and c of the law would in these two cases differ from each other. This connection between two descriptions, as well as the question, as to how the values of K, B and c are obtained in the description of the frequency distribution related to those obtained in frequency structure, will be discussed in our paper "The organization of repetition in a coherent text: two alternative approaches" (in preparation) The present study is a part of a research program on the quantitative-systemic study of musical composition/musical language carried out within the framework of the project "Language Synergetics" at the Institute for Language Study (Sprachwissenschaftliches Institut) of the Ruhr-University Bochum, West Germany.

Table 1

F. Mendelssohn. Prelude and Fugue for Organ, op. 35, e-moll

The frequency F	The number V_F of different Formal motifs (F-motifs) each of which occurs F times in the text
F	V_F
1	152
2	48
3	18
4	9
5	11
6	8
7	4
8	3
9	2
10	2
11	2
12	1
15	1
16	1
19	2
20	4
22	1
26	2
27	1
28	1
31	1
32	1
43	1
49	1
52	1
61	1
62	1
67	1
153	1

been made to fit the observed values V_F, $F = 1, 2, \ldots, F_{\max}$, by their theoretical prognoses V_F^*,

$$V_F^* = \frac{K}{(B+F)^c} \qquad (2*)$$

– i.e. to find the optimal values of the parameters K, B and c in (2) so that the value of V_F as little as possible deviated from V_F^*. For finding these optimal values, the optimization-procedure suggested in Neddler & Mead (1965) has been used. This procedure changes the parameters K, B and c successively, until the sum

$$S = \sum_{F=1}^{n} (V_F - V_F^*)^2,$$

(where n is the total number of different frequencies in the Table 1) of squares of the deviations of V_F from V_F^* is minimal.

As a goodness-of-fitting criterion, that commonly used in statistics has been used:

$$Krit = 1 - \sum_{F=1}^{n} \frac{(V_F - V_F^*)^2}{(V_F - \hat{V}_F)^2} \qquad (4)$$

(where $\hat{V}_F = 1/n \sum_{F=1}^{n} V_F$.) In general, the fitting is considered to be good if the value of *Krit* is greater than 0.9.

As the investigations showed, the suggested approach made it possible to estimate the parameters K, B and c so that the value of the criterion *Krit* was very close to 1 – i.e. the predicted frequency distribution proved to be very close to the actual one (see Table 2).

The investigation have demonstrated especially good agreement between the actual value of V_F and its theoretical prognosis for the F-motifs rarely occurred in the text – i.e. for a group of F-motifs which always form a very large part of the F-motif vocabulary of the musical composition. In particular, the suggested approach made it possible to predict with high accuracy the number of the *unrepeated* F-motifs in the composition (i.e. those occurring in it only once) which occupy nearly half of its F-motif vocabulary. In particular, for $F = 1, 2, 3$, the average of the relative deviations

Table 2

Material	L	V	K	B	c	*Krit*
J.S. Bach. Prelude and Fugue No 20, WTC, 1	1422	296	127.1683	-.1862	1.9999	.9987
J.S. Bach. Prelude and and Fugue No 2, WTC, 2	671	168	60.6725	-.2517	1.5836	.9944
J.S. Bach. Prelude and and Fugue No 22, WTC, 2	1430	310	109.2483	-.2586	1.8851	.9987
D. Scarlatti. Sonata No 1,(Keller/Weismann)	250	72	10.7947	-.7686	.8416	.9520
D. Scarlatti.Sonata No 9, (Goldenweiser)	156	53	1435.3000	3.5876	2.8066	.9713
D. Scarlatti. Sonata No 13, Nikolaev	568	190	8826.0000	2.3467	3.7076	.9974
G.F. Händel. Fugue for Organ, G-dur	765	201	43.3675	-.5311	1.3343	.9940
J. Haydn. Symphony No 45 ("Abschieds-)	1304	340	520.3695	.7103	2.1473	.9968
W.A. Mozart. Fugue d-moll	731	199	57.2817	-.3974	1.4438	.9967
A. Albrechtsberger. Fugue c-moll	693	168	21.1717	-.7725	1.1190	.9968
L.v. Beethoven. Sonata No 19	584	151	$1.1728 \cdot 10^6$	6.126	5.0236	.9840
L.v. Beethoven. Rondo op. 51 No 1	624	162	1666.0000	1.4526	2.9427	.9920
F. Chopin. Ballade No 1	902	206	141.7478	.1524	1.9040	.9909
F. Chopin. Sonata No 3	2366	504	9872.4000	2.1786	3.2238	.9983
F. Mendelssohn. Preludium and Fugue for Organ op. 35 No 6	1393	282	242.7958	.2557	2.0523	.9970

Table 2 (continued)

K. Saint-Saens. Introduction and Rondo-Capriccioso for Violin and Orchestra	1200	222	4990.5000	2.2129	3.3118	.9985
S. Prokofiev. Sonata for Violin Solo	1519	398	$81498 \cdot 10^6$	8.2634	8.8953	.9951
P. Hindemith. Sonata No 1 for Violin Solo	1452	374	1031.2000	.8547	2.5906	.9985
D. Šostakovič. Prelude and Fugue op. 87 No 4	983	237	42.0539	-.6291	1.2410	.9982
D. Šostakovič. Prelude and Fugue op. 87 No 9	677	146	58.8933	-.1412	1.5056	.9960
D. Šostakovič. Prelude and Fugue op. 87 No 12	1292	316	49.6455	-.6557	1.3614	.9989
D. Kabalevskij. Rondo op.59 for Piano	625	171	126.6199	.2644	1.7528	.9936
Ju. Levitin. Sonatina for Solo Flute	760	159	63.0779	-.1215	1.4344	.9921

$$D_F = \frac{(V_F - V_F^*)}{V_F^*}$$

of V_F from V_F^* in all of the compositions investigated, as well as the average of the absolute values abs(D_F) of D_F and the respective standard deviations proved to be the following:

Table 3

F	D_F	S_D	abs(D_F)	$S_{\text{abs}(D)}$
1	-.1525	.3407	.1836	.3242
2	3.1910	4.8580	4.5209	3.5886
3	-15.2501	18.0640	19.2214	13.5350

It is interesting to compare these data with those obtained in models of Zipf-Mandelbrot's Law where the values of K, B and c are hypothesized to be connected with the text length L, the size V of its vocabulary and the maximal frequency $F_{\max}$ of the unit under study in it:

a) within the framework of the model suggested in Orlov (1976) (where $K = 1/\ln F_{\max}$, $B = K \cdot L/F_{\max-1}$, $V_F^* = B \cdot F_{\max}/(F \cdot (F+1))$ and c is hypothesized to be equal to 1):

Table 4a

F	D_F	S_D	abs(D_F)	$S_{\text{abs}(D)}$
1	29.5488	44.8247	39.5094	36.0631
2	-6.6429	36.1729	30.6506	19.6474
3	7.6691	14.6231	12.4409	10.7223

b) within the framework of the generalized model of the Zipf-Mandelbrot's Law suggested in Orlov (1976) where

$$\begin{aligned} (K &= \frac{aL^a}{F_{\max}^a - 1})^a, \\ B &= (\frac{KL}{F_{\max}})^{1-a}, \\ V_F^* &= B \cdot (F_{\max}^{1-a} - 1) \cdot (F^{-a} - (F+1)^{-a}), \\ a &= 1 - 1/c, \end{aligned} \tag{4}$$

and c is estimated from the equation $V^* = B \cdot (F_{\max} - 1)^{1/c}$ where V^* is the actual size of the F-motif vocabulary of the text:

Table 4b

F	D_F	S_D	abs(D_F)	$S_{\text{abs}(D)}$
1	15.7650	23.2206	37.4112	33.9588
2	25.8242	43.6820	23.8915	14.3932
3	-25.0403	30.7071	33.9514	20.0328

c) within the framework of the model suggested by Krylov (1987) where K and B are considered to be functions of both L, V and $F_{\max}$, and where

$$V_F = \frac{A}{F^2} \exp^{a/F} \tag{5},$$

with the values of a and A being iteratively estimated from the equations

$$k_i = A_i(\exp\frac{F_{\max}}{A_i} - 1), \quad i = 0, 1, \ldots \tag{6a}$$

$$a_{i+1} = \ln(\frac{V}{A_i} - b_2 - a_i \cdot b_3 - \frac{a_i^2}{2} \cdot b_4 - \frac{a_i^3}{6} \cdot b_5 + \frac{1}{k_i}) \tag{6b}$$

$$A_{i+1} = \frac{L}{\ln k_i + b_1 + \exp a_{i+1} + a_{i+1} \cdot b_2 + \frac{a_{i+1}^2}{2} \cdot b_3 + \frac{a_{i+1}^3}{6} \cdot b_4} \tag{6c}$$

with $A_0 = V$ and $a_0 = 0$ as the respective initial values and b_i, $i = 1, 2, 3, 4, 5$, being the constants connected with the logarithmic derivative of the gamma-function:

Table 4c

F	D_F	S_D	abs(D_F)	$S_{\text{abs}(D)}$
1	-1.3180	16.9290	13.6306	9.8373
2	14.2297	38.1662	28.7952	28.4675
3	-22.6091	30.3260	32.4467	19.0023

As can be seen, the suggested approach made it possible to predict the size of the groups of the rare F-motifs in compositions more accurately than the models discussed in which the parameters K, B and c are hypothesized to be functionally dependent on the quantitative characteristics of text – as a matter of fact, of the concrete text under study.

It is natural to ask: how are the parameters K, B and c related to each other in the totality of the musical composition investigated? In other words, how "regular", in general, are the relations between frequencies of the F-motifs?

At the first sight, there is no visible order in the relations between K, B and c in Table 3: the values of K deviate enormously from each other – from 10.8 up to 8.10^{10} – and their average; the values of B and c also deviate in an apparently irregular way, or, at any rate, independently of text length or vocabulary-size.

However, more thorough analysis reveals certain regularities. For instance, it could be seen that the values of c in general, increase with the increase of B; in a more vague form the same thing could be observed in the relations between K and c. All this makes it reasonable to study the correlations – at least the linear ones – between the three parameters. The study has revealed the following:

Table 5

r(K,B) =	-6868	(statistically significant by P < 0.05)
r(K,c) =	.8163	(statistically significant by P < 0.001)
r(B,c) =	.9561	(statistically significant by P < 0.001)
r(ln K,B) =	.9525	(statistically significant by P < 0.001)
r(ln K,C) =	.9957	(statistically significant by p < 0.001)

In other words, all three parameters of the law proved to be positively correlated with each other; the respective coefficients of correlation are not only statistically significant but also quite high.

It is to be noted that the parameters B and c are very highly correlated with each other: the respective correlation-coefficient is very close to 1, which means practically linear relations between these parameters. It might be supposed that the organization of the F-motif repetitions in musical text can be described by Zipf-Mandelbrot's Law with only two parameters – specifically, K and B. It is to be noted that the study of the correlations between the $\ln K$ and B, resp. $\ln K$ and c (which has been considered reasonable due to the great variance of the parameter K in the investigated musical texts discussed above) has revealed very high positive correlations in both cases: $r(\ln K, B) = .9525$, $r(\ln K, c) = .9957$. Such very high correlations – practically, a linear functional dependence – could lead to hypotheses about the "logarithmic" relations in the organization of repetitions in a text. However, such speculations, as well as those on the possibility of describing the repetition-structure in the text by the two-parametric – or even one-parametric – model of Zipf-Mandelbrot's Law exceed the frames of the present paper, the main aim of which was to discuss an "alternative approach" to Zipf-Mandelbrot's Law in coherent texts, making it possible to prove, on the basis of certain statistical tests, whether it is fulfilled in specific, concrete cases.

References

Altmann, G. (1988), *Wiederholungen in Texten.* Bochum: Brockmeyer.

Arapov, M.V. (1988), *Kvantitativnaja Lingvistika.* Moskva: Nauka.

Boroda, M.G. (1973), "K voprosu o metroritmičeski elementarnoj edinice v muzyke." *Bull. of the Ac. Sci.*, vol. 71, No 3, 745-748.

Boroda, M.G. (1974), "O častotnoj strukture muzykal'nych soobščenij." *Bull. of the Ac. Sci. of the Georgian SSR*, vol. 76, No 2, 178-202.

Boroda, M.G. (1978), "Ob opredelenii informacionnoj melodičeskoj edinicy tipa frazy v muzyke." *Bull. of the Ac. Sci. of the Georgian SSR*, vol. 89, No 1., 57-60.

Boroda, M.G. (1979), *Principy Organizacii Povtorov na Microurovne Muzykal'nogo Teksta.* Diss., Tbilisi, typescript, 362 p.

Boroda, M.G. (1982), "Die melodische Elementareinheit." In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (Hrsg.), *Sprache, Text, Kunst: Quantitative Analysen.* Bochum, Brockmeyer, 205-221.

Boroda, M.G. (1986), *Problema Segmentacii v Muzyke. Strukturnye edinicy musykal'noj reči.* Dep. No 1203 Gos. Biblioteka SSSR, Dep. 1203, Moskva, 358 p.

Boroda, M.G. (1990), "The organization of repetitions in the musical composition: Towards a quantitative-systemic approach." *Musikometrika* 2, Bochum, Brockmeyer, 53-106.

Boroda, M.G., Polikarpov, A.A. (1988), "Zipf-Mandelbrot's Law and units of different text levels." *Musikometrika* 1, Bochum, Brockmeyer, 127-158.

Detlovs, V.K. (1984), "Obobščennye formal'nye motivy." *Algebra i diskretnaja matematika.* Riga.

Detlovs, V.K. (1988), "On microsegmentation of tunes. Alternative units to F-motif." *Musikometrika* 1, Bochum, Brockmeyer, 69-82.

Kalinin, V.M. (1965), "Funkzionaly, svjazannye s raspredeleniem Puassona, i statističeskaja struktura teksta." *Trudy matematičeskogo instituta imeni Steklova.* vyp. 79, Moskva-Leningrad.

Krylov, Ju.K. (1987), "Stacionarnaja model' porozhdenija svjaznogo teksta." *Acta et Commentationes Universitatis Tartuensis* 774, Tartu, 55-66.

Lebedev, A.N. (1983), "Zakonomernosti povtorenija slov v reči." *Psychologičeskij žurnal*, No 5, 11-22.

Lebedev, A.N. (1986), "Nejrofiziologičeskie predely pamjati čeloveka." *Acta et Commentationes Universitatis Tartuensis* 745, Tartu, 95-108.

Orlov, Ju.K. (1970), "Obobščenie zakona Zipfa-Mandel'brota." *Bull. Ac. Sci. of the Georgian SSR*, vol. 57, No 1, 37-40.

Orlov, Ju.K. (1976), "Obobščennyj zakon Zipfa-Mandel'brota i častotnye struktury informacionnych edinic različnych urovnej." *Vycislitel'naja lingvistika*, Moskva, 179-202.

Orlov, Ju.K. (1982), "Ein Modell der Häufigkeitsstruktur des Vokabulars." In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.S. (Hrsg.), *Sprache, Text, Kunst: Quantitative Analysen.* Bochum, Brockmeyer, 118-192.

Orlov, Ju.K. (1988), "Unsichtbare Harmonie." *Musikometrika* 1, Bochum, Brockmeyer, 263-270.

Tuldava, Ju.A. (1987), *Problemy i Metody Kvantitativno-Systemnogo Issledovanija Leksiki.*. Tallinn: Valgus.

Vološin, B.A., Orlov, Ju.K. (1982), "Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Verteilung der Anteile von Farbflaechen in der Malerei." In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.S. (Hrsg.), *Sprache, Text, Kunst: Quantitative Analysen.* Bochum, Brockmeyer, 263-270.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 61–71

The Constants of the Menzerath-Altmann Law

Ludek Hrebicek
Prague

The constants occurring in the formula of Menzerath-Altmann's law are derived here on the basis of certain empirical presumptions. We present their verification on the text structures called aggregations. We are not able to verify the validity of constants A^x and b^x on the other linguistic levels due to the absence of necessary data. Nevertheless, we believe that Menzerath-Altmann's law represents a general principle able to distinguish and constitute linguistic levels. As the professsion of the author of the present paper is Turkology, the theory is verified on the Turkish texts.

1. Menzerath-Altmann's law: a short recapitulation

P. Menzerath (1928) for the first time published his observation providing that a sound is the shorter the longer the whole in which it occurs; he stated that the more sounds in a syllable the shorter is its relative length. Later he ascertained that the relative number of phonemes decreases with the increasing number of syllables in the word, cf. P. Menzerath (1954).

This excellent observation remained almost unnoticed in linguistics till the publication of its mathematical formulation by G. Altmann (1980). His model starts from the following supposition:

> "The longer a language construct the shorter its components (constituents)." (G. Altmann 1980,1).

He wrote the differential equation:

$$\frac{y'}{y} = -c + \frac{b}{x},$$

where y is the mean length of the constituent, x is the length of the construct, and b and c are constants.

Altmann presented three possible solutions to this equation, from which the following one appeared the most favourable for applications:

$$y = Ax^b, \tag{1}$$

where A and b are constants. For example, when the length of a syllable (in the number of sounds or phonemes) is treated as the function of the length of the word (in the number of syllables), y is the mean length of syllable and x the length of the word. If Menzerath-Altmann's law holds, the value of b must be negative. The general formulation by Altmann leads to suppositions concerning the equilibrium of the language system and its subsystems.

According to our conviction, Altmann's contribution to the formulation of this law is substantial; therefore we propose to name it Menzerath-Altmann's law.

Altmann verified the theory on the length of syllables in morphemes of Indonesian, and on the length of syllables in English words as well as in Bachka-German words. Important developments of the theory are contained in the works by R.Grotjahn (1982), G.Altmann, E.Beöthy, K.-H.Best (1982) and by R.Köhler, G.Altmann (1986). A new monography announced and due to appear soon has not been available to us. The works cited contain references to other works relevant to the problem.

Now let us recapitulate our considerations concerning the higher structures of a text.

2. Vehicle aggregations

Text seems to be a rather vague and variable unit in comparison with other linguistic units. Linguistics arrived at a relatively rich and definitive description of sentences; however, between sentence and text there is a structural vacancy. We feel that there should exist a further linguistic level, i.e., units consisting of sentences.

In connection with Menzerath-Altmann's law, text itself cannot be treated as consisting directly of sentences; the supposition that the longer the text the shorter the sentence is unacceptable even in jest. The law used as an instrument for introduction or constitution of a new linguistic level features the following advantages:

1. The existence or non-existence of a new level is verified by Menzerath-Altmann's law.
2. The range of its application is enlarged.
3. A new instrument for differentiating subsequent levels is proved.

It can be argued that this is a way leading from the law to the level and, at the same time, from the level to the law. However, the validity of the law is verified on certain levels, the existence of which is beyond dispute. The literature quoted above offers reliable support for taking this law as a general principle of natural languages.

How shall we conceive of the new text level?

Sentences of each text consist of words; the number of words indicates the sentence length. A certain lexical unit occurs in sentences as a word. Let us call *vehicle aggregation* (v. aggretation) each set of sentences of a text, in which a certain lexical unit occurs.

The sentences with more than one occurrences of a lexical unit are treated in the same way as the sentences with one occurrence.

The following variables can be introduced: the length of aggregation x (in the number of sentences) and the mean sentence length y of the sentences belonging to a given aggregation (in the number of words). The following relation is derived from Menzerath-Altmann's law:

The longer the aggregation, the shorter the mean sentence length.

In Tables 1 and 2 the data concerning v.aggregations obtained from two Turkish texts are presented together with the values y_c indicating the mean sentence length computed according to (1). Constants A and b were computed from the observed values of y by the method of least squares, cf. G.Altmann (1980:4).

The two Turkish texts are Z.Gökalp (1970:46-51) and A.Gölpinarli (1965:81-82), the latter being a poem by Yunus Emre.

The mutual relation of y and y_c was studied from the viewpoint of their variances, means and the total concordance of their distributions with the help of F-test, t-test and Wilcoxon test. The following results were obtained from Tab.1:

$$F = \frac{s_1^2}{s_2^2} = 10.121 > 2.86 = F_{0.025}(15;15); H_0 : s_1^2 = s_2^2 \text{ is rejected.}$$

The means $\bar{y} = 12.41$ and $\bar{y}_c = 12.20$ are evidently in concordance.

Wilcoxon test: $T = 63 > 30 = T_{0.05}(16)$; there is no reason to reject H_0, saying that both the distributions are in concordance.

Irrespective of the difference between the variances, we can conclude that the curve Ax^b expresses the properties of the observed values of y.

Table 1
V. aggregations in Z. Gökalp (1970)

x	g	w	y=w/(xg)	y_c	y_c^*
1	238	3118	13.10	13.86	13.10
2	70	1712	12.23	13.28	12.33
3	31	1263	13.58	12.95	11.90
4	5	355	17.75	12.72	11.60
5	10	666	13.32	12.55	11.38
6	5	247	8.23	12.41	11.20
7	3	256	12.19	12.29	11.05
8	5	475	11.88	12.19	10.92
9	2	266	14.78	12.10	10.81
11	1	132	12.00	11.96	10.62
12	2	326	13.58	11.89	10.54
14	2	233	8.32	11.78	10.40
20	1	278	13.90	11.52	10.08
23	2	490	10.65	11.42	9.95
26	1	350	13.46	11.34	9.85
45	1	432	9.60	10.96	9.39

x the length of aggregation (in the number of sentences)
g the number of aggregations with length x
w the total number of words occurring in aggregations with length x
y the mean sentence lenght (in the number of words)
$y_c = Ax^b$, $A = 13.86$
b = -0.0616005 b^* = -0.0875917
k = 105 which is the number of sentences of the text
n = 915 which is the number of words of the text

Table 2
V. aggregations in A. Gölpinarli (1965)

x	g	w	y=w/(xg)	y_c	y_c^*
1	114	409	3.59	3.70	3.59
2	21	154	3.66	3.81	3.55
3	4	46	3.83	3.88	3.52
4	2	31	3.88	3.92	3.50
5	2	47	4.70	3.96	3.49
15	1	60	4.00	4.15	3.42
17	1	68	4.00	4.17	3.41

$A = 3.70$ $k = 66$ $b = 0.0426841$ $n = 220$ $b^* = -0.0177053$

The following results were obtained by testing the values of y and y_c from Table 2:
$F = 4.55 < 5.82 = F_{0.025}(6;6)$; there is no reason to reject $H_0 : s_1^2 = s_2^2$.
The means $\bar{y} = 3.95$ and $\bar{y}_c = 3.94$ are in evident concordance.

Wilcoxon test: $T = 7 > 2 = T_{0.05}(7)$; there is no reason to reject H_0, both the distributions are in concordance. In this case coefficient b is positive; consequently, the presupposition of Menzerath-Altmann's law is not fulfilled. Among eight texts analyzed up to now (of which seven are Turkish, one is an English text translated from Turkish, six are in prose and two in verses) this is the only exception with positive b. It must be stressed that this text is relatively short; it cannot demonstrate a balanced tendency of the random variables. Nevertheless, the theory must also include such cases. Therefore we introduce another type of aggregation.

3. Sign aggregations

It is generally accepted that each word in a text is a vehicle of sign, a carrier of meaning. When a text is produced by a producer (speaker, author) and when it is received by a hearer or reader, signs of different levels are interpreted by both the participants of communication; their interpretations correspond to their own structures of the communication memories.

As to the interpretation of the lexical signs, it can be understood as a process in two opposite directions:

(a) One and the same lexical unit occurring more than once in a text can be interpreted as carrying different meanings.

(b) Two or more different lexical units of a text can be interpreted as carrying an identical meaning.

Not only lexical units, but also lexical signs occur in a text in its sentences. Thus we arrive at *sign aggregations* (s. aggregations). It must be stressed that the process of interpretation is completely subjective, issuing from the communicants' intuition.

The Turkish texts analyzed above were interpreted by the author of the present paper. A more correct approach requires an interpretation by a group of competent speakers; however, we could not carry out such an experiment. Even a lower competence can offer one of possible interpretations.

S. aggregations and the related data of the two Turkish texts are presented in Table 3 and 4.

Table 3
S. aggregations in the text Z. Gökalp (1970)

x	g	w	y=w/(xg)	y_c	y_c^*
1	236	3183	13.49	12.81	13.49
2	77	1747	11.34	12.17	12.70
3	35	1336	12.72	11.81	12.25
4	10	466	11.65	11.56	11.95
5	9	526	11.69	11.37	11.72
6	4	213	8.88	11.22	11.53
8	4	377	11.78	10.99	11.24
11	1	111	10.09	10.73	10.93
12	1	145	12.08	10.66	10.85
14	4	574	10.25	10.54	10.71
23	1	269	11.70	10.16	10.25
25	1	233	9.32	10.10	10.18
51	1	479	9.39	9.59	9.56

A = 12.807 b = -0.0736913 b^* = -0.0875917

The results of testing the data from Table 3 are: $F = 2.44 < 3.28 = F_{0.025}(12;12)$; there is no reason to reject $H_0 : s_1^2 = s_2^2$.

The means are in an evident concordance: $\bar{y} = 11.106, \bar{y}_c = 11.055$.

Wilcoxon test: $T = 36 > 17 = T_{0.05}(13)$; there is no reason to reject H_0, both the distributions are in concordance.

Table 4
S. aggregations in the text A. Gölpinarli (1965)

x	g	w	y=w/(xg)	y_c	y_c^*
1	111	395	3.56	3.73	3.56
2	25	165	3.30	3.62	3.52
3	3	38	4.22	3.56	3.49
4	1	20	4.00	3.51	3.47
5	1	16	3.20	3.48	3.46
17	1	49	2.88	3.30	3.39
47	1	157	3.34	3.16	3.33

A = 3.73 b = -0.0429834 b^* = -0.0177053

The results of testing the data from Table 4 are: $F = 5.852 > 5.82 = T_{0.025}(6;6)$; the hypothesis $H_0 : s_1^2 = s_2^2$ is to be rejected. The means $\bar{y} = 3.50$ and $\bar{y}_c = 3.48$ are evidently in concordance. Wilcoxon test: $T = 13 > 2 = T_{0.05}(7)$; both the distributions are concordant. H_0 is not rejected.

In a paper by L. Hrebicek (in print) for the same texts and their s. aggregations, a somewhat different data are presented; the reason consists in new interpretations in which also the aggregations of the length $x = 1$ were taken into consideration.

The interpretation of the text A. Gölpinarli (1965), cf. Tabs. 2 and 4, resulted in s. aggregations which are distributed in accordance with Menzerath-Altmann's law. The equilibrium of the short text appeared unstable, but interpretation revealed its tendency.

The same instability was indicated by the results of F-tests. Constant b is always close to zero, but mostly under the zero value. Analogical results were obtained from the entire set of eight texts mentioned above.

4. Prediction of the constants

The computation of y_c requires the observated data of y, from which constants A and b can be ascertained. In this way, y cannot be predicted by y_c. Here is presented one procedure enabling the estimation of b.

In formula (2), x means a variable which can obtain an arbitrary positive value, though the linguistic facts necessitate only integers, i.e., $x = 1, 2, 3, \ldots$. When only integers are taken into account, the following progression is obtained:

$$y_1 = A1^b$$
$$y_2 = A2^b$$
$$y_3 = A3^b$$
$$\ldots$$
$$y_x = Ax^b$$

The ultimate term of this progression describing a text with k sentences is

$$y_k = Ak^b$$

An aggregation of lenght $x = k$ means that there is a lexical unit or lexical sign occurring in each sentence of the respective text. Thus the aggregation contains all sentences of the text, the whole text.

The value of the first term of this progression is evidently:

$$y_1 = A \tag{2}$$

As far as the ultimate term is concerned, if a text is supposed having k sentences and length n in the number of words, the value of this term is evidently:

$$y_k = Ak^b = \frac{n}{k} \tag{3}$$

Hence it follows:

$$\ln A + b \ln k = \ln n - \ln k$$

and

$$b = \frac{\ln n - \ln A}{\ln k} - 1 \tag{4}$$

The values of A and b estimated with the help of (2) and (4) are marked A^* and b^* in Tabs. 1 – 4. This approach requires for the estimation of A^* the observed data concerning aggregations with length $x = 1$. This can be understood as a reflection of the importance of *hapax legomena* observed by philologists long ago.

The results of testing the concordance between y and y_c^* are as follows:

V. aggregations

Table 1

$F = 6.182 > 2.86 = F_{0.025}(15;15)$; $H_0 : s_1^2 = s_2^2$ must be rejected. The concordance of means $\bar{y} = 12.41$ and $\bar{y}_c^* = 10.945$ was tested by t-test: $t = 2.183 > 2.131 = t_{0.05}(15;15)$; $H_0 : m_1 = m_2$ must be rejected. However, if the probability level $\alpha = 0.01$ is selected, the result is as follows: $t = 2.183 < 2.947 = t_{0.01}(15;15)$; there is no reason to reject $H_0 : m_1 = m_2$. Wilcoxon test: $T = 20 > 16 = T_{0.01}(15)$; both the distributions are in concordance.

Table 2

In this case from F-test and Wilcoxon test negative results were obtained; the distributions of y and y_c^* are not in concordance. This is understandable with respect to the rising values of y. The means are $\bar{y} = 3.951$ and $\bar{y}_c^* = 3.497$; the testing indicated their concordance.

S. aggregations

Table 3

$F = 1.642 < 3.28 = F_{0.025}(12;12)$; there is no reason to reject $H_0 : s_1^2 = s_2^2$. The evident concordance of the means $\bar{y} = 11.106$ and $\bar{y}_c^* = 11.34$ was also verified by the t-test. Wilcoxon test: $T = 30 > 14 = T_{0.05}(12)$; both the distributions are in concordance.

Table 4

$F = 36.044 > 5.82 = F_{0.025}(6;6)$; $H_0 : s_1^2 = s_2^2$ must be rejected. The evident concordance of the means $\bar{y} = 3.50$ and $\bar{y}_c^* = 3.46$ was verified also by t-test. Wilcoxon test: $T = 9 > 0 = T_{0.05}(6)$; both the distributions can be supposed to be concordant.

The analogical results obtained from the above- mentioned set of texts – among which the results related to Tab. 2 represent an exception – enable us to suppose A^* and b^* to be acceptable estimates of the constants.

We suppose that in this way the theoretical importance of variables n and k (for the two analyzed texts their observed values are presented in Tabs. 1 and 2) is corroborated. They are not simple quantitative indicators, but quantities influencing deeply linguistic structures.

We may try to formulate the following general assertion:

The criterion for the validity of Menzerath-Altmann's law is given by the expression

$$\frac{\ln n - \ln A}{\ln k} < 1,$$

where n is now understood as the sum of lengths of constituents (for example, the total sum of the lengths of words; these lengths are measured in the number of sounds or phonemes), k is the number of constituents (for example, the total number of syllables in a set of words) and A is the constant coefficient of Menzerath-Altmann's law, which can be estimated from the mean length of constituents observed on the shortest construct (for example, the mean length of syllables, in the number of sounds, observed on monosyllabic words).

This assertion has a general form; however, we could verify it here only on aggregations and sentences as their constituents.

References

Altmann, G. (1980), "Prolegomena to Menzerath's law". *Glottometrika* 2, 1-21.

Altmann, G., Beöthy, E., & Best, K.-H. (1982), "Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz". *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 35, 5, 537-543.

Gökalp, Z. (1970), *Türkçülüğün Esaslari*, Millî Eğitim, Istanbul, 46-51.

Gölpinarli, A. (1965), *Yunus Emre*, Risâlat al-nushiyya ve Dîvân, Sulhi Garan, Istanbul, 81-82.

Grotjahn, R. (1982), "Ein statistisches Modell für die Verteilung der Wortlänge". *Zeitschrift für Sprachwissenschaft* 1, 1, 44-75.

Hrebicek, L., *Menzerath-Altmann's law on the semantic level.* (in print)

Köhler, R. (1986), "Synergetische Aspekte der Linguistik". *Zeitschrift für Sprachwissenschaft* 5, 2, 253-265.

Menzerath, P. (1928), "Über einige phonetische Probleme". *Actes du premier congres international de linguistes.* Leiden, Sijthhoff, 104-105.

Menzerath, P. (1954), *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 73–83

Überprüfung einer Hypothese zur Kompositabildung (an polnischem Sprachmaterial)

Rolf Hammerl
Bochum

1. Einführung in den Untersuchungsgegenstand

Synergetische Untersuchungen der Lexik natürlicher Sprachen (unter lexikalischem, morphologischen, semantischen usw. Aspekt) sind in den letzten Jahren nicht selten gewesen; es wurden grundlegende Eigenschaften lexikalischer Einheiten untersucht, deren Verteilungen und gegenseitigen Abhängigkeiten modelliert (vgl. z.B. Altmann 1980, Köhler 1986, Köhler,Altmann 1989).

Altmann (1988) macht in seinem Artikel auf eine Teilmenge lexikalischer Einheiten aufmerksam, auf Komposita, für die – ausgehend von den Erkenntnissen bisher untersuchter Sprachgesetze – spezielle Gesetzeshypothesen formuliert wurden, von denen bisher nur eine einzige an deutschem Sprachmaterial überprüft wurde (vgl. Rothe 1988):

> Je größer die Polysemie eines Wortes, desto mehr Komposita kann es bilden (wo dieses Wort selbst Teil dieser Komposita ist).

Hier soll eine zweite Hypothese über Komposita anhand des Polnischen untersucht werden, die den Zusammenhang zwischen der Länge L der Grundwörter, die Komposita bilden, und der Zahl n_L der von diesen Grundwörtern der Länge L gebildeten Komposita betrifft. Unsere allgemeine Untersuchungshypothese lautet:

> Die Zahl der Komposita n_L (deren Grundwörter die Länge L besitzen) ist eine Funktion der Länge L deren Grundwörter.

Diese Hypothese stellt eine Verallgemeinerung der von Altmann (1988, 104) angeführten Hypothese

"the shorter a word, the more frequently it occurs in compounds"

dar.

Diese Verallgemeinerung wurde deshalb vorgenommen, weil wir in unseren Untersuchungen zur polnischen Sprache Lexeme eines Textwörterbuchs (Słownictwo 1974-1977) untersuchen, in welchem alle Lexeme aufgeführt werden (ohne Eigennamen, Kurzwörter, Zitate), die in einem Textkorpus mit einem Umfang von 500.000 Wörtern auftraten, und nicht etwa die Lexeme eines einsprachigen Lexikons. Wie wir schon in Hammerl,Sambor (1988) gezeigt haben, können sich die Verteilungen von Lexemen hinsichtlich quantitativer Eigenschaften im Textwörterbuch und im Lexikon signifikant unterscheiden. Es wird vermutet, daß auch die Abhängigkeiten zwischen bestimmten quantitativen Eigenschaften von Lexemen im Textwörterbuch und im Lexikon von unterschiedlicher Spezifik sind.

2. Zur Kennzeichnung der untersuchten Komposita

In den Untersuchungen zur polnischen Sprache (vgl. Sambor 1988) werden fünf Klassen von Komposita unterschieden, von denen in unseren weiteren Analysen lediglich die Klasse K_1 der semantisch regulären, nicht erweiterbaren eigentlichen Komposita berücksichtigt wurde. Dies deshalb, weil es sich nur hier unter synchronischem Aspekt um eine "lebende" Kompositaklasse handelt (synchronische Motivation), deren Komposita paraphrasiert werden können.

Die Komposita der Klasse K_3 können in diesem Sinne nicht paraphrasiert werden und stellen "versteinerte" Formen dar. Die Komposita der Klasse K_2 betreffen nur Adjektive, die jederzeit "bei Bedarf" gebildet werden können und beliebig erweiterbar sind.

Nun sollen die Komposita der Klasse K_1, die in unseren weiteren Untersuchungen Berücksichtigung finden, näher gekennzeichnet werden:

a) Diese Komposita enthalten charakteristische Kompositionsmorpheme des Typs -o-, -i-, z.B. *par-o-wóz = lokomotywa poruszana parą.*

b) Diese Komposita sind semantisch regulär, d.h., es besteht semantische Übereinstimmung zwischen der Bedeutung des ganzen Kompositums und der Bedeutung mindestens eines seiner Glieder in der Paraphrase.

c) Diese Komposita sind nicht erweiterbar, d.h. weder durch weitere Glieder am Ende des Kompositums noch durch weitere Glieder zwischen dessen Elementen (Bestandteilen).

Unsere Untersuchungen stützen sich auf eine Menge von 1010 Komposita der Klasse K_1, die innerhalb einer Menge von 30059 Lexemen des polnischen Häufigkeitswörterbuchs (Słownictwo 1974- 1977) auftragen. Für jedes Kompositum war dessen Frequenz in einem Textkorpus bekannt, welches 500.000 Wörter (Texte aus 5 Funktionalstilen zu je 100.000 Wörtern) umfaßte.

3. Empirische Untersuchung

3.1. Zusammenhang zwischen der Länge L von Wörtern und der Zahl der Komposita n_L, die diese Wörter der Länge L bilden

Der Zusammenhang zwischen L und n_L wird in den Tabellen 1a und 1b dargestellt, wobei in Tabelle 1a Wörter aller Wortarten berücksichtigt wurden und in Tabelle 1b dieselben Wörter ohne Numeralien. Die Notwendigkeit einer solchen Unterscheidung wird im Kapitel 3.2. motiviert.

Tabelle 1a
Abhängigkeit zwischen L und n_L (alle Wortarten)

L (in Silben)	n_L
1	257
2	482
3	200
4	55
5	14
6	2

Tabelle 1b
Abhängigkeit zwischen L und n_L
(alle Wortarten ohne Numeralien)

L (in Silben)	n_L
1	252
2	474
3	200
4	55
5	14
6	2

Wie man schon an den Daten der Tabelle 1 erkennt, kann die obige Hypothese von Altmann über den Zusammenhang zwischen L und n_L für polnische Komposita der Klasse K_1 im Textwörterbuch nicht bestätigt werden. Die größte Zahl von Komposita bilden im polnischen Textwörterbuch Wörter mit einer Länge von 2 Silben, weit weniger Wörter mit 1 Silbe, 3, 4, 5 und 6 Silben. Dieses Ergebnis wird im Kapitel 4 diskutiert.

3.2. Zusammenhang zwischen der Frequenz F der Komposita und der Zahl der Komposita n_F mit der Frequenz F

Das analysierte Textwörterbuch (Slownictwo 1974-1977) führt auch für alle Lexeme, die in einem Textkorpus von 500.000 Wörtern der polnischen Schriftsprache aufgetreten sind, deren Auftrittshäufigkeit (Frequenz F) auf. In Tabelle 2 wird der Zusammenhang zwischen dieser Frequenz F und der Zahl von Komposita mit der Frequenz F dargestellt.
Innerhalb des Gesamtvokabulars der Untersuchungen zur polnischen Sprache von 30059 Lexemen (vgl. Sambor 1988, Tab.2) gilt: 3,4% der Lexeme sind Komposita; innerhalb des entsprechenden Textkorpus, welches 500.000 Wörter umfaßt, gilt:

a) 0,87% der Textwörter sind Komposita (der Klasse K_1).
b) Wie man aus Tabelle 2 ersehen kann, nimmt die Zahl der Komposita auch mit wachsender Texthäufigkeit ab. Etwa 65% der

Tabelle 2
Abhängigkeit zwischen F und n_F für polnische Komposita der Klasse K_1 aus dem Textwörterbuch (alle Wortarten)

F	n_F	$F \cdot n_F$		F	n_F	$F \cdot n_F$		F	n_F	$F \cdot n_F$
1	656	656		16	4	64		32	1	32
2	160	320		17	2	34		34	2	68
3	63	189		18	2	36		38	1	38
4	35	140		19	2	38		51	1	51
5	15	75		20	2	40		67	1	67
6	9	54		21	–	–		79	1	79
7	12	84		22	1	22		85	1	85
8	2	16		23	1	23		96	1	96
9	9	81		24	–	–		101	1	101
10	4	40		25	1	25		102	1	102
11	4	44		26	1	26		123	1	123
12	–	–		27	1	27		124	1	124
13	2	26		28	–	–		183	1	183
14	2	28		29	1	29		229	1	229
15	1	15		30	1	30		262	1	262
								572	1	572
									1010	4374

Komposita tritt im untersuchten Textkorpus von 500.000 Wörtern nur einmal auf, etwa 16% zweimal, 6% dreimal usw. Die überwiegende Mehrzahl der Komposita gehört dem seltenen Vokabular an. Sambor (1975, 9/10) berücksichtigte als Kriterium für die Abgrenzung des seltenen und häufigen Vokabulars die mittlere Häufigkeit $\bar{F}$, die für das Vokabular eines Textkorpus von 100.000 Wörtern (welches Teil des Textkorpus von 500.000 Wörtern ist) etwa 9 betrug. Nehmen wir nun zur reinen Illustration, ohne es exakt beweisen zu müssen, für unser Vokabular von 30059 Lexemen eine mittlere Häufigkeit von $5 \cdot 9 = 45$ an, so gilt: 95% der Komposita gehören dem seltenen Vokabular an.

Analysiert man die häufigsten Komposita hinsichtlich ihrer Wortartzugehörigkeit, so stellt man fest, daß es sich hier fast ausschließlich um

Numeralien handelt. Unter den 10 häufigsten Komposita tritt nur ein Substantiv und ein Adjektiv auf; die restlichen Komposita sind Numeralien:

1.	F=572 dziewięćset (neunhundert)	Numerale
2.	F=268 pięćdziesiąt (fünfzig)	Numerale
3.	F=229 sześćdziesiąt (sechzig)	Numerale
4.	F=183 samokształcenie się (das Sich-Selbst-Bilden)	Substantiv
5.	F=124 osiemdziesiąt (achtzig)	Numerale
6.	F=123 siedemdziesiąt (siebzig)	Numerale
7.	F=102 pięćset (fünfhundert)	Numerale
8.	F=101 trzysta (dreihundert)	Numerale
9.	F= 96 współczesny (gegenwärtig)	Adjektiv
10.	F= 85 osiemset (achthundert)	Numerale

Für die polnische Sprache ist es somit typisch, daß die häufigsten Komposita der Klasse K_1 Numeralien sind. Da die Zahl dieser Numeralien theoretisch sehr groß sein kann (im Polnischen stellt jede Kardinalzahl ab 20 ein Kompositum dar), soll nun noch die Abhängigkeit zwischen F und n_F für alle Wortarten außer den Numeralien untersucht werden.

Tabelle 3
Zusammenhang zwischen F und n_F für polnische Komposita der Klasse K_1 (alle Wortarten zusammen außer den Numeralien)

F	n_F	F	n_F	F	n_F	F	n_F
1	654	10	4	19	2	28	–
2	160	11	4	20	2	29	1
3	63	12	–	21	–	30	1
4	35	13	2	22	1	32	1
5	15	14	2	23	–	34	2
6	9	15	1	24	–	38	1
7	12	16	4	25	1	79	1
8	2	17	2	26	1	96	1
9	9	18	2	27	1	183	1
						2633	

Wie man aus einem Vergleich der Tabellen 2 und 3 erkennen kann, treten Numeralien (Komposita) in ihrer überwiegenden Mehrzahl unter den häufigsten Komposita auf und haben somit eine sehr spezielle Häufigkeitsverteilung (vgl. Tabelle 4).

Tabelle 4
Rang-Häufigkeitsverteilung der Numeralien (Komposita)

r	F_r	r	F_r	r	F_r	r	F_r
1	572	4	124	7	101	10	51
2	262	5	123	8	85	11	23
3	229	6	102	9	67	12	1
						13	1

Soll nun die Abhängigkeit zwischen F und n_F bzw. die Rang- Häufigkeitsverteilung der Komposita modelliert (d.h. mathematisch beschrieben) werden, so müssen für jede Wortart getrennt spezielle Daten gesammelt und dann verglichen werden. Dies geht über den Rahmen dieses Artikels hinaus und soll Gegenstand einer gesonderten Arbeit sein.

3.3. Zusammenhang zwischen der Länge L der Grundwörter, die Komposita bilden, und der Gesamtfrequenz F_g aller Komposita der Grundwörter mit der Länge L

Tabelle 5a
Zusammenhang zwischen L und F_g (alle Wortarten)

L	F_g	F_g/n_L
1	1695	6,60
2	2129	4,42
3	426	2,13
4	101	1,84
5	21	1,50
6	2	1,00
	4374	

Tabelle 5b
Zusammenhang zwischen L und F_g
(alle Wortarten außer den Numeralien)

L	F_g	F_g/n_L
1	934	3,71
2	1149	2,42
3	426	2,13
4	101	1,84
5	21	1,50
6	2	1,00
	2633	

Tabelle 5a und 5b zeigt, daß die Komposita, deren Grundwörter eine Länge von 2 Silben haben, nicht nur innerhalb der Klasse aller Komposita am häufigsten vertreten sind (vgl. Tabelle 1), sondern auch in realen Texten (obwohl die Komposita, deren Grundwort eine Länge von 1 Silbe hat, im Durchschnitt häufiger im Text auftreten – dies beweist der Zahlenwert für F_g/n_L – dafür gibt es aber eben relativ wenige dieser Komposita).

4. Diskussion

Aus Tabelle 1 wird ersichtlich, daß zwischen der Zahl der Komposita n_L und der Länge deren Grundwörter L ein Zusammenhang besteht.

Wie wir aus den Untersuchungen Köhlers (1986) und besonders von Hammerl,Maj (1988) schon wissen, bestehen zwischen den Eigenschaften

Länge, Frequenz, Polylexie, Polytextie

lexikalischer Einheiten gegenseitige Abhängigkeiten. D.h. die Frequenz lexikalischer Einheiten ist z.B. eine Funktion sowohl der Länge, als auch der Polylexie und Polytextie, d.h. F=f(L,PL,PT). Es müssen jedoch noch mehr Abhängigkeiten dieser Art zwischen diesen 4 Eigenschaften und noch vielen anderen, in den Untersuchungen Köhlers nicht berücksichtigten Eigenschaften lexikalischer Einheiten (z.B. deren Abstraktheit, morphologischen Status) bestehen, die Gesetzescharakter tragen. Dies ist eine Grundvoraussetzung für eine effektive und schöpferische Gestaltung sprachlicher Kommunikate auf Sprecherseite und deren Dekodierung auf Hörerseite, d.h. eine Grundvoraussetzung der sprachlichen Kommunikation überhaupt.

Bezogen auf den oben untersuchten Zusammenhang zwischen n_L und L würde das bedeuten, daß wir es hier eigentlich nur mit der Randverteilung $n = f(L)$ einer zumindest 4-dimensionalen Verteilung n=f(L,PT,PL,F) zu tun haben, die dann, wenn sich z.B. PT, PL, F als Funktionen von L ausdrücken lassen, auch als eindimensionale Verteilung dargestellt werden kann.

Der Typ dieser Randverteilung $n_L = f(L)$ wird einerseits durch das übergeordnete "principle of least effort" (Zipf 1943) bei der sprachlichen Kommunikation gesteuert und zweitens durch die gesetzmäßigen Abhängigkeiten mit anderen Eigenschaften lexikalischer Einheiten.

Beim heutigen Wissensstand ist es uns noch nicht möglich, diese Randverteilung aus einem allgemeinen synergetischen Modell der natürlichen Sprachen als Spezialfall abzuleiten. In Tabelle 6 wird gezeigt, daß die untersuchte Verteilung z.B. mit der Hyperpoisson- bzw. Hyperpascal-Verteilung beschrieben werden kann. Inwieweit aber eine dieser Verteilungen theoretisch begründet und an entsprechenden Untersuchungen anderer Sprachen bestätigt werden kann, müssen weiterführende Untersuchungen zeigen. Es wird vermutet, daß der Typ dieser Verteilung

sowohl von der Wortart des Grundwortes als auch von der Wortart des Kompositums abhängig ist; dies wurde schon oben anhand der Wortart der Numeralien (als Komposita) deutlich gemacht.

Tabelle 6
Anpassung der Hyperpoisson- und der Hyperpascal-Verteilung an die empirischen Daten aus Tabelle 1

L	n_L	Hyperpoisson	Hyperpascal
1	257	252.76	264.31
2	482	474.05	478.68
3	200	215.67	188.99
4	55	55.83	57.97
5	14	10.10	15.93
6	2	1.58	4.12
		$a = 0.6007$	$k = 1.1606$
		$b = 0.3203$	$m = 0.1326$
		$\chi^2_3 = 2.97$	$q = 2.34$
		$p = 0.40$	$\chi^2_2 = 2.34$
			$p = 0.31$

Literatur

Altmann, G. (1980), "Prolegomena to Menzerath's law". *Glottometrika* 2, 1-10.

Altmann, G. (1988), "Hypotheses about compounds". *Glottometrika* 10, 100-107.

Hammerl, R., Maj, J. (1988), "Anmerkungen zum Modell von R. Köhler: Struktur und Dynamik der Lexik". *Glottometrika* 10, 1-31.

Hammerl, R., Sambor, J. (1988), "Vergleich der Längenverteilungen der Lexeme nach der Silbenzahl im Lexikon und im Textwörterbuch." *Glottometrika* 10, 198-204.

Köhler, R. (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R., Altmann, G. (1989), "Synergetic modelling of language phenomena". In: Köhler, R. (Ed). *Studies in Language Synergetics.* (erscheint).

Rothe, U. (1988), "Polylexy and Compounding". *Glottometrika* 9, 121-134.

Sambor, J. (1975), *O słownictwie statystycznie rzadkim.* Warszawa, PWN.

Sambor, J. (1988), "Polnische Version des Projektes 'Sprachliche Synergetik. Teil I. Quantitative Lexikologie'. *Glottometrika 10*, 171-197.

Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom I-V. Warszawa 1974-1977.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 85–92

Zur Verteilung japanischer wort- bzw. satzsyntaktischer Postpositionen in einem vollendeten Text

Undine Roos
Bochum

1. Einleitung

Das Japanische ist typologisch gesehen eine agglutinierende, polysyllabische Sprache. Die Beziehungen zwischen Wörtern und Sätzen werden durch die sog. *Postpositionen* hergestellt, die nach Lewin, Müller-Yokota & Fujiwara (1983,324) folgendermaßen definiert sind:

> "[Die Postpositionen] bilden die Funktionsformen der nominalwertigen Wörter und verändern zudem die Funktionen der verbalwertigen Wörter im Satz."

Sie sind mit den deutschen Präpositionen, Konjunktionen, Kasusendungen etc. vergleichbar.

Untersucht werden sollen nun folgende Hypothesen:

1. Die Wahrscheinlichkeitsverteilung der wortsyntaktischen Postpositionen ga_1 und *o* sowie der satzsyntaktischen Postpositionen ga_2 und *te* (vgl. Kapitel 2) unterscheiden sich signifikant.[1]
2. Die Verteilung der Häufigkeit der genannten Postpositionen auf einzelne Sätze eines vollendeten Textes erfolgt gesetzmäßig, d.h., sie kann durch eine entsprechende Wahrscheinlichkeitsverteilung beschrieben werden.

[1] Es sei darauf hingewiesen, daß ga_1 als wortsyntaktische Postposition und ga_2 als satzsyntaktische Postposition gezählt wurde.

2. Beschreibung der untersuchten Postpositionen

Bevor im nächsten Kapitel auf die Verteilungen der vier Postpositionen eingegangen wird, soll zunächst eine kurze Beschreibung ihrer grammatischen Funktion gegeben werden.[2]

ga が

Wortsyntaktische Postposition; Funktion: Bezeichnung des Subjekts

Beispiel:[3] 春が来た。
haru **ga** kita
(der Frühling ist gekommen)

Satzsyntaktische Postposition; Funktion: adversativ, koordinativ

Beispiel: 今日は学校が休みだが何をしようかなあ。
kyō wa gakkō ga yasumi da **ga**, nani o shiyō ka nā?
(heute ist kein Unterricht, was machen wir?)

o を

Wortsyntaktische Postposition; Funktion: Bezeichnung des Objekts

Beispiel: 山を見る。
yama **o** miru
(den Berg sehen)

[2] Diese Beschreibung ist nach Lewin, Müller-Yokota & Fujiwara (1983,324ff). Sie stellt nur eine sehr grobe Klassifizierung dar. Detaillierte Informationen findet man z.B. in Lewin (1974).

[3] Die Beispiele stammen allesamt wörtlich aus Lewin, Müller-Yokota & Fujiwara (1983,324ff).

te て

Satzsyntaktische Postposition; Funktion:

1. temporal ("und dann")

Beispiel: ユースがあって、ビールもある。
jūsu ga **atte** bīru mo aru
(es gibt Saft und dann auch noch Bier).

2. adversativ

Beispiel: これは大きくて、それは小さい。
kore wa ōkiku**te**, sore wa chiisai
(dies hier ist groß, das da aber klein)

3. konditional

Beispiel: 歩いて、十分かかる。
aruite, jippun kakaru
(wenn man zu Fuß geht, dauert es zehn Minuten)

4. kausal

Beispiel:
学生寮が小さくて、住む人が少ない。

gakuseiryō ga chiisaku**te**, sumu hito ga sukunai
(da das Studentenwohnheim klein ist, wohnen wenig Leute dort)

3. Empirische Untersuchung

In dieser Arbeit wurde ein sprachwissenschaftlicher Text mit einem Umfang von insgesamt 296 Sätzen[3] unter dem Gesichtspunkt der Häufigkeit des Auftretens der wort- bzw. satzsyntaktischen Postpositionen

$$ga_1,\ o,\ ga_2,\ te$$

untersucht.

[3] Bei dem Text handelt es sich um "Schreibweise und Wortbildung der Sprache" von S. Mizutani (1987).

Die Zählungen des Auftretens der vier Postpositionen in den Sätzen des genannten Textes ergab folgende Werte (Tabelle 1):

Tabelle 1
Häufigkeiten der vier Postpositionen

Häufigkeit	ga_1	o	ga_2	te
0	153	135	246	263
1	105	119	47	29
2	30	28	2	3
3	5	9	1	1
4	3	3	0	0
5	0	0	0	0
6	0	2	0	0

3.1. Überprüfung von Hypothese 1

Zunächst wurden die in Tabelle 1 aufgelisteten Daten für alle vier Postpositionen durch den $2I$-Test (vgl. Altmann 1981,143) auf Homogenität überprüft. Benutzt wurde die folgende Formel:

$$2I = 2\sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij} \ln n_{ij} + 2n \ln n - 2\sum_{i=1}^{r} n_{i.} \ln n_{i.} - 2\sum_{j=1}^{k} n_{.j} \ln n_{.j} \quad (1)$$

Als Ergebnis erhält man einen Wert von

$$2I = 220.53$$

mit 12 Freiheitsgraden. Da dieses Resultat den Tabellenwert von

$$\alpha_{.05;12} = 21.03$$

deutlich übersteigt, kann Homogenität der Daten nicht angenommen werden, d.h. die Postpositionen wurden nicht ein und derselben Grundgesamtheit entnommen. Deshalb wurde nun überprüft, ob innerhalb der jeweiligen Kategorien (wort- bzw. satzsyntaktisch) Homogenität gegeben ist. Tabelle 2 zeigt das Ergebnis.

Tabelle 2
$2I$-Werte für die beiden Kategorien der Postpositionen

Kategorie	$2I$	FG	krit. Tabellenwert bei $\alpha = 0.05$
wortsyntaktisch	3.22	4	9.49
satzsyntaktisch	5.07	3	7.82

Wie man sieht, liegt der Wert von $2I$ in beiden Fällen unter dem kritischen Tabellenwert. Es kann somit angenommen werden, daß sich die Häufigkeiten der beiden wort- bzw. satzsyntaktischen Postpositionen nicht signifikant unterscheiden.

3.2 Überprüfung von Hypothese 2

Nun soll die Wahrscheinlichkeitsverteilung der wortsyntaktischen Postpositionen ga_1 und o sowie der satzsyntaktischen Postpositionen ga_2 und te untersucht werden. Die Überprüfung von Hypothese 1 hat erbracht, daß es sich hier um 2 verschiedene Verteilungen handelt (wobei zunächst nicht ausgeschlossen werden kann, daß für jede Postposition ga_1, o, ga_2 und te eine andere Verteilung vorliegt). Die unterschiedliche grammatische Funktion der Postpositionen determiniert die Verteilung der Postpostionen. Wir untersuchen hier also zwei Subhypothesen:

Hypothese 2a:

Die Verteilung der wortsyntaktischen Postpositionen ga_1 und o kann aus einem Modellansatz hergeleitet werden, den Hammerl (1989) zur Verteilung der Wortarten im Text angewandt hat, wobei natürlich die dort verwendeten Konstanten hier eine andere Interpretation erhalten müssen. Dieses Problem kann aber im Rahmen dieser Arbeit noch nicht geklärt werden.

Der genannte Modellansatz führt auf die sogenannte modifizierte Zipf-Verteilung:

$$P_x = \begin{cases} \alpha & x = 1 \\ \frac{(1-\alpha)x^{a+b\ ln\ x}}{T} & x = 2,3,\ldots,n \end{cases} \qquad (2)$$

mit $\alpha = P_1$
und

$$T = \sum_{x=2}^{n} x^{a+b\ ln\ x},$$

wo a und b Konstanten darstellen.

Hypothese 2b:

Die Verteilung der satzsyntaktischen Postpositionen ga_2 und *te*, die – wie man aus Tabelle 1 ersehen kann – wesentlich seltener auftreten als die wortsyntaktischen Postpositionen, können mit der "Verteilung seltener Ereignisse", d.h. der Poissonverteilung, beschrieben werden (vgl. Altmann, Hammerl 1989,34):

$$P_x = \frac{a^x e^{-a}}{x!} \qquad \text{für } x = 0,1,2\ldots, \qquad (3)$$

wo a und e Konstanten sind ($e \approx 2,71828$).

Zur Überprüfung von Subhypothese 2a wurde die modifizierte Zipf-Verteilung an die empirische Verteilung der wortsyntaktischen Postpositionen (ga_1 und *o* zusammen) angepaßt (vgl. Tabelle 3) und entsprechend zur Überprüfung der Subhypothese 2b die Poissonverteilung an die empirische Verteilung der satzsyntaktischen Postpositionen (ga_2 und *te* zusammen; vgl. Tabelle 4).

In diesen Tabellen bedeutet x die Häufigkeit der jeweiligen Postposition pro Satz, n_x die empirisch ermittelte Zahl der Sätze mit x Postpositionen und NP_x die entsprechende theoretische Zahl der Sätze, die durch Anwendung des jeweiligen mathematischen Modells ermittelt wurde.

Tabelle 3
Verteilung der wortsyntaktischen Postposition ga_1 und o (modifizierte Zipf-Verteilung)

x	n_x	NP_x
0	288	288.00
1	224	224.28
2	58	56.28
3	14	16.20
4	6	5.31
≥ 5	2	1.93

$\alpha = 0.48648 \quad a = 1.0376 \quad b = 1.3241$
$\chi^2(1) = 0.45$

Tabelle 4
Verteilung der satzsyntaktischen Postpositionen ga_2 und te (Poisson-Verteilung)

x	n_x	NP_x
0	509	508.18
1	76	77.58
2	5	5.92
3	2	0.31

$a = 0.1527 \quad \chi^2(1) = 0.13$

Wie man an den obigen Tabellen sehen kann, werden auch die Subhypothesen 2a und 2b bestätigt, d.h., die wortsyntaktischen Postpositionen (ga_1 und *o*) folgen der modifizierten Zipf-Verteilung und die satzsyntaktischen Postpositionen (ga_2 und *te*) der Poisson-Verteilung.

Zusammenfassend kann man also folgendes feststellen:

- Die wort- und satzsyntaktischen Postpositionen (ga_1, *o*, ga_2, *te*) unterscheiden sich bzgl. der Häufigkeitsverteilungen in einem vollendeten Text.

– Die Postpositionen derselben Kategorie (wortsyntaktisch [ga_1, o]/ satzsyntaktisch [ga_2, te]) folgen derselben Wahrscheinlichkeitsverteilung.

Untersucht werden müßte nun, ob sich besonders Hypothese 2 auch dann bewahrheitet, wenn weitere Postpositionen der jeweiligen Kategorie hinzugezogen werden, ob analoge Feststellungen auch für wort- und satzmodifizierende Postpositionen gelten und ob sie auch dann zutreffen, wenn es sich in stilistischer Hinsicht um verschiedenartige Texte handelt. Dies soll Gegenstand einer anderen Arbeit sein.

Literatur

Altmann, G. (1981), "The Homogenity of Metric Patterns in Hexameter". In: Grotjahn, R. (Ed.), *Hexameter Studies.* Bochum, Brockmeyer, 137 – 150.

Altmann, G., Hammerl, R. (1989), *Diskrete Wahrscheinlichkeitsverteilungen I.* Bochum: Brockmeyer.

Hammerl, R. (1989), "Untersuchungen zur Verteilung der Wörtarten im Text". *Glottometrika* 11, 142-156.

Lewin, B. (1974), *Abriss der japanischen Grammatik.* Wiesbaden: Harrassowitz.

Lewin, B., Müller-Yokota & Fujiwara, M. (1983), *Einführung in die japanische Sprache.* Wiesbaden: Harrassowitz.

Mizutani, S. (1987). "Go no hyōki to gokōsei" In: Mizutani, S. (Ed.), *Chōsō nihongo shinkōza 6: moji, hyōki to gokōsei* Japan, Chōsō shoten, 1–19.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 93–105

Assoziationen und Dissoziationen zwischen dem Genus und der phonetischen Struktur der einsilbigen deutschen Nomina

U. Rothe
Bochum

1. Die Zuweisung eines Genus bei den deutschen Nomina wird allgemein als recht arbiträr angesehen (vgl. Brinkmann 1962, Admoni 1970, Greenberg 1978, Eisenberg 1986). Während die Genusdetermination in einigen Sprachen reglementiert ist (etwa nach phonetischen Regeln streng im Lateinischen, weniger streng in den modernen romanischen Sprachen, oder nach semantischen Klassen im Englischen und im Suahili), kann in anderen Sprachen, wie im Deutschen, das Genus anscheinend nicht in einem konsistenten Regelapparat bestimmt werden. Allerdings haben dennoch verschiedene Untersuchungen der Vergangenheit gezeigt, daß es semantische (Spitz 1965, Wienold 1967, Fleischer 1975, Wellmann 1975, Beito 1976 und Bechert 1982), morphologische und phonetische (Helbig/Buscha 1972, Altmann/Raettig 1973, Zubin/Köpcke 1981, 1984) Faktoren unterschiedlicher Gewichtigkeit gibt, die im Deutschen bestimmte Genuszuweisungen zum Nomen motivieren. Einen guten Überblick über den derzeitigen Stand der Genusforschung bietet Köpcke (1987), dessen Datensammlung zu den deutschen Einsilbern auch Grundlage des vorliegenden Beitrags ist.

2. Für die Untersuchung von genustypischen Strukturmerkmalen der Nomina auf phonetischer Ebene sind die Anordnungen der Konsonanten und Vokale im Wort relevant, wobei unterschieden wird nach den Strukturen von Anlaut, Inlaut, Auslaut und dem ganzen Wort. Man kann die Stärke der Assoziationen und der Dissoziationen zwischen den Strukturmerkmalen und den drei Genera in einem Wahrscheinlichkeitsmodell bestimmen, wie es Altmann/Raettig (1972) gezeigt haben.

Dieses Modell ist geeignet, wenn die Genuszuweisung generell homogen ist, d.h., wenn sich im Lexikon (bzw. hier bei den Einsilbern) die Maskulina, Feminina und Neutra auf die gesamten Nomina mit gleicher Häufigkeit verteilen. Dies ist aber bei den vorliegenden Daten nicht der Fall, wie in einem Homogenitätstest mit Hilfe der Informationstatistik 2I nach Kullback (1959) zunächst überprüft wurde. Die Formel des Tests lautet:

$$2I = 2\sum\sum_{j=1}^{k} n_j \ln n_j + 2\sum_{j=1}^{k} n_j \ln n_k - 2\sum_{j=1}^{k} n_j \ln n_j. \qquad (1)$$

n_j bezeichnet dabei die Häufigkeiten eines Genus, k die Zahl der Genera insgesamt (=3) und n die Zahl der katalogisierten Einsilber (1466). Es ergibt sich ein Wert von $2I = 603,95$ mit 2 Freiheitsgraden. Die bereits aus den beobachteten Daten zu erkennende Inhomogenität der Genusverteilung über alle 1466 katalogisierten Nomina (940 Maskulina, 205 Feminina, 321 Neutra) bestätigte sich in dem Test, da

$$603,95 > \chi^2_{.05(2)}.$$

Der geringe Anteil der femininen Nomina an den Einsilbern wurde auch von Koepcke angesprochen mit dem Hinweis, daß bei Mehrsilbern diese Minderzahl durch mehrere produktive feminine Suffixe wie *-heit, -keit, -ung* etc. sowie durch den nur in Mehrsilbern auftauchenden meist femininen Auslaut *-e* wieder ausgeglichen wird (p. 45).
Anschließend wurden hier in Anlehnung an Altmann/Lehfeldt (1981) und Schulz/Altmann (1988) die Werte in Kontingenztafeln erfaßt und anschließend mit dem Test nach Rehák, Reháková 1980 (vgl. Altmann/Lehfeldt 1981:301) ausgewertet. Dieser Test erlaubt eine Beurteilung dahingehend, ob sich bestimmte Merkmale mit bestimmten Genera signifikant assoziieren bzw. dissoziieren oder nicht. Die Berechnungsformel lautet:

$$z = \frac{n_{ij} - E_{ij}}{[((n_{i.}n_{.j})(n - n_{i.})(n - n_{.j}))/n^2(n-1)]^{1/2}}. \qquad (2)$$

Die Ergebnisse des Tests sind die Quantile der Normalverteilung, die angeben, ob bestimmte Zellen signifikant unter- bzw. überbelegt sind.

Diese Werte sind in den nachfolgenden Tabellen jeweils in der 6. bis 8. Spalte ("Quantile") aufgeführt.

Die qualitative Bewertung der Quantile der Normalverteilung in Tab. 2 erfolgt aufgrund des Existenzkriteriums (nij = 0, nij > 0) und der Signifikanz der Abweichung der beobachteten Zellenhäufigkeit von der theoretischen wie folgt:

M = marginal, wenn $n_{ij} > 0$ und $n_{ij} < E(n_{ij})$, signifikant abweichend
A = aktuell, wenn $n_{ij} > 0$ und $n_{ij} = E(n_{ij})$, nicht signifikant abweichend
P = präferiert, wenn $n_{ij} > 0$ und $n_{ij} > E(n_{ij})$, signifikant abweichend
V = virtuell, wenn $n_{ij} = 0$ und $n_{ij} = E(n_{ij})$, nicht signifikant abweichend
I = unzulässig, wenn $n_{ij} = 0$ und $n_{ij} < E(n_{ij})$, signifikant abweichend.

Die für unsere Untersuchung interessanten Bewertungen sind M und P. Als marginal (M) bewertet wird eine Zelle, wenn ihre Häufigkeit signifikant unter dem Erwartungswert liegt, ein Genus also eine bestimmte Lautkonfiguration dissoziiert. Als präferiert (P) bewertet wird eine Zelle, wenn ihre Häufigkeit signifikant über dem Erwartungswert liegt, was bedeutet, daß ein Genus ein bestimmtes Lautbild assoziiert. Aktuell (A) ist eine Zelle, wenn ihre Häufigkeit weder nach oben noch nach unten signifikant vom Erwartungswert abweicht. Leere Zellen (Nullbelegung) führen zu den Beurteilungen virtuell (V), wenn der entsprechende Erwartungswert ebenfalls Null ist bzw. unzulässig (I), wenn der Erwartungswert über Null liegt.

3. Nachfolgend werden die Kontingenztafeln und ihre Auswertungen aufgeführt. Es wird differenziert nach phonetischer Struktur der Anlautkonsonanten (Tab. 1), der Auslautkonsonanten (Tab. 2), der Inlautvokale und -diphthonge (Tab. 3) sowie schließlich der Gesamtstruktur (Tab. 4).

In der ersten Hauptspalte wird jeweils die phonetische Struktur aufgelistet, in der zweiten werden Beispiele aufgeführt. Die drei letzten Hauptspalten beinhalten nacheinander gelesen – jeweils für die 3 Genera – die beobachteten Häufigkeiten nach Köpcke (Spalte 3-5),

Tabelle 1
Kontingenztafel der Aufteilung der Anlautstrukturtypen über die drei Genera

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
b -V-F	Bad	38	12	16	-1.18	1.04	0.51	A	A	A
d "		25	1	10	0.64	-1.95	0.89	A	A	A
f "		35	9	14	-0.66	0.37	0.45	A	A	A
g "	Geld	19	11	17	-3.49	1.92	2.44	M	A	P
h "		42	9	20	-0.94	-0.30	1.35	A	A	A
j "	Jod	2	2	6	-2.94	0.56	2.94	M	A	P
k "		47	11	16	-0.16	0.25	-0.02	A	A	A
l "		27	11	18	-2.58	1.27	1.93	M	A	A
m "		39	13	17	-1.40	1.22	0.60	A	A	A
n "	Naht	9	9	5	-2.55	3.53	0.00	M	P	A
p "		38	10	10	0.18	0.76	-0.85	A	A	A
r "		41	4	19	-0.06	-1.80	1.58	A	A	A
t "		36	4	10	1.14	-1.22	-0.30	A	A	A
v "		33	11	15	-1.38	1.08	0.70	A	A	A
z "		34	6	12	0.15	-0.50	0.24	A	A	A
s "		40	9	10	0.56	0.31	-0.91	A	A	A
bl "	Blatt	4	0	5	-1.25	-1.21	2.47	A	V	P
br "		8	5	8	-2.53	1.33	1.83	M	A	A
dr "		14	2	1	1.56	-0.25	-1.59	A	A	A
dz "		3	0	0	1.29	-0.70	-0.91	A	V	V
fj "		2	0	0	1.05	-0.57	-0.75	A	V	V
fl "		19	6	6	-0.36	0.89	-0.32	A	A	A
fr "	Fracht	8	5	1	-0.57	2.38	-1.33	A	P	A
gl "		3	1	4	-1.59	-0.11	1.94	A	A	A
gn "		2	0	1	0.08	-0.70	0.49	A	V	A
gr "		20	2	6	0.79	-1.04	-0.04	A	A	A
kl "		16	3	4	0.52	-0.12	-0.51	A	A	A

Tabelle 1
Fortsetzung

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
kn "	Knick	14	0	1	2.35	-1.56	-1.42	P	V	A
kr "		17	3	6	0.11	-0.35	0.17	A	A	A
kv "		7	3	4	-1.13	0.82	0.62	A	A	A
pf "		9	0	4	0.37	-1.45	0.79	A	V	A
pl "		7	1	2	0.37	-0.36	-0.13	A	A	A
pr "		9	3	3	-0.36	0.69	-0.16	A	A	A
ps "	Psalm	1	0	0	0.74	-0.40	-0.53	A	V	V
sf "		1	1	0	-0.43	1.48	-0.75	A	A	V
sk "		5	0	1	0.97	-0.99	-0.30	A	V	A
sl "		2	0	0	1.05	-0.57	-0.75	A	V	V
sm "		1	0	0	0.74	-0.40	-0.53	A	V	V
∫t "		1	0	1	-0.43	0.57	0.97	A	V	A
∫v "		1	0	0	0.74	0.40	-0.53	A	V	V
tr "	Tran	33	3	0	3.46	-0.97	-3.20	P	A	I
ts "		18	6	9	-1.19	0.72	0.78	A	A	A
vr "		0	0	1	-1.35	-0.40	1.90	V	V	A
∫l "	Schleim	22	3	1	2.17	-0.35	-2.23	P	A	M
sm "		14	2	1	1.56	-0.25	-1.59	A	A	A
∫n "	Schnee	7	1	0	1.37	-0.11	-1.49	A	A	V
∫p "	Span	25	2	3	2.19	-1.15	-1.57	P	A	A
∫r "		10	1	0	1.84	-0.46	-1.75	A	A	V
∫t "	Sturm	45	2	3	3.85	-2.06	-2.75	P	M	M
sv "	Schwan	21	0	4	2.07	-2.02	-0.70	P	I	A
pfl "	Pflock	4	1	0	0.73	0.40	-1.18	A	A	V
pfr "		2	0	0	1.05	-0.57	-0.75	A	V	V
skr "		0	0	1	-1.35	-0.40	1.90	V	V	A
spl "		1	0	0	0.74	-0.40	-0.53	A	V	V
tsv "		7	2	0	0.84	0.73	-1.59	A	A	V
spl "		4	0	0	1.49	-0.80	-1.06	A	V	V
spr "		6	2	1	0.14	0.73	-0.78	A	A	A
∫tr "	Strand	18	1	1	2.41	-1.16	-1.83	P	A	A
0 -V-F	Ulk	26	10	20	-2.86	0.88	2.59	M	A	P

Tabelle 2
Kontingenztafel der Aufteilung der Auslautstrukturtypen über die drei Genera

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
I-V-b	Grab	2	0	0	1.06	-0.57	-0.75	A	V	V
" ç	Blech	8	0	5	-0.20	-1.46	1.45	A	V	A
" f		33	2	9	1.52	-1.82	-0.24	A	A	A
" g		1	0	0	0.75	-0.40	-0.53	A	V	V
" k	Stock	73	5	17	2.66	-2.52	-0.98	P	M	A
" l		51	5	23	0.07	-2.00	1.59	A	M	A
" m		40	6	6	1.95	-0.51	-1.84	A	A	A
" n		50	7	14	1.13	-1.01	-0.46	A	A	A
" p		40	1	11	1.95	-2.55	-0.13	A	M	A
" r	Haar	27	19	27	-4.97	3.06	3.19	M	P	P
" s	Gras	59	5	43	-2.02	-2.87	4.75	M	M	P
" t	Rad	40	17	35	-4.27	1.30	3.86	M	A	P
" x		16	1	10	-0.54	-1.55	1.92	A	A	A
" s		17	1	2	1.96	-1.16	-1.30	A	A	A
" ñ	Schwung	23	1	2	2.61	-1.50	-1.77	P	A	A
" çt	Gicht	4	4	3	-1.93	2.16	0.43	A	P	A
" ft	Gruft	7	8	4	-2.50	3.57	-0.09	M	P	A
" ks		22	5	3	1.06	0.44	-1.59	A	A	A
" kt		4	2	1	-0.39	1.12	-0.49	A	A	A
" lç	Dolch	10	1	0	1.86	-0.46	-1.76	A	A	V
" lf		4	2	2	-0.84	0.91	0.21	A	A	A
" lk		10	2	2	0.57	0.04	-0.69	A	A	A
" lm		13	1	1	1.83	-0.82	-1.43	A	A	A
" ln		1	0	0	0.75	-0.40	-0.53	A	V	V
" lp		4	1	3	-0.84	-0.12	1.07	A	A	A
" ls		5	0	1	0.98	-0.99	-0.31	A	V	A
" lt		12	4	9	-1.70	0.30	1.72	A	A	A
" ls		1	0	0	0.75	-0.40	-0.53	A	V	V
" mp		6	0	1	1.19	-1.07	-0.49	A	V	A
" ms		3	1	2	-0.73	0.19	0.68	A	A	A

Tabelle 2
Fortsetzung

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
" mt	Amt	2	0	3	-1.13	-0.90	2.06	A	V	P
" ms		1	0	0	0.75	-0.40	-0.53	A	V	V
" nç		1	0	0	0.75	-0.40	-0.53	A	V	V
" nf		2	1	0	0.09	0.97	-0.92	A	A	V
" ns	Gans	1	2	0	-1.11	2.64	-0.92	A	P	V
" nt		33	6	12	0.08	-0.45	0.28	A	A	A
" ns	Mensch	4	0	1	0.74	-0.90	-0.10	A	V	A
" pf	Zopf	13	0	0	2.71	-1.46	-1.92	P	V	V
" ps	Mops	29	1	1	3.45	-1.74	-2.54	P	A	M
" pt		1	0	2	-1.11	-0.70	1.88	A	V	A
" rç	Lurch	3	0	0	1.30	-0.70	-0.92	A	V	V
" rf		8	0	2	1.05	-1.28	-0.15	A	V	A
" rk		8	2	4	-0.55	0.04	0.61	A	A	A
" rl		5	1	1	0.40	0.03	-0.49	A	A	A
" rm		12	3	2	0.56	0.45	-1.02	A	A	A
" rn		10	1	4	0.20	-0.82	0.45	A	A	A
" rp		3	1	1	-0.19	0.39	-0.10	A	A	A
" rs		3	1	0	0.45	0.64	-1.06	A	A	V
" rt		21	3	6	0.67	-0.63	-0.26	A	A	A
" rs	Marsch	7	2	0	0.85	0.72	-1.59	A	A	V
" st	Wurst	27	13	4	-0.39	3.04	-2.09	A	P	M
" ts		34	2	4	2.79	-1.65	-1.85	P	A	A
" t∫	Matsch	11	2	3	0.38	-0.17	-0.31	A	A	A
" xt	Macht	1	18	0	-5.39	10.24	-2.32	M	P	I
" ∫t		1	1	0	-0.42	1.47	-0.75	A	A	V
" ñk	Schrank	21	2	1	2.40	-0.80	-2.12	P	A	M
" ñs		0	0	1	-1.34	-0.40	1.89	V	V	A
" çts		0	0	1	-1.34	-0.40	1.89	V	V	A
" kst		1	1	0	-0.42	1.47	-0.75	A	A	V
" lps		1	0	0	0.75	-0.40	-0.53	A	V	V
" lst		2	1	0	0.09	0.97	-0.92	A	A	V
" lts		7	3	3	-0.78	0.96	0.10	A	A	A

Tablle 2
Fortsetzung

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
" mpf	Dampf	10	0	0	2.37	-1.28	-1.68	P	V	V
" mps		1	0	0	0.75	-0.40	-0.53	A	V	V
" nst	Gunst	2	3	1	-1.58	2.56	-0.31	A	P	A
" nts		6	1	0	1.19	0.03	-1.40	A	A	V
" nts		3	1	0	0.45	0.64	-1.06	A	A	V
" nft	Zunft	0	2	0	-1.89	3.52	-0.75	V	P	V
" pst		2	0	1	0.09	-0.70	0.48	A	V	A
" rçt	Furcht	0	1	0	-1.34	2.49	-0.53	V	P	V
" rft	Werft	1	2	0	-1.11	2.64	-0.92	A	P	V
" rkt		1	0	0	0.75	-0.40	-0.53	A	V	V
" rps		1	0	0	0.75	-0.40	-0.53	A	V	V
" pst		6	1	0	1.19	0.03	-1.40	A	A	V
" rts	Herz	9	1	3	0.38	-0.65	0.10	A	A	A
" tst		0	0	1	-1.34	-0.40	1.89	V	V	A
" ñks		1	1	0	-0.42	1.47	-0.75	A	A	V
" ñkt		1	0	0	0.75	-0.40	-0.53	A	V	V
" ñst		1	1	0	-0.42	1.47	-0.75	A	A	V
" rnst	Ernst	1	0	0	0.75	-0.40	-0.53	A	V	V
" rpst		1	0	0	0.75	-0.40	-0.53	A	V	V
" lpst		0	0	1	-1.34	-0.40	1.89	V	V	A
" rtst		1	0	0	0.75	-0.40	-0.53	A	V	V
I-V- 0	Frau	33	22	22	-4.01	3.81	1.45	M	P	A

Tabelle 3
Kontingenztafel der Aufteilung der Inlautstrukturtypen über die drei Genera

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
I-a -F	Satz	186	40	39	2.28	0.58	-3.12	P	A	M
e	Bett	91	16	46	-1.27	-1.33	2.58	A	A	P
i		111	21	36	0.56	-0.59	-0.16	A	A	
o	Trotz	106	11	25	2.75	-2.26	-1.30	P	M	A
ö		3	1	0	0.45	0.64	-1.06	A	A	V
u	Schutt	123	31	13	2.73	1.81	-4.68	P	A	M
ü		7	3	2	-0.42	1.10	-0.44	A	A	A
a:		50	14	24	-1.47	0.54	1.26	A	A	A
n:		4	1	2	-0.39	0.02	0.43	A	A	A
e:	Mehl	25	7	20	-2.46	-0.11	2.94	M	A	P
i:	Biest	33	5	25	-1.99	-1.41	3.49	M	A	P
o:	Lob	40	3	27	-1.25	-2.40	3.46	A	M	P
ö:	Flöz	4	2	6	-2.23	0.27	2.36	M	A	P
u:	Wut	38	20	11	-1.61	3.68	-1.23	A	P	A
ü:		3	2	3	-1.57	0.90	1.07	A	A	A
ai		67	10	23	0.62	-1.19	0.28	A	A	A
au		47	14	13	-0.11	1.26	-0.92	A	A	A
oü	Deut	2	4	6	-3.44	1.94	2.36	M	A	P

Tabelle 4
Kontingenztafel der Aufteilung der Gesamtstrukturtypen über die drei Genera

Struktur	Beispiel	Häufigkeit			Quantile			Auswertung		
		m	f	n	m	f	n	m	f	n
V	Ei,Au	0	1	1	-1.89	1.47	0.96	V	A	A
VK	Aus	7	1	14	-3.18	-1.29	4.77	M	A	P
KV	See	16	12	11	-3.05	3.06	0.97	M	P	A
KVK	Rad	220	47	126	-4.68	-1.58	6.76	M	A	P
V2K	Ort	17	6	4	-0.13	1.25	-0.90	A	A	A
2KV	Knie	16	6	8	-1.24	0.96	0.64	A	A	A
KV2K	Lust	228	63	63	0.13	2.38	-2.14	A	P	M
2KVK	Pfiff	223	21	54	4.32	-3.87	-1.77	P	M	A
3KV	Stroh	1	3	2	-2.43	2.55	0.68	M	P	A
KV3K	Werft	37	12	6	0.50	1.71	-2.01	A	A	M
2KV2K	Klotz	113	23	16	2.78	0.43	-3.58	P	A	M
3KVK	Strick	28	1	0	3.68	-1.65	-2.88	P	A	I
2KV3K	Knirps	18	5	2	0.83	0.87	-1.69	A	A	A
3KV2K	Strang	12	2	1	1.29	-0.07	-1.43	A	A	A
3KV3K	Strumpf	1	0	0	0.75	-0.40	-0.53	A	V	V
KV4K	Hengst	1	0	1	-0.42	-0.57	0.96	A	V	A
V3K	Axt	0	2	2	-2.68	2.08	1.36	I	P	A
V4K	Angst	2	0	0	2.06	-0.57	-0.75	A	V	V

Tabelle 5
Assoziationen zwischen den Genera und der phonetischen Struktur der Einsilber

Phonet. Typen	m	f	n
Anlauttypen	kn,tr,sl, sp,st,sv, str	n,fr,	g,j,bl, 0-V-F
Auslauttypen	k,n̄,pf,ps, ts,n̄k,mpf	r,çt,ft, ns,st,xt, nst,nft, rçt,rft, I-V-0	r,s,t,mt
Inlauttypen	a,o,u	u:	e,e:,i:, o:,ö:,oü
Gesamtstrukturtypen	2KVK, 2KV2K, 3KVK	KV,KV2K, 3KV,V3K	VK,KVK

Tabelle 6
Dissoziationen zwischen den Genera und der phonetischen Struktur der Einsilber

Phonet. Typen	m	f	n
Anlauttypen	g,j,l,n, br,0-v-F	st	ʃl,ʃt
Auslauttypen	r,s,t,ft, xt,I-V-0	k,l,p,s	ps,st,n̄k
Inlauttypen	e:,i:,ö:, oü	o,o:	a,u
Gesamtstrukturtypen	VK,KV, KVK,3KV	2KVK	KV2K, KV3K, 2KV2K

die Quantile der Normalverteilung nach dem Test von Rehák/Reháková (1980) (Spalte 6-8) und schließlich die qualitative Auswertung (Spalte 9-11).

Die in Tabelle 1 bis 3 in der ersten Spalte auftauchenden Großbuchstaben V, I und F bedeuten Vokal bzw. Initialkonsonant oder -konsonantencluster bzw. Finalkonsonant oder – konsonantencluster. In Tabelle 4 bedeutet der Buchstabe K Konsonant, hinzugefügte Ziffern bezeichnen die Clustergröße.

4. Es ergeben sich bei allen Strukturtypen mehrere Stellen, an denen signifikante Assoziationen (P) bzw. Dissoziationen (M) bestehen, die im einzelnen in Tab. 5 und 6 zusammengefaßt sind.

Die o.a. Resultate deuten darauf hin, daß im Deutschen Trends bestehen, bestimmte phonetische Muster an die Speicherung der Genuszuordnungsregeln zu koppeln, die Zuweisung also nicht arbiträr erfolgt. Überlappungen mit morphologischen Eigenschaften der Nomina sind vorerst weitgehend ausgeschlossen, da bei Einsilbern Ableitungsmorpheme entfallen.

Nach der vorausgehenden Untersuchung kann man also durchaus signifikante Angaben darüber abgeben, welche phonetischen Muster bevorzugt an welche Genusmarkierung gekoppelt sind. Deshalb dürfte es nun interessant sein, die weitergehende Frage zu stellen, wodurch solche Assoziationen und Dissoziationen entstehen. Zum einen ist hier zu klären, inwieweit solche Lautbild-Genuskoppelungen dem Vereinfachungsprinzip entspringen, wonach der Sprechende dieselben Muster vorzugsweise an dieselben Kategorien bindet, um Speicherungsbemühungen (hier der Genuszuweisungregeln) einzusparen. Zum anderen darf vermutet werden, daß universelle Trends in der Sprachverwendung eine Rolle spielen, möglicherweise dahingehend, dem Lautbild metaphorisierende Eigenschaften zuzusprechen, wie es u.a. Fónagy (1963) sehr eindrucksvoll dargelegt hat.

Die erste Frage könnte anhand eines theoretisch und empirisch fundierten zweidimensionalen Verteilungsmodells geklärt werden. Für die Beantwortung der letzten Frage müßten viele (Genuszuweisung praktizierende) Sprachen in o.a. Weise parallel untersucht und anschließend die auffälligen phonetischen Muster verglichen werden. Eine Bestätigung der Vermutung universeller metaphorisierender Trends könnte aufschlußreiche Hinweise für die Theorie des sprachlichen Denkens und der Sprachpsychologie liefern.

Literatur

Admoni, W., (1970), *Der deutsche Sprachbau.* München: Beck'sche Verlagsbuchhandlung.

Altmann, G., Lehfeldt, W., (1980), *Quantitative Phonologie.* Bochum: Brockmeyer.

Beöthy, E., Altmann, G., (1984), "The diversification of meaning of Hungarian verbal prefixes". *Finnish-Ugrische Mitteilungen* 8, 29-37.

Brinkmann, H., (1954), "Zum grammatischen Geschlecht im Deutschen. Festschrit für E. Öhmann". *Annales Acad. Scientiarium Finnicae B.* 84, 371-428.

Brinkmann, H., (1962). *Deutsche Sprache. Gestalt und Leistung.* Düsseldorf: Schwann.

Eisenberg, P., (1986), *Grundriß der deutschen Grammatik.* Stuttgart: Metzler.

Fleischer, W., (1975), *Wortbildung der deuschen Gegenwartssprache.* Tübingen: Niemeyer.

Fónagy, I., (1963), *Die Metaphern in der Phonetik. Ein Beitrag zur Entwicklungsgeschichte des wissenschaftlichen Denkens.* The Hague: Mouton.

Greenberg, J.H., (1979), "How does a Language Acquire Gender Markers?" In: Greenberg, J.H. (Ed.), *Universals of Human Language, Vol. 3: Word Structure.* Stanford, 47-82.

Hirsch-Wiercbicka, L., (1971), *Funktionelle Belastung und Phonemkombination am Beispiel einsilbiger Wörter der deutschen Gegenwartssprache.* Hamburg: Buske.

Köhler, R., Altmann, G., (1986), "Synergetic modelling of language phenomena". *Zeitschrift für Sprachwissenschaft* 5, 253-265.

Köpcke, K.-M., (1982), *Untersuchungen zum Genussystem der deutschen Gegen- wartssprache.* Tübingen: Niemeyer.

Wellmann, H., (1975), *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. 2. Hauptteil: Das Substantiv.* Düsseldorf: Schwann.

Wienold, G., (1967), *Genus und Semantik.* Meisenheim a. Glan, Hain.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 107–114

Verteilung der Suffixe denominaler Verben nach ihren semantischen Wortbildungsmustern

Ursula Rothe
Bochum

Eine Untersuchung zur Verteilung von Präfixen denominaler Verben ergab, daß sich die semantischen Felder, denen Verbalpräfixe angehören, nach dem Diversifikationsmodell verteilen (vgl. Rothe 1990). Der vorliegende Beitrag untersucht, ob der Diversifikationsprozeß auch bei der Verteilung der Suffixe denominaler Verben nach ihren semantischen Wortbildungstypen nachgewiesen werden kann. Der Beitrag stützt sich auf dieselben Ergebnisse, die unter dem Titel "Deutsche denominale Verben" von Vl. D. Kaliuščenko (1988) veröffentlicht wurden und im o.a. Beitrag von Rothe bereits unter dem Aspekt der Präfix-Diversifikation untersucht wurden.

Denominale Verben sind Verben, die aus Nomina gebildet werden. Diese Verb-Neubildungen entstehen durch Präfigerung (*Eis – vereisen*) eines Nomens, durch Suffigierung (*Bagatelle – bagatellisieren*), oder auch durch affixlosen Wechsel von der Wortart Nomen zur Wortart Verb (*Löffel – löffeln*).

2. Wie bereits anläßlich des Artikels zu den Präfixen (vgl. Rothe 1990) festgestellt wurde, handelt es sich hier um eine Diversfikation nicht nach den Suffix-Bedeutungen im lexikographischen Sinne, sondern nach den semantischen Motiven der denominalen Verbalisierungen. Es ist also eine Frage, die nicht den aktuellen (semantischen) Status des Suffixes, sondern direkt seinen Beitrag bei der Entstehung einer Neubildung betrifft. Dies ist insofern von Wichtigkeit, als diese letztgenannte Art von Diversifikation einerseits das Lexikoninventar und die Lexemlänge vergrößern, andererseits hingegen die syntaktische Struktur vereinfachen, verkürzen. Das Gegenteil ist bei der Diversifikation von Einheiten nach ihren Bedeutungen im lexikographischen Sinne (vgl. Köhler 1986:78) der Fall.

Z.Zt. sind Suffixe als Mittel zur Verb-Neubildung weniger produktiv als Präfixe. Deshalb ist zu erwarten, daß – in Einklang mit der synergetischen Grundidee der konkurrierenden Sprecher-/Hörerkräfte – weniger Suffixe zur Verfügung stehen, die als Motivierung für eine Neubildung in Frage kommen als bei den Präfixen, sich aber deshalb die möglichen semantischen Muster konzentrierter auf die wenigen zur Verbneubildung bereitstehenden Suffixe verteilen.

Eine bloße Betrachtung der Gesamtzahl in Frage kommender Klassen (max. 16 bei den Präfixen *(ver-)* und max. 20 *(-ier)* bei den Suffixen) läßt allerdings eine solche Spekulation zunächst nicht zu. Zunächst sei noch einmal die Tabelle der von Kaliuščenko (1988:113) zusammengefaßten Deutungsformeln mit Beispielen in Tabelle 1 angeführt.

3. Insgesammt konnten die drei Suffixe *-ier (kompostieren), -isier (kapitalisieren)* und *-(e)l (witzeln)* für die Untersuchung verwendet werden. Die Verteilung ihrer Deutungsformeln wurde mit dem sog. modifizierten Zipfschen Modell, welches auch für die Präfixe die besten Anpassungen ergab, verglichen. Die linguistische Interpretation und die synergetische Ableitung der entsprechenden Verteilung kann in Hammerl (1989) nachgelesen werden. Die Formel lautet:

$$P_x = \begin{cases} \alpha & \text{für } x = 1 \\ \frac{(1-\alpha)x^{a+b\ln x}}{T} & \text{für } x = 2,3,\ldots,n \end{cases} \qquad (1)$$

wobei $\alpha = f_1/N, T = \sum_{x=2}^{n} x^{a+b\ln x}$.

Die Tabellen 2-4 beinhalten die entsprechenden Berechnungen für die Suffixe *-ier, -isier* und *-(e)l* für das Nhd. nach Kaliuščenko (1988:113-115). Tabelle 5 enthält zum Vergleich die reflexiven Verben *(sich zieren)*.

Die einzelnen Spalten der Tabellen geben von links nach rechts gelesen wider:

1. die Schlüsselzahl der Deutungsformel (Bed.), gemäß Tab.1.,
2. die Ränge der Klassen x, nach abnehmender Häufigkeit geordnet,
3. die beobachteten Häufigkeiten n(x) der einzelnen Klassen,
4. die berechneten Häufigkeiten P(x).

Tabelle 1
Semantische Klassen (Deutungsformeln) denominaler Verben nach Kaliuscenko 1988

Klasse	Deutungsformel	*Beispiele*
(1)	S1 ist (wie) Sm	*gärtnern, wurmen, spiegeln*
(2)	S1 wird zu Sm	*verdummteufeln, sich vernarren*
(3)	S1 macht S2 zu Sm	*adeln, mopsen, kapitalisieren*
(4)	Sm erscheint bei/auf S1	*verlausen*
(5)	S1 verliert Sm	*nadeln*
(6)	S1 versieht S2 mit Sm	*grundieren, verursachen*
(7)	S1 nimmt S2 den Gegenstand Sm	*köpfen, entnerven*
(8)	S1 schafft Sm	*bahnen*
(9)	S1 handelt typisch an Sm	*wildern*
(10)	S1 schafft ein Sm des Gegenstands S2	*formen*
(11)	S1 wirkt auf Sm = Teil von S2	*blättern*
(12)	S1 stellt S2 aus Sm her	*knibbeln*
(13)	S1 wirkt auf S2 mit Hilfe von Sm	*bürsten*
(14)	S1 handelt mit Hilfe von Sm	*ankern*
(15)	S1 führt eine Handlg. Sm an S2 aus	*alarmieren*
(16)	S1 führt eine Handlg. Sm aus	*pokern*
(17)	S1 befindet sich im Zustand Sm	*verunglücken*
(18)	S1 löst bei S2 ein Gefühl Sm aus	*verseuchen*
(19)	S1 sagt Sm (zu einer Person S2)	*beichten*
(20)	Sm = Beziehung (zw. S1 und S2)	*reimen*
(21)	S1 befindet sich am Ort Sm	*zelten*
(22)	S1 gerät an den Ort Sm	*sich einschiffen*
(23)	S1 entfernt sich von Sm	*ausufern*
(24)	S1 plaziert S2 am Ort Sm	*zetteln*
(25)	S1 entfernt S2 aus Sm	*ausstopfen*
(26)	Sm = Art und Weise der Hdlg. v. S1 an S2	*pachten*
(27)	Sm = Zeit ("Periode")	*übersommern*
(28)	Sm = Situation	*tagen*
(29)	(Vereinzelte denom. Verben)	*akademisieren, zetern, basteln*

Die unteren 4 Zeilen der Tabellen geben von oben nach unten gelesen wider:

1.u.2. die Parameter a bzw. b der Verteilung,
3. den χ^2 - Wert mit der Zahl der Freiheitsgrade in Klammern,
4. die Wahrscheinlichkeit P, mit der ein so kleines oder ein noch kleineres χ^2 zu erwarten ist .

Tabelle 2
Verteilung des Suffixes *-ier* nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
6	1	83	83.00
3	2	31	37.94
15	3	30	31.40
16	4	29	26.07
1	5	21	21.90
13	6	21	18.64
26	7	19	16.04
10	8	17	13.94
24	9	16	12.23
19	10	14	10.81
21	11	12	9.62
18	12	11	8.61
2	13	9	7.75
29	14	6	7.01
9	15	5	6.37
17	16	5	5.81
20	17	4	5.32
14	18	3	4.88
7	19	1	4.50
12	20	1	4.15
a = 0.0000			
b = 0.2604			
$\chi^2_{(16)} = 15.85$			
P = 0.463487			

Tabelle 3
Verteilung des Suffixes *-isier* nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
3	1	18	18.00
1	2	10	11.22
6	3	10	8.52
19	4	7	6.50
26	5	5	5.05
10	6	3	3.99
18	7	3	3.21
29	8	3	2.62
15	9	2	2.16
16	10	2	1.81
24	11	2	1.53
2	12	1	1.30
20	13	1	1.11
a = 0.0000			
b = 0.3786			
$\chi^2_{(9)} = 1.00$			
P = 0.999427			

4. Die Resultate sind für alle drei Suffixe zufriedenstellend. Da sich gute Anpassungen auch bei den Präfixen ergaben, kann man darauf schließen, daß die modifizierte Zipf-Verteilung als Modell der Diversifikation von Affixen nach semantischen Wortbildungsmustern angemessen ist. Bei *-ier* und bei *-isier* ist der Parameter a gleich Null, er scheint demnach für die Interpretation keine Rolle zu spielen. Ähnliches war bereits bei der Verteilung der Präfixe zu beobachten.

In den früheren Untersuchungen zur semantischen Diversifikation hat sich die negative Binominalverteilung als adäquates Modell erwiesen. Bei den Affixen ergaben Testberechnungen mit dieser Verteilung fast ausschließlich nicht signifikante Resultate.

Möglicherweise kommt darin die Tatsache zum Ausdruck, daß es sich hier um einen komplizierteren Fall von Diversifikation handelt. Es sind nicht nur einzelne Wörter bzw. Morpheme und ihre lexikalischen

Tabelle 4
Verteilung des Suffixes *-(e)l* im Nhd. nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
1	1	13	86.00
29	2	11	14.64
14	3	6	6.86
6	4	5	4.97
13	5	5	3.85
9	6	3	3.11
2	7	2	2.59
3	8	2	2.21
17	9	2	1.91
19	10	2	1.68
21	11	2	1.50
16	12	1	1.34
26	13	1	1.21
28	14	1	1.10
$a = 0.9861$ $b = 0.0535$ $\chi^2_{(10)} = 0.99$ $P = 0.999835$			

Bedeutungen betroffen, sondern sowohl semantische wie auch morphologische und syntaktische Aspekte. Während bei der semantischen Diversifikation Einheiten eines lexikalischen Systems zu Ungunsten anderer Einheiten desselben Systems mit zusätzlichen Bedeutungen beladen werden, werden bei der soeben untersuchten Art der Diversifikation von Affixen Einheiten eines lexikalischen Systems morphologisch und syntaktisch verändert, um auf syntaktischer Ebene Kodierungs- und Strukturierungsbemühungen zu sparen. Es werden also zugunsten einer syntaktischen Vereinfachung komplexere Einheiten in das lexikalische System gebracht.

Inwieweit solche syntaktischen und morphologischen Aspekte in dem Modell interpretiert werden müssen, in welcher Weise sie die Zahl der möglichen Klassen von Bedeutungen und damit den Diversifikationsprozeß beeinträchtigen, wird in weiteren Untersuchungen an wortbildenden Einheiten zu klären sein.

Tabelle 5
Verteilung des *Reflexiva* nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
2	1	51	51.00
4	2	19	17.91
18	3	11	13.71
5	4	10	10.54
29	5	9	8.25
17	6	8	6.56
1	7	7	5.31
22	8	5	4.36
14	9	3	3.62
27	10	3	3.04
21	11	2	2.58
23	12	2	2.21
20	13	1	1.90
$a = 0.0001$ $b = 0.3676$ $\chi^2_{(9)} = 2.33$ $P = 0.985099$			

Literatur

Altmann, G., "Diversification processes of the word". In: Köhler, R. (ed.), *Studies in Language Synergetics.* (erscheint)

Altmann, G., "Two models for word association data." In: Köhler, R. (Ed.), *Studies in Language Synergetics.* (erscheint)

Beöthy, E., Altmann, G. (1984a), "The diversification of meaning of Hungarian verbal prefixes. II." *Finnisch-Ugrische Mitteilungen 8*, 29-37.

Beöthy, E., Altmann, G. (1984b), "Semantic diversification of Hungarian verbal prefixes. III. "föl-", "el-", "be-"." *Glottometrika 7*, 45-56.

Beöthy, E., Altmann, G., (1990), "Semantic diversification of Hungarian verbal prefixes. I. "meg-"." In: Rothe, U. (Ed.), *Diversification Processes in Language: Grammar.* (im Druck)

Best, K.-H., (1990), "Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen." *Glottometrika 11*, 107-110.
Dolinskij, V.A., (1988), "Raspredelenie reakcij v experimentach po verbal'nym associacijam". *Acta et Commentationes Universitatis Tartuensis 827*, 89-101.
Erben, J., (1975), "Einführung in die deutsche Wortbildungslehre." Berlin: Schmidt.
Fillmore, Ch., (1968),"The case for case". In: Bach, E., Harms, R.T. (Eds.), *Universals in Linguistic Theory.* Holt, Rinehart and Winston.
Hammerl, R., (1989),"Untersuchungen zur Verteilung der Wortarten im Text." *Glottometrika 11*, 142-156.
Kaliuščenko, Vl.D., (1988), *Deutsche denominale Verben.* Tübingen: Narr.
Köhler, R., (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Rothe, U., (1990), "Semantische Beziehungen zwischen den Präfixen deutscher denominaler Verben und den motivierenden Nomina". *Glottometrika 11*, 111-121.
Rothe, U., (1990), "Diversification Processes in Language. An Introduction". In: Rothe, U. (Ed.), *Diversification Processes in Language: Grammar.* (im Druck)
Zipf, G. K., (1949), *Human Behavior and the Principle of Least Effort.* Cambridge, Mass.: Addison-Wesley.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 115–124

Ein methodischer Beitrag zum Piotrowski-Gesetz

K.-H.Best, Göttingen
E.Beöthy, Amsterdam
G.Altmann, Bochum

1. Der Ansatz

Die Ableitung und Überprüfung des sogennaten Piotrowski- Gesetzes, mit dem man Veränderungen sprachlicher Entitäten, Vokabularbereicherung durch Fremdwörter oder das Anwachsen des Vokabulars modelliert (vgl. Altmann, v. Buttlar, Rott, Strauß 1983; Beöthy, Altmann 1982, Altmann 1983, Best 1983a; Best, Altmann 1986; Kohlhase 1983; Best, Kohlhase 1983; Imsiepen 1983; Müller-Hasemann 1983; Tuldava 1984), zeigte, daß dieser Ansatz fruchtbar ist und weiter entwickelt werden kann. Diese Entwicklung ist aus zwei Gründen angezeigt.

(1) Theoretischer Grund

In der Theorie sind drei Ansätze gemacht worden, je nachdem, ob es sich um eine *vollständige* Veränderung einer Entität A in Entität B handelt:

$$p_t' = ap_t(1 - p_t), \tag{1}$$

um eine *unvollständige* Veränderung

$$p_t' = ap_t(c - p_t), \tag{2}$$

oder eine *reversible* Veränderung

$$p_t' = a_t p_t(1 - p_t), \tag{3}$$

wobei a_t nicht konstant, sondern eine Funktion der Zeit ist. Hier ist überall p die Proportion der alten Formen (A), die mit der Proportion

der neuen Formen $1 - p$ in Interaktion stehen, und $p_t' = dp/dt$, d.h. die Veränderungsrate von p nach der Zeit ist.

Bei der reversiblen Veränderung wurde der Ansatz $a_t = b - kt$ verwendet (vgl. Altmann 1983). Bedenkt man aber, daß *vor* und *während* der Veränderung noch nicht bekannt ist, wie sie verlaufen wird, dann ist es sowohl aus theoretischen als auch aus empirischen Gründen adäquater, alle drei Verläufe mit einer einzigen Kurve abzudecken, die den Proportionalitätsfaktor (a_t) im Laufe des Wandels verändert und als Asymptote eine allgemeine Konstante hat, die im Spezialfall (1) zu 1 werden kann.

Vereinigt man alle drei Ansätze, so erhält man

$$p_t' = a_t p_t (K - p_t) \tag{4}$$

mit der Lösung

$$p_t = \frac{K}{1 + exp(-K \int a_t dt)} \tag{5}$$

Als erste Approximation zur Spezifizierung von a_t scheint eine lineare Funktion $a_t = b - kt$, wie oben angegeben, plausibel zu sein, weil dieser Ansatz am einfachsten ist, und man am Anfang der Forschung die Sachen möglichst nicht unnötig komplizieren sollte. Die Integration und Reparametrisierung ergibt

$$p_t = \frac{K}{1 + Aexp(Bt + Ct^2)} \tag{6}$$

In der Linguistik wurde des öfteren beobachtet, daß logarithmische Zusammenhänge keine Ausnahmen bilden, so daß man auch den Ansatz $a_t = b - k\ln t$ verwenden könnte, was zu der Kurve

$$p_t = \frac{K}{1 + At^{Bt}e^{Ct}} \tag{7}$$

führt.

Bei der Überprüfung werden wir uns auf die Kurve (6) beschränken, die alle an Daten bisher beobachteten "glatten" Verläufe erfaßt.

(2) Empirischer Grund

In der Empirie hat man sehr heterogene Daten untersucht, mit recht zufriedenstellenden Resultaten. Es wurde aber beobachtet, daß die Daten nicht immer den "glatten" Verlauf mit "optisch" beschränkter Fluktuation aufweisen, wie man es in zufälligen Stichproben erwarten würde. Auch wenn das Testen positiv ausging, war man mit den Resultaten nicht besonders zufrieden (vgl. Imsiepen 1983). Man muß hier bedenken, daß Beobachtungen noch nicht Daten sind, daß Daten von uns produziert werden, daß Beobachtungen nur im Lichte einer Theorie zu Daten werden können. Im Falle des Piotrowski-Gesetzes ist die Datenerhebung besonders kritisch. Man muß als Quellen immer geschriebene Texte benutzen, wobei man für frühere Jahrhunderte nicht immer Zufallsstichproben erheben kann, sondern eben das nehmen muß, was es gibt, was noch zugänglich ist. Dadurch entstehen vier Arten von Verzerrungen:

(a) Ein bestimmter gewählter Text kann dialektal, sozio-lektal usw. so gestaltet werden, daß er eher eine evolutionäre Verzögerung oder Beschleunigung des untersuchten Merkmals beinhaltet, die sich dann als eine starke Abweichung von dem Kurvenverlauf erweist.

(b) Die Einteilung der Daten in zeitliche Intervalle ist willkürlich und kann zu ähnlichen Verzerrungen führen wie die (nicht zufällige) Wahl eines Textes. Ein Text kann viele Jahre vor seinem Erscheinen geschrieben worden sein, so daß er automatisch in einem falschen zeitlichen Intervall liegt.

(c) In der Sprachgeschichte gibt es Zeipunkte, wo eine Autorität oder ein Ereignis die Sprachnorm plötzlich verändert. Im Datenverlauf können an diesen Stellen Brüche, starke Abweichungen oder Umkehrungen des Trends erschienen. Solche Phänomene sind keine Falsifikationsinstanzen der Theorie – da die ceteris paribus Bedingung geändert wurde. Im Gegenteil, gerade solche Abweichungen der Daten von der Theorie geben Anlaß, nach den Ursachen in der Sprachgeschichte zu suchen.

(d) Die Untersuchung von Daten in einem großen Sprachgebiet kann automatisch zu Verzerrungen führen, wenn ein Dialekt in den Daten übermäßig vertreten wird. Dieses Problem läßt sich leich-

ter bewältigen als die Tatsache, daß in den Daten, besonders älteren, nur eine soziale Schicht vertreten ist, nämlich die gehobenere, die schreiben konnte.

Um die ersten zwei Arten der Verzerrungen einigermaßen zu eliminieren, empfiehlt es sich, eine Glättung der Daten mit Hilfe von gleitenden Durchschnitten durchzuführen. Durch die Glättung wird eine textbedingte Beschleunigung oder Verzögerung der Merkmalveränderung in einem Zeitintervall durch die Ausprägungen in den benachbarten Zeitintervallen etwas nivelliert, und gleichzeitig schwächt man auch den Einfluß der gegebenen (willkürlichen) Zeiteinteilung ab. Man kann ruhig sagen, daß man dadurch die Beobachtungen in adäquatere Daten überführt. Dies ist besonders bei solchen Daten von Bedeutung, wo man die Beobachtungen nicht in regelmäßigen Zeitabständen (z.B. Jahr für Jahr) machen kann.

Betrachten wir die Daten von Imsiepen (1983) zur e-Epithese, wie sie sich bei starken Verben im Deutschen entwickelt hat (vgl. Tabelle 1). Die Mittelwerte der Zeitintervalle t' wurden in diskrete Punkte transformiert, indem wir von ihnen 1465 subtrahierten und durch 10 dividierten, so daß wir $t = 1, 2, \ldots$ erhielten. Die Anzahlen der Verben mit der e-Epithese in den verwendeten Texten (f_t) und die Anzahl aller Verben in diesen Texten (N_t) sind in der dritten und vierten Spalte der Tabelle zu sehen. Die relative Häufigkeit $p_t = f_t/N_t$ ist in der vorletzten Spalte der Tabelle angegeben.
Wie man auf der Abbildung 1 sieht, ist die Fluktuation der p_t Werte sehr groß. Passt man an diese Daten die Funktion (6) an, so erhält man die Kurve

$$p_t^* = \frac{0.6265}{1 + 53.58exp(-0.4297t + 0.0115t^2)},$$

deren Werte in der letzten Spalte von Tabelle 1 zu finden sind. Für den Determinationskoeffizienten, denn wir als

$$R = 1 - \frac{\sum(p_t - p_t^*)^2}{\sum(p_t - \bar{p})^2} \qquad (9)$$

definieren, wo p_t die theoretischen Werte nach (6) sind und $\bar{p}$ die durschnittliche Proportion $\bar{p} = (1/n)\sum p_t$ ist, den Wert

Tabelle 1
Zahl der Verben mit der e-Epithese
in deutschen Texten (nach Imsiepen 1983)

Jahre t'	t	Verben mit e-Epithese f_t	Alle Verben N_t	Relative Häufigkeit p_t	Kurve (6) p_t^*
1470-79	1	20	1387	.0144	.0171
1480-89	2	53	850	.0624	.0251
1489-99	3	110	1057	.1041	.0358
1500-09	4	20	1081	.0185	.0497
1510-19	5	92	667	.1379	.0670
1520-29	6	92	815	.1129	.0875
1530-39	7	68	626	.1086	.1110
1540-49	8	48	761	.0631	.1367
1550-59	9	181	1008	.1796	.1636
1560-69	10	174	795	.2189	.1906
1570-79	11	119	786	.1508	.2167
1580-89	12	160	660	.2424	.2410
1590-99	13	149	420	.3548	.2627
1600-09	14	72	434	.1659	.2814
1610-19	15	351	1356	.2588	.2967
1620-29	16	306	1087	.2815	.3084
1630-39	17	103	523	.1969	.3166
1640-49	18	248	916	.2707	.3212
1650-59	19	313	877	.3569	.3221
1660-69	20	266	643	.4137	.3194
1670-79	21	182	848	.2146	.3132
1680-89	22	426	862	.4942	.3033
1690-99	23	212	856	.2477	.2898
1700-09	24	251	683	.3675	.2729
1710-19	25	322	822	.3917	.2527
1720-29	26	284	1267	.2242	.2297
1730-39	27	203	1076	.1867	.2045
1740-49	28	96	1004	.0956	.1778
1750-59	29	124	1008	.1230	.1506
1760-69	30	47	1003	.0469	.1241
1770-79	31	10	1102	.0091	.0995
1780-89	32	17	1064	.0160	.0773

$$R = 0.66,$$

der nicht besonders hoch ist.

Eine Glättung erzielen wir beispielsweise dadurch, daß wir immer aus jeweils drei benachbarten Werten von p_t den Durchschnitt bilden. So ergibt sich beispielsweise für $t' = 1465$

$$p = \frac{20 + 53 + 110}{1387 + 850 + 1057} = 0.0556.$$

usw. Diese Werte und die theoretische Kurve sind in Tabelle 2 angegeben. Die Zeitvariable kann wieder durch die Werte $t = 1, 2, 3, \ldots$ laufen, weil die konkreten Jahre hier zunächst keine Rolle spielen. Wir bekommen natürlich um 2 Werte weniger.

Je mehr Werte man zusammenfaßt, desto "glatter" werden die empirischen Daten und desto bessere Anpassung ist zu erwarten. Die Anpassung der Kurve (6) nach Zusammenfassung von jeweils 5 und jeweils 7 Werten ist in Tabellen 3 und 4 dargestellt. Die Kurven und die Determinationskoeffizienten lauten:

Für Glättung durch 3 Werte

$$p_t^* = 0.7030/(1 + 54.62exp(-0.4380t + 0.0124t^2))$$

$$R = 0.8579,$$

für Glättung durch 5 Werte

$$p_t^* = 0.6887/(1 + 31.72 + exp(-0.3922t + 0.0116t^2))$$

$$R = 0.8784,$$

für Glättung durch 7 Werte

$$p_t^* = 0.6597/(1 + 17.99exp(-0.3473t + 0.0109t^2))$$

$$R = 0.9117.$$

Tabelle 2
Glättung der Daten von Imsiepen:
Gleitende Durchschnitte von jeweils 3 Werten

x	p_t	p_t^*
1	0.0556	0.0190
2	0.0612	0.0279
3	0.0791	0.0399
4	0.0796	0.0553
5	0.1195	0.0744
6	0.0945	0.0970
7	0.1240	0.1225
8	0.1572	0.1502
9	0.1829	0.1788
10	0.2019	0.2073
11	0.2290	0.2343
12	0.2517	0.2590
13	0.2588	0.2806
14	0.2534	0.2985
15	0.2562	0.3125
16	0.2601	0.3224
17	0.2867	0.3280
18	0.3395	0.3294
19	0.3214	0.3266
20	0.3714	0.3195
21	0.3196	0.3083
22	0.3703	0.2929
23	0.3325	0.2737
24	0.3092	0.2511
25	0.2556	0.2255
26	0.1742	0.1979
27	0.1370	0.1692
28	0.0886	0.1408
29	0.0581	0.1137
30	0.0234	0.0891

Tabelle 3
Glättung der Daten von Imsiepen:
Gleitende Durschnitte von jeweils 5 Werten

x	p_t	p_t^*
1	0.0585	0.0305
2	0.0821	0.0428
3	0.0900	0.0583
4	0.0810	0.0772
5	0.1241	0.0992
6	0.1406	0.1239
7	0.1483	0.1503
8	0.1699	0.1775
9	0.2132	0.2044
10	0.2176	0.2300
11	0.2326	0.2533
12	0.2623	0.2738
13	0.2568	0.2909
14	0.2502	0.3044
15	0.2776	0.3140
16	0.3055	0.3198
17	0.2921	0.3217
18	0.3461	0.3196
19	0.3424	0.3135
20	0.3435	0.3036
21	0.3422	0.2899
22	0.3330	0.2726
23	0.2704	0.2519
24	0.2383	0.2284
25	0.1988	0.2028
26	0.1407	0.1758
27	0.0924	0.1486
28	0.0567	0.1223

Die Entscheidung, wie weit man bei der Glättung gehen soll, kann nicht objektiv herbeigeführt werden. Im allgemeinen reicht uns, wenn der Determinationskoeffizient über 0.8 ist. Bei 0.9 kann man das Resultat

Tabelle 4
Glättung der Daten von Imsiepen:
Gleitende Durschnitte von jeweils 7 Werten

x	p_t	p_t^*
1	0.0702	0.0487
2	0.0825	0.0650
3	0.1016	0.0843
4	0.1173	0.1064
5	0.1417	0.1306
6	0.1544	0.1560
7	0.1777	0.1819
8	0.1855	0.2071
9	0.2208	0.2309
10	0.2402	0.2524
11	0.2391	0.2712
12	0.2574	0.2868
13	0.2747	0.2991
14	0.2843	0.3079
15	0.2830	0.3132
16	0.3204	0.3149
17	0.3167	0.3131
18	0.3339	0.3077
19	0.3527	0.2987
20	0.3249	0.2863
21	0.2931	0.2706
22	0.2731	0.2517
23	0.2222	0.2301
24	0.1934	0.2063
25	0.1491	0.1811
26	0.1038	0.1552

schon als sehr gut betrachten. Man bedenke, daß die Linearisierung der Kurve und die Anwendung des F-tests keine objektivere Entscheidung beinhaltet, weil die Signifikanzebene konventionell bestimmt wird.

Literatur

Altmann, G. (1983), "Das Piotrowski-Gesetz und seine Verallgemeinerungen". In: Best, Kohlhase 1983a, 54-90.

Altmann, G., v.Buttlar, H., Rott, W., Strauß, U. (1983), "A law of change in language". In: Brainerd, B.(ed.), *Historical Linguistics.* Bochum, Brockmeyer, 104-115.

Beöthy, E., Altmann, G. (1982), "Das Piotrowski-Gesetz und der Lehnwortschatz". *Zeitschrift für Sprachwissenschaft* 1, 171-178.

Best, K.-H. (1983), "Zum morphologischen Wandel einiger deutscher Verben". In: Best, Kohlhase (1983a), 107-118.

Best, K.-H., Altmann, G. (1986), "Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen". *Folia Linguistica Historica* 7, 31-41.

Best, K.-H., Kohlhase, J. (1983), "Der Wandel von ward zu wurde". In: Best, Kohlhase 1983a, 91-102.

Best, K.-H., Kohlhase, J. (1983a), *Exakte Sprachwandelforschung.* Göttingen: Herodot.

Imsiepen, U. (1983), "Die e-Epithese bei starken Verben im Deutschen". In: Best, Kohlhase 1983a, 119-141.

Kohlhase, J. (1983), "Die Entwicklung von ward zu wurde beim Nürnberger Chronisten Heinrich Deichsler". In: Best, Kohlhase 1983a, 103-106.

Müller-Hasemann (1983), "Das Eindringen englischer Wörter ins Deutsche ab 1945". In: Best, Kohlhase 1983a, 143-160.

Tuldava, J.A. (1984), "K voprosu o modelirovanii rosta slovarja." *Acta et Commentationes Universitatis Tartuensis* 689, 147-156.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 125–130

Fast high-order monkeys and a fast algorithm for calculating high-order character entropies

Jacques B.M. Guy
Clayton (Australia)

William R. Bennett Jr. delved at length into the topic of monkeys and the meaning and calculation of the character entropy of texts in chapter 4 of his 1976 *Scientific and engineering problem-solving with the computer.* His solution, alas, calls for an n-dimensional matrix to calculate the n-th order character entropy or simulate a sample input text to the n-th order. With an alphabet of 27 characters (26 letters plus a space) one needs an array of 19,683 cells for computing the third-order entropy, and of 531,441 cells for the fourth-order entropy. As for character entropies beyond the fourth order, they require far too much time and storage to be tackled on anything but a mainframe computer.

With the algorithm presented here it takes only seconds on a personal computer to calculate the character entropy of a text to any order up to an arbitrarily set upper limit. At the same time it functions as a very fast monkey. It is a simple programming task to modify it to compute the word entropy of texts to any order and to provide an equally fast word-monkey.

I discovered this algorithm when trying to optimize for speed a class of monkeys described some forty years ago by Shannon in his *Mathematical Theory of Communication*:

> "... one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5) and (6)." [respectively third-order character, first-order and second-order word approximations] (Shannon 1949).

In the above quote Shannon describes the workings of a second-order character monkey. Shannon's monkey is computationally simple and elegant. A more general statement of it could be:

> "Let one select a string of n tokens at random from a text file. This string is recorded and stored in a variable called *the context*. The file is then accessed at random and read until a string identical to the context is encountered. The succeeding token is then recorded; it is appended to the context while the first token of the context is deleted. The file is again accessed at random, searched for this context and the succeeding token recorded, etc."

Replace "token" by "character" in the above recipe and you have a character monkey of order $n + 1$; replace "token" by "word" and you have a word monkey of order $n + 1$.

Implementing Shannon's monkey to any order is elementary programming. But the sequential search for the right context makes for very slow higher-order monkeys. In contrast with the sluggishness of Shannon's original monkeys, the algorithm described below was so fast that it generated text faster than one could read: An idle loop had to be inserted in the routine that outputs the characters selected.

The Monkey Algorithm

The algorithm is best demonstrated by an example. Say we wish to construct a fourth-order character monkey to ape this text: "abababababc".

Append $n - 1$ null characters (.) to the text: "abababababc...". Draw up a list of all eleven four-character strings contained in it:

1 abab
2 baba
3 abab
4 baba
5 abab
6 baba
7 abab
8 babc
9 abc.
10 bc..
11 c...

Sort them into alphabetical order:

1 abab
2 abab
3 abab
4 abab
5 abc.
6 baba
7 baba
8 baba
9 babc
10 bc..
11 c...

Randomly pick a string that does not contain a null character, say, "abab". This string provides the context "aba" and the first character to be output, "b". Output them both and update the context, which now becomes "bab". The first and last occurrences of the context "bab" are then searched for in the table (since strings are sorted alphabetically a binary search followed by a scan forward and backward does nicely and speedily). One occurrence of the three-character context string is then chosen randomly and the following character (the fourth character of the string selected) is output while the context is updated. The first and last occurrences of the new context are searched for, one is randomly selected, the succeeding character output, etc.

The process is eventually bound to grind to a halt when, the current context being "bab", string # 9 is selected and the succeeding character, "c", is output, as "c" only occurs at the end of the sample text. One could consider the text closed on itself, that is, that its last character is followed by its first. I prefer to have my monkey recognize when it is lost for words (or, rather, characters), mumble " ... er ..." and then start afresh.

An Even Faster Monkey

The search for the current context and the selection of one occurrence of it can be speeded up through a simple scheme: record the frequency of occurrence of each context string in the table at its place of first

occurrence, and record a negative offset to the first occurrence in the other cells. Where a string contains a null character enter a frequency of 0. Thus:

Index	Frequency of context	
1	4	abab
2	-1	abab
3	-2	abab
4	-3	abab
5	0	abc.
6	4	baba
7	-1	baba
8	-2	baba
9	-3	babc
10	0	bc..
11	0	c...

The first and last occurrences of the context being searched for are now found without carrying out a sequential search after the binary search. If the binary search yields a string with a negative frequency, then the first occurrence of it is to be found at Base Index = Current Index + Frequency at Current Index. The last occurrence is found at Index = Base Index + (Frequency at Base Index) -1.

Calculating the Entropy

The fourth-order entropy of the sample text is calculated very simply from the table above:

Index	Frequency of context (or offset)		Character in context	Contribution to total entropy
1	4	abab	b after aba	
2	-1	abab	b after aba	
3	-2	abab	b after aba	
4	-3	abab	b after aba:	4/8 * 4/4 * (-Log2(4/4)) = 0.0000
5	0	abc.	nil	
6	4	baba	a after bab	
7	-1	baba	a after bab	
8	-2	baba	a after bab:	4/8 * 3/4 * (-Log2(3/4)) = 0.1556
9	-3	babc	c after bab:	4/8 * 1/4 * (-Log2(1/4)) = 0.2500
10	0	bc..	nil	
11	0	c...	nil	
$\sum$	8			Total entropy: 0.4056

To calculate lesser-order entropies one needs only to adjust the length of the context string and to recompute the context frequencies. Thus for instance for the third-order entropy the length of the context string being two characters:

Index	Frequency of context (or offset)		Character in context	Contribution to total entropy
1	5	abab	a after ab	
2	-1	abab	a after ab	
3	-2	abab	a after ab	
4	-3	abab	a after ab:	5/9 * 4/5 * (-Log2(4/4)) = 0.1431
5	-4	abc.	c after ab:	5/9 * 1/5 * (-Log2(1/5)) = 0.2580
6	4	baba	b after ba	
7	-1	baba	b after ba	
8	-2	baba	b after ba:	
9	-3	babc	b after ba:	4/8 * 4/4 * (-Log2(4/4)) = 0.0000
10	0	bc..	nil	
11	0	c...	nil	
$\sum$	9			Total entropy: 0.4011

Remarks on the Software Implementation of the Algorithm

The sorting process being the most time-consuming part of the algorithm when long sample texts are involved, it is best to carry out a sort only once on strings of, say, twenty characters. The entropy of the text can then be calculated to any order right up to twenty, and the corresponding monkey constructed with minimal recomputations, which take but a few seconds on a personal computer. The character strings themselves are of course not stored in the table (that would consume far too much central memory); only the sample text is stored, with two-byte integers kept in the table, pointing to the starting position of the strings in the sample text. The sample text is read through a filter into a character array. The filter, which is partly specifiable by the user of the program, allows lower-case letters to be translated into upper case, characters to be disregarded (or replaced by spaces), and collapses consecutive spaces into a single space.

Bibliography

Bennett, W.R., Jr. (1976), *Scientific and engineering problem-solving with the computer.* Prentice-Hall, Englewood Cliffs, New Jersey.

Shannon, C.E., Warren, W. (1949), *The mathematical theory of communication.* Urbana, Chicago, London:University of Illinois Press.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 131–149

On the internal structure of the diskos of Phaistos text

D. Rumpel
Duisburg

0. Abstract

The diskos text is transformed into a neutral alpha-code. This code provides a broad view of the text and can also be processed by computer. By these means the "roots" and particles already isolated by Ipsen (1929) are here examined with respect to their systematic and contextual connection. It can be made plausible that the text *applies a declension* of the general form "Prefix-Root-Suffix", where the prefixes assign cases and the suffixes assign gender and number. A tentative assignment of grammatical meaning to the particles is derived.

1. Text representation

The diskos, a circular clay tablet bearing a die-stamped text on both sides (see Fig. 1), was found by Pernier (1908) in the remains of the Minoan palace of Phaistos on Crete. It is almost generally accepted that the symbols represent syllables and the frames enclose words. The reading direction of the text is controversial. (Neumann 1968)

In this paper, the diskos-text is transformed into a readable neutral code by assigning arbitrary open syllables to the characters. (The assignments are stated in Tab. 1. The open syllables were chosen so as not to coincide with the open syllables found in a "Pelasgic" or "Aegaean" language.) When the code of Tab. 1 is applied, the diskos text takes the form given in Tab. 2.

The text is read beginning from the centre. This assumption is based on the somewhat subjective observation already mentioned by Schertel (1948) and later Grumach (1962), namely that the structure of the text seems to "make some sense" when read in this direction, but

Face A

Face B

Figure 1
The diskos text

Table 1
Evans' code number, the selected alpha-code syllable and the symbol

1	Ba	10	Be	19	Bi	28	Bo	37	Bu		
2	Ca	11	Ce	20	Ci	29	Co	38	Cu		
3	Fa	12	Fe	21	Fi	30	Fo	39	Fu		
4	Ga	13	Ge	22	Gi	31	Go	40	Gu		
5	Ha	14	He	23	Hi	32	Ho	41	Hu		
6	La	15	Le	24	Li	33	Lo	42	Lu		
7	Va	16	Ve	25	Vi	34	Vo	43	Vu		
8	Xa	17	Xe	26	Xi	35	Xo	44	Xu		
9	Ya	18	Ye	27	Yi	36	Yo	45	Yu		

refuses admittance when read in the opposite direction. The reader can check the validity of this observation himself by arranging the contents of Tab. 2 backwards.

In Tab. 2, the positions of the "strokes" on the diskos are marked by bold letters, but the strokes are not considered below. The question mark indicates one or two destroyed symbols.

The arrangement of lines in Tab. 2 is arbitrary, but takes account of repetitions in the text. The resulting line structure corresponds closely to the periodic structure of the diskos-text found by W. Nahm (1979) by autocorrelation of wordlengths. Line numbers will be used

Table 2
Diskos-text in neutral alpha-code

			Face A		
1)	CuFaBe	BaGe	FiBuXoYiYiFeCa		
2)	CuFaBe	**Xo**BiHi	BaGeFeCa		
3)	FeXiGo	BiXeYeLa	YiYeHoHeYiFeCa		
4)	**Xi**GoFeCa	**Ba**Bo	YeHiBeViYiCa		
5)	**Xi**GoFeCa	HiLo	FiBuXoYiYiFeCa		
6)	**Xi**GoFeCa	**Ba**Bo	YeHiBeViYiCa	CeFu	**Cu**HiHoFeCa
7)	VaGuHuBa	XoBiHuFeCa	XoXiGo	[?]YeLaFeCa	
8)	XaXuYi	FeVaYuYi	LoGuGaFeCa		
9)	VoCoCo	**Va**YuCo	FeGuLi	**Ye**BaGeFeCa	
			Face B		
10)	**Va**Yu	ViHiVoCo	VaHiXoLaCa	VaYeFuFoYa	**Xa**VaYoCoGi
11)	LiYeHiVa	**Va**YuVa	XoYeVa	ViHiVoYi	**Xa**VaYoCoGi
12)	**Va**YuCo	GeXaCo	**Xa**VaYoCo	BaYiYaCa	LoFuHoXoLa
13)	BaLoCo	YeHeVe	XoCiLiLiCo		
14)	BaCuViYi	GuYoXiCa	XoGuLiVa	ViLuBuGi	YeBaGeVaLe
15)	LoFuBaGe	**Vu**YeHiVe	FeCiLiLo	YiViGi	
16)	**Ha**HiBuCa	XoVaYuYi	VaGuGiFeCa		

for referencing in the following analysis.

If the original text comprises only open syllables and vowels, as proposed by many scholars (see G. Neumann, 1968), and also assumed in the later part of the following analysis, then Tab. 2 already yields a replica of the basic phonetic structure of the text. Otherwise, the coding also retains its neutrality for different interpretations.

2. Analysis for repetition of particles and "roots"

The analysis of the present section is mainly a recapitulation of Ipsen's (1929) fundamental work; the differences lie in our use of the opposite reading direction and maintainance of a broad overview of the text. The analysis follows the steps:

1. Search for "roots" reappearing with different attachments in the text,
2. Identification of prefixes and suffixes,
3. Search for further words bearing the identified prefixes and suffixes.

Table 3 shows the result of step 1 applied to the text of Tab. 2.

Table 3
Appearance of "roots" in the diskos-text line in Tab. 2 — plausible roots, ... ambiguous roots

"Root"	Occurence
VaYu	(8) FeVaYuYi, (9) VaYuCo, (10) VaYu, (11) VaYuVa, (12) VaYuCo, (16) XoVaYuYi
ViHiVo[1]	(10) ViHiVoCo, (11) ViHiVoYi
XaVaYoCo	(10) XaVaYoCoGi, (11) XaVaYoCoGi, (12) XaVaYoCo
BaGe	(1) BaGe, (2) BaGeFeCa, (9) YeBaGeFeCa, (14) YeBaGeVaLe, (15) LoFuBaGe
XiGo	(3) FeXiGo, (4) XiGoFeCa, (5) XiGoFeCa, (6) XiGoFeCa, (7) XoXiGo
GuLi	(9) FeGuLi, (14) XoGuLiVa
XoLa	(10) VaHiXoLaCa, (12) LoFuHoXoLa
YeHi	(3) YeHiBeViYiCa, (6) YeHiBeViYiCa, (11) LiYeHiVa, (15) VuYeHiVe
CiLi	(13) XoCiLiLiCo, (15) FeCiLiLo
YeLa	(3) BiXeYeLa, (7) [?]YeLaFeCa
VaGu	(7) VaGuHuBa, (16) VaGuGiFeCa

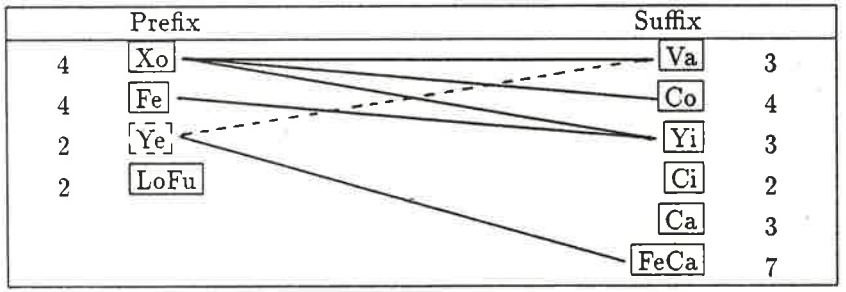
[1] Preferred against VoCo: (9) VoCoCo,(10) ViHiVoCo

Scanning table 3, it becomes obvious that most of the leading or following attachments to the "roots" do not appear at random, but show a repetitive pattern, as characteristic for particles with a specific sense.

In step 2, by applying the criterion that a suspected particle must occur at least twice in characteristic positions in Tab.3, prefixes and suffixes are identified. This criterion is not sufficient to identify all particles which can possibly occur, but provides a reasonable certainty that the identified particles are truly a subset of the particles which occur.

Identified particles are entered in Table 4. The numbers attached show the frequency of their occurence in Table 3. In the table, the entries are marked for their type of appearance: Singly (encased), and in prefix/suffix combinations (connection lines).

Table 4
Identified particles



From Tab. 4 it can be seen that more than half of the particles display the ability to appear alone, as well as in combination with other particles of the opposite kind (prefix or suffix). Such a general structure

Prefix – "Root" – Suffix,

in which prefixes and suffixes are combinable and may be missing in individual cases, is an obvious feature of the diskos-text. This is illustrated by Tab. 5, which contains the diskos text coded as in Tab. 2. This time the identified particles are marked by enclosure in square brackets; the "roots" are underlined according to Tab. 3.
The particle identification process recognizes only prefixes and suffixes according to the criterion of their initial or final position in the word, but neglects other possible particle positions. Scanning the text of Tab. 5 for occurrences of the code of an identified particle within words,

Table 5
Diskos Text with markers

	Face A
1)	CuFaBe BaGe FiBuXoYiYi[FeCa]
2)	CuFaBe [Xo]BiHi BaGe[FeCa]
3)	[Fe]XiGo BiXeYeLa YiYeHoHeYi[FeCa]
4)	XiGo[FeCa] BaBo YeHiBeViYi[Ca]
5)	XiGo[FeCa] HiLo FiBuXoYiYi[FeCa]
6)	XiGo[FeCa] BaBoYeHiBeViYi[Ca] CeFu CuHiHo[FeCa]
7)	VaGuHuBa [Xo]BiHu[FeCa] [Xo]XiGo [?]YeLa[FeCa]
8)	XaXu[Yi] [Fe]VaYu[Yi] LoGuGa[FeCa]
9)	VoCo[Co] VaYu[Co] [Fe]GuLi [Ye]BaGe[FeCa]
	Face B
10)	VaYu ViHiVo[Co] VaHiXoLa[Ca] VaYeFuFoYa XaVaYoCo[Gi]
11)	LiYeHi [Va] VaYu[Va] [Xo]Ye[Va] ViHiVo[Yi] XaVaYoCo[Gi]
12)	VaYu[Co] GeXa[Co] XaVaYoCo BaYiYa[Ca] [LoFu]HoXoLa
13)	BaLo[Co] [Ye]HeVe [Xo]CiLiLi[Co]
14)	BaCuVi[Yi] GuYoXi[Ca] [Xo]GuLi[Va] ViLuBu[Gi] [Ye]BaGeVaLe
15)	[LoFu]BaGe VuYeHiVe [Fe]CiLi Lo YiVi[Gi]
16)	HaHiBu[Ca] [Xo]VaYu[Yi] VaGuGi[FeCa]

"roots", or in opposite positions, supplies the examples in Tab. 6. The relevant positions are underlined.
Table 6 has a large number of such examples, a number which is increased by the fact that many words include more than one particle code in a non-particle position.

Assuming that the particles have a special meaning, still leaves the following possibilities open:

a) that the particles are semantic determinatives which are phonetically mute, or bear no uniquely assigned phonetic value. Then it must be expected that the particles mainly retain their special

Table 6
Particle codes in non-particle positions

Prefix	
Xo:	1) and 5) FiBuXoYiYi[FeCa], 10) VaHiXoLa[Ca], 12) [LoFu]HoXoLa
Fe:	none, except [FeCa]
Ye:	3) BiXeYeLa, YiYeHoHeYi[FeCa], 4) and 6) YeHiBeViYi[Ca], 7) [?]YeLa[FeCa], 10) VaYeFuFoYa, 11) LiYeHi[Va], [Xo]Ye[Va], 15) VuYeHiVe
LoFu:	none
Suffix	
Va:	7) VaGuHuBa, 8) [Fe]VaYu[Yi], 9) VaYu[Co], 10) VaYu, VaHiXoLa[Ca], VaYeFuFoYa, 10) and 11) XaVaYoCo[Gi], 11) VaYu[Va], 12) VaYu[Co], XaVaYoCo, 14) [Ye]BaGeVaLe, 16) [Xo]VaYu[Yi], VaGuGi[FeCa]
Co:	9) VoCo[Co], 10) and 11) XaVaYoCo[Gi], 12) XaVaYoCo
Yi:	1) and 5) FiBuXoYiYi[FeCa], 3) YiYeHoHeYi[FeCa], 4) and 6) YeHiBeViYi[Ca], 12) BaYiYa[Ca] 15) YiVi[Gi]
Gi:	16) VaGuGi[FeCa]
Ca:	none, except [FeCa]
FeCa:	none

meaning, independent of their position in the word.

b) that the particle-code only expresses a phonetic value, and that the special meaning is coupled to the phonetic value at a special position in the word.

The high number of variegated occurrences in Table 6 can only be explained by possibility b) and, in general, rules out possibility a). A determinative hypothesis, because of the lack of counterexamples, may still be sustained for the particles [Fe], [LoFu], [Ca] and [FeCa], but is improbable, especially for the particles coded by double-symbols.

There are only four cases in Table 6 which call for closer inspection. For XaVaYoCo the text-sequence

12) VaYu[Co] GeXa[Co] XaVaYoCo,

by correspondence of the end-syllables, indicates that also the last Co has a particle-status. The search process, however, by means of comparison with 10) and 11) XaVaYoCoGi, assigned this Co to the "root". In fact, XaVaYoCo ⟶ XaVaYo[Co] is made plausible, and XaVaYoCo[Gi] would attain the structure XaVaYo[Co][Gi], with a double suffix. This may also apply to 14) [Ye]BaGeVaLe ⟶ [Ye]BaGe[Va]Le(?). Further particles, in rare cases, might be embedded in composite words, as possibly 1) and 5) FiBuXoYiYi[FeCa] ⟶ FiBu[Xo]YiYi[FeCa](?), or in the peculiar word [Xo]Ye[Va] ⟶ [Xo][Ye][Va](?).

Step 3, now, comprises the task of collecting, from the text, all words which bear the identified particles, or contain nude "roots" identified by their appearance together with particles in other occurrences.

Table 7 shows these words arranged in a matrix according to prefix and suffix-attachment. The three best proven "roots" VaYu, BaGe and XiGo are enclosed in a box, while the other "roots" are marked as in Tab. 3.

3. Grammatical observations

The 45 examples of the root/particle structure in Tab.7 (due to repetitions in the text) cover 51, or 83%, of the 61 words making up the diskos text. If we rule out the faint possibility that the author of the text, in order to achieve some (humorous?) effect, has deliberately chosen words which by chance display the structure in question, then it must be a structure inherent in the language; i.e. a grammar or, in respect to words, an inflection.

We can expect an inflection to show some sort of declension and conjugation. Verbal forms have a low occurrence and cannot be readily assumed to cover 83% of the words of an extended text. If the search-process of part 2 has hit on verbal forms at all, they must have been outnumbered by nouns and adjectival forms. Therefore, the matrix of Tab. 7 must in essence represent a declension, not a conjugation.

Assuming a declension, we can expect to find inflections for the three independent dimensions: gender,number and case. With regard to the general structure derived above (and for the moment disregard-

Table 7
Matrix of Diskos-words according to prefix/suffix attachment

Suffix:	none	Va	Co	Vi	Gi	Ca	FeCa
Prefix:							
	VaYu	VaYu Va	VaYu Co	ViHiVoYi	XaVaYoCoGi	YeHiBeViYiCa	BaGe FeCa
		LiYeHiVa	ViHiVoCo	BaCuViYi	ViLuBuGi	VaHiXoLaCa	XiGo FeCa
	BaGe		GeXaCo	XaXuYi	YiViGi	BaYiYaCa	[?]YeLaFeCa
none			BaLoCo			GuYoXiCa	VaGuGiFeCa
			VoCoCo			HaHiBuCa	FiBuXoYiYiFeCa
			XaVaYoCo				YiYeHoHeYiFeCa
							CuHiHoFeCa
							LoGuGaFeCa
	XdXiGo	XoGuLiVa	XoCiLiLiCo	Xo VaYu Yi			XoBiHuFeCa
Xo	XoBiHi	XoYeVa					
	FdXiGo			Fe VaYu Yi			
Fe	FeGuLi						
	FeCiLiLo						
	Ye BaGe	VaLe?					Ye BaGe FeCa
Ye	YeHeVe						
	LoFu BaGe						
LoFu	LoFuHoXoLa						

ing the special case of a probable double suffix), there are only two particle positions per word to convey the sense of the three dimensions. In order to ensure a free combinability of the particles, we therefore have to assume that one particle position – either the prefix or the suffix – consistently bears one dimension, and the other one the residual two dimensions. There is no other way to distribute 3 over 2 as integers.

The matrix of Tab. 7 has 5 lines and 7 columns. The set of gender distinctions taken alone, or of number-distinctions taken alone, cannot readily be expected to amount to 5 or 7. The number of case-distinctions alone, on the other hand, could well attain these values. The most plausible assumption, therefore, is that the particle in the position with one dimension expresses case, the other gender and number in combination.

The matrix shows occurrences with no prefix (29 examples, upper line) and no suffix (11 examples, left column). This would mean that these examples either carry "no case", or "no gender/number-combination". "No gender/number" could for example specify a (masculine) singular whereas "no case" could characterize both the nominative and the accusative case. Only the latter assumption offers a reasonable explanation for the high preponderance of the examples in the upper line of the matrix, and at the same time indicates that the lines (prefixes) represent cases. Then the suffixes have to be assumed to define the gender/number combination.

It should be emphasized that up to this point of the paper the analysis is neutral as to the reading-direction. It could as well be applied on the text in inverted direction and would result in a hypothetical language which first defines gender/number by a prefix, and then case by a suffix at the end of the word.

4. Stylistic observations

In Table 2 and 5, the diskos-text was put into lines regarding repetitions in the text. These repetitions are obviously much more frequent than would be expected in a narrative text. On the other hand, they display a greater grammatical and positional variability than can reasonably be expected from a listing. They therefore have to be assumed to be of a poetic intention.

Face A starts with a very rhythmically designed, probably invocatory part in lines 1) to 6), broken only by line 3) and followed by a short and less elaborate part comprising lines 7) to 9). Roughly the same structure is visible in face B, where lines 10) and 11) and possibly 12), by repetitions of VaYu, VaYu[Va], VaYu[Co], ViHiVo[Co], ViHiVo[Yi] and XaVaYo[Co][Gi], XaVaYo[Co], display a hymnal style, whereas the following lines return to plain text (see also Schertel 1948).

From Tab. 2 it is obvious that the text-lines, derived on the basis of repetitions, show a preference for a length of 3 words on face A, whereas face B prefers 5 words. This observation is supported by the findings of W. Nahm (1979), whose autocorrelations also show a periodicity of 3 words for face A and 5 words for face B. Nahm, actually, based his studies on the opposite reading direction of the Diskos-text, but his results from autocorrelation are independent of the reading direction.

Considerable differences between the two faces also appear in the distribution of "roots" and particles. In order to make this obvious, Table 8 shows the Diskos-text once more; this time the frequent "roots" BaGe, XiGo, VaYu and the suffixes FeCa and Gi are *enhanced*. The "root" XiGo is only apparent on face A, where it clusters in the central part. VaYu is preponderant on face B, whereas BaGe is almost evenly distributed between the two faces. The suffix Gi is a speciality of face B; FeCa is highly preponderant on face A.

If we assume the "roots" to denote notions and the particles to denote grammatical relations, then the "root"-distribution indicates a certain change in topics between the two sides, but not a total break, as there is a considerable overlap in the BaGe- and VaYu-distribution.This change in topics seems to be related to an alteration of the grammatical expressions (as indicated by the suffixes) and of the poetical form (as indicated by the number of words per line).

The most frequent "root" VaYu appears 4 times on face B (Line 10,11,12,16) in the "hymnal" part as well as in the plain text; it also appears twice (Line 8,9) in the short plain text of face A. The zero form of this "root" keeps the initial position of face B, the repetitions show a high flexional variability. A reasonable assumption to cover these phenomena is that VaYu in the pronominal form, e.g. "I" or "he", denotes a main person dealt with on the diskos. Inflections (also possessive inflections) "me, my" or "him, his" would then be expressed by

Table 8
Distribution of the "roots" BaGe, XiGo, VaYu and of the suffixes FeCa and Gi in the Diskos text

	Face A						
1)	CuFaBe	BaGe	FiBuXoYiYi	FeCa			
2)	CuFaBe	XoBiHi	BaGe	FeCa			
3)	Fe XiGo	BiXeYeLa	YiYeHoHeYi	FeCa			
4)	XiGo	FeCa	BaBo	YeHiBeViYiCa			
5)	XiGo	FeCa	HiLo	FiBuXoYiYi	FeCa		
6)	XiGo	FeCa	BaBo	YeHiBeViYiCa	CeFu	CuHiHo	FeCa
7)	VaGuHuBa	XoBiHu	FeCa	Xo XiGo	[?]YeLa	FeCa	
8)	XaXuYi	Fe VaYu	Yi	LoguGa	FeCa		
9)	VoCoCo	VaYu	Co	FeGuLi	Ye BaGe	FeCa	
	Face B						
10)	VaYu	ViHiVoCo	VaHiXoLaCa	VaYeFuFoYa	XaVaYoCo	Gi	
11)	LiYeHiVa	VaYu Va	XoYeVa	ViHiVoYi	XaVaYoCo	Gi	
12)	VaYu Co	GeXaCo	XaVaYoCo	BaYiYaCa	LoFuHoXoLa		
13)	BaLoCo	YeHeVe	XoCiLiLiCo				
14)	BaCuViYi	GuYoXiCa	XoGuLiVa	ViLuBu	Gi	Ye BaGe VaLe	
15)	LoFu BaGe	VuYeHiVe	FeCiLiLo	YiVi	Gi		
16)	HaHiBuCa	Xo VaYu	Yi	VaGuGi	FeCa		

the attached particles. Related to these phenomena is the appearance of the suffix [Gi] four times on face B only (Line 10,11,14,15), where it is present in the "hymnal", as well as in the plain text part. [Gi] could easily fulfil the verbal notation of what the main person "is" or "has done". This assumption will also explain the double-suffix word XaVaYo[Co][Gi] as e.g. "Something-adjectival-he-is".

If these observations are interpreted correctly, then face B would contain a solemnly introduced report on a feat of the main person, face A could be a corollary prayer or hymn invoking deities (BaGe, BaGe[FeCa], also the double YiYi in the third word of line 1 is interesting in this respect) on that occasion.

To start the text with face B would give the initial VaYu its rightful position.

5. Tentative assignment of grammatical sense

The text (Table 5) contains several examples of word sequences with corresponding suffixes, as in lines:

8) XaXu[Yi] [Fe]VaYu[Yi]
9) VoCo[Co] VaYu[Co]
11) LiYeHi[Va] VaYu[Va] [Xo]Ye[Va]
12) VaYu[Co] GeXa[Co] XaVaYo[Co]
13) BaLo[Co]____[Xo]CiLiLi[Co]
14) [Xo]GuLi[Va]____[Ye]BaGe[Va]Le?

This relatively frequent correspondence of suffixes is the best candidate for the denotation of adjectival reference. This would mean that [Yi],[Co] and [Va] are adjectival particles, and that adjectival reference in the diskos-language couples only by gender/number, but not by case.

A sequence of corresponding prefixes[1] is only present in line

7) [Xo]BiHu[FeCa] [Xo]XiGo.

If adjectival reference does not require a coincidence of cases, the most probable assumption for this coincidence is a chain of genitives. Thus, the prefix [Xo] must be suspected to denote the genitive case.

For assigning cases to the remaining prefixes [Fe], [Ye] and [LoFu], no clear help can be found in the text, but we still can experiment with the full text and look for the best overall fit.

In this way, the assignments for the next frequently expected cases

[Fe] = dative case
[Ye] = instrumental case

[1] Line 10) VaHiXoLa[Ca] VaYeFuFoYa is not regarded, because initial Va is not established to be a particle.

were established. The relevant passages are contained in lines 3), 8), 9), 13), 14) and 15).

In line 3) [Fe]XiGo, taken as a dative, would represent a good object, to whom the probable deities, invoked in lines 1) and 2), are to give something. Taken as an instrumental, it makes no special sense.

If the stylistic interpretation, as given in section 4, is correct, then line

9) VoCo[Co] VaYu[Co] [Fe]GuLi [Ye]BaGe[FeCa],

refers to the main person, as well as to the deities, and is the closing of the whole text. For such a clause, "His something to something (the land?) by the deities" makes better sense than "His something by something (the priest?) to the deities"

By the assignment [Ye] = instrumental, the deities also appear instrumental in line 14), which is embedded in the probable "report of a feat" of the main person. This seems reasonable, especially if a warlike action is being reported. The assignment of [LoFu] remains open.

Tentative assignments of gender/number to suffixes mainly depend on the interpretation of VaYu, as given in section 4. If we assume the main person to be male, then the first two words of face B: VaYu ViHiVo[Co] could contain a meaning as "I or he, the king/ ruler" which would render ViHiVo[Co] an adjectival noun and [Co] the adjectival male/singular particle. A plausible female counterpart is found in line 11) as ViHiVo[Yi], which means [Yi] is the adjectival female/(probably) singular. For the third adjectival candidate [Va], no plausible gender/number combination could be derived from the text. It may be a quite general coupling particle.

The suffixes [Gi], [Ca] and [FeCa] which do not show obvious coupling properties, remain. From its formal position in XaVaYo[Co][Gi] (see section 2) [Gi] is suspected to be of a special kind; for stylistic reasons, it is suspected to be verbal (see section 4). Thus, [Ca] and [FeCa] remain for the largely absent (male and female) plurals. No formal reason for the final assignment can be found in the text, but the archaeological evidence shows that almost all Minoan deities were female. If, therefore, the stylistic interpretation of an invocation of the deities is valid for the first lines of face A, then BaGe[FeCa] is very likely to be a female plural. This leads to the assignment

Table 9
Diskos-text with grammatical equivalents inserted

	Face A
1)	CuFaBe BaGe FiBuXoYiYi[ae][1]
2)	CuFaBe [of]BiHi BaGe[ae]
3)	[to]XiGo BiXeYeLa YiYeHoHeYi[ae]
4)	XiGo[ae] BaBo YeHiBeViYi[i]
5)	XiGo[ae] HiLo FiBuXoYiYi[ae][1]
6)	XiGo[ae] BaBo YeHiBeViYi[i] CeFu CuHiHo[ae]
7)	VaGuHuBa [of]BiHu[ae] [of]XiGo [?]YeLa[ae]
8)	XaXu[sa] [to]VaYu[sa] LoGuGa[ae]
9)	VoCo[sus] VaYu[sus] [to]GuLi [by]BaGe[ae]
	Face B
10)	VaYu ViHiVo[sus] VaHiXoLa[i] VaYeFuFoYa XaVaYo[sus][Gi]
11)	LiYeHi[Va] VaYu[Va] [of]Ye[Va][2] ViHiVo[sa] XaVaYo[sus][Gi]
12)	VaYu[sus] GeXa[sus] XaVaYo[sus] BaYiYa[i] [LoFu]HoXoLa
13)	BaLo[sus] [by]HeVe [of]CiLiLi[sus]
14)	BaCuVi[sa] GuYoXi[i] [of]GuLi[Va] ViLuBu[Gi] [by]BaGeVaLe[3]
15)	[LoFu]BaGe VuYeHiVe [to]CiLiLo YiVi[Gi]
16)	HaHiBu[i] [of]VaYu[sa] VaGuGi[ae]

[1] FiBu[of]YiYi[ae]?
[2] [of][by][Va]?
[3] [by]BaGe[Va]Le?

[Ca] = male plural
[FeCa] = female plural

The text does not enforce the introduction of more than two genders.

To reintroduce the tentative grammatical assignments visibly into the text, the prefixes were replaced by their equivalents in English

[Xo]=[of], [Fe]=[to], [Ye]=[by], none = none

and the suffixes by their Latin equivalents:

[Co]=[sus], [Yi]=[sa], [Ca]=[i], [FeCa]=[ae]

The resulting text is given in Tab. 9. For face A, regarding footnote 1, an apparently consistent grammatical sequence is achieved, wherein 7) [?]YeLa[ae] and 8) LoGuGa[ae] seem to be participal constructions referring to 4), 5), 6) XiGo[ae]. Face B contains three ambiguities as for 12) BaYiYa[i], 14) BaCuVi[sa] GuYoXi[i] and 16) [of]VaYu[sa] VaGuGi[ae].

6. Further work

If Linear A is in the same language as the diskos, the same grammatical structures, with possibly the tentative assignments suggested in this paper, should be present.

Parallelizing Linear A includes the problem of parallelizing diskos-symbols to Linear A characters, which might in turn profit from parallelizing structures. On the other hand, the short listings of Linear A might be rather poor in grammar.

At present, the author can give only one example:

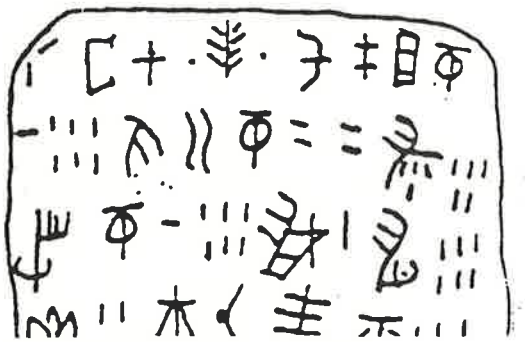


Figure 2
Clay table HT 116, beginning

Fig. 2 shows the beginning of a list with the following structure:

Caption – **isolated character "Branch" (transaction sign)**

– Word – Ware determinans –Number
– Word – Ware determinans –Number
– Ware determinans –Number
– Word – Ware determinans –Number
– Ware determinans –Number
– Ware determinans –Number
etc.

The interesting feature is the isolated Linear-A character "Branch". We can tentatively parallelize this character with the diskos-symbol "Branch" = 35 = Xo. Xo again is the code for the proposed genitive particle [of].

Introducing this hypothesis into the listing gives a clear fit, and at the same time an explanation for the isolated state of the particle. It allows an interpretation of the listing as

Contributions [of]:

Name – Ware –Quantity
Name – Ware –Quantity
– Ware –Quantity
Name – Ware –Quantity
– Ware –Quantity
– Ware –Quantity
etc.

Agglutinating the genitive particle to the names would have required unnecessary repetitions and possibly obscured the reading of the names.

Literature

Faucounau, J. (1975), "Le sens de l'ecriture du Disque de Phaistos". In: *Kadmos*, 14, 94-101.
Grumach, E. (1962), "Die Korrekturen des Diskos von Phaistos". In: *Kadmos*, 1, 1-26.
Ipsen, G. (1929), "Der Diskos von Phaistos. Ein Versuch der Entzifferung". In: *Indogermanische Forschungen*, 47, 1-41.
Marinatos, S. (1935), "Ausgrabungen und Funde auf Kreta 1934-1935". In: *Archäologischer Anzeiger*, 244-259.

Nahm, W. (1979), "Zum Diskos von Phaistos II". In: *Kadmos*, 18, 1-25.
Neumann, G. (1968), "Zum Forschungsstand beim "Diskos von Phaistos". In: *Kadmos*, 7, 27-44.
Pernier, L. (1909), "Il Disco di Phaestos con caratteri pittografici". In: *Ausonia*, 3, 255-304.
Schachermeyr, F. (1979), *Die minoische Kultur des alten Kreta*. Berlin, Köln, Mainz: Kohlhammer.
Schertel, E. (1948), "Der Diskos von Phaistos Wege zu seiner Entzifferung". In: *Würzburger Jahrbücher für die Altertumswissenschaft*, 3, 334-365.
Schürr, D. (1973), "Der Diskos von Phaistos und Linear A". In: *Kadmos*, 12, 6-19.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 151–173

The Phaistos Disc and the Devil's Advocate

On the Apories of an Ancient Topic of Research[1]

Michael Trauth

I. Introduction

When Luigi Pernier found an unconspicuous, chalk-incrusted terracotta disc in the ruins of the Minoan palace at Phaistos on Crete in the year 1908, he was hardly aware of having opened the flood-gates to numerous and thoroughly controversial studies of this artefact. To very few documents of the early Greek world has as much attention (within and outside the circle of experts) been paid as to the the 'Phaistos Disc',[2] which is covered with unidentified symbols. And once more, as is so often the case, the intellectual challenge of a seemingly unsolvable riddle has inspired competent and less competent investigations to highly heterogenous attempts at solution – efforts (comparable to the decoding of the hieroglyphics) ranging from meticulous attempts at revealing basic knowledge to poor speculation. This state of affair has opened up an avenue of approach to the debate leading to a rich field of activity for any self-appointed *advocatus diaboli*:[3] examinations

[1] The present contribution has come into being at to the kind instigation of Prof. Dr. Burghard Rieger. I am indebted to my colleagues from the special fields of Ancient History, History of Philology and Language Statistics, whose readiness to offer advice has been of great assistance. In particular I want to thank my friends Heinrich P. Delfosse and Robert Hübner for their kind help and many valuable suggestions. I am grateful to Mrs. Frauke Birtsch, cand. phil. for her patient translation of my article.

[2] Ernst Grumach counted more than 250 publications on this topic until 1965 (Grumach 1963); since then approximately seventy more have been added.

[3] It is little known that 'advocatus diaboli'is a term – only used in jest – for the 'promotor fidei' in the canonization of saints in the Catholic church: It is his task to articulate any points mitigation against the canonozation. Thus, the term has gained currency for any destruction carried out *ex officio*.

which, for lack of alternatives, operate on the basis of a stock of dispersed 'possibilities' and thus reach their results in an inductive, often not canonical by methodical way, and much too easily fall prey to the temptation of "schoolmasterly" criticism. Such criticism is often burdened with the hypothecation of sterility, because the central problem, the lack of assured surrounding knowledge, puts implicitly into question the Archimedian point of differentiation between mere speculation and serious progress in understanding.

This reservation seems necessary to me as a preliminary remark in order to leave no doubt about the complexity of what follows. My exposition is meant to be read in conjunction with the subsequent contribution by Dieter Rumpel, "On the Internal Structure of the Diskos of Phaistos Text", an outline of the state of research as well as an attempt to point out the problematic aspects of the results achieved. Rumpel has made a fresh attempt to extract plausible grammar elements from an examination of the text of the Disc – a goal requiring certain axiomatic assumptions about the nature of the text. From a pragmatic point of view this doubtlessly is a sensible procedure – and one which is also supported by the surprising results *ex eventu* – although it will be made more transparent and be placed in a broader context below for the reader who is little familiar with the many open questions concerning the Disc. Without opposing Rumpel's linguistic results with an antithesis of my own, it is merely my intention as an 'advocatus diaboli' to point out the existence of several alternatives to the axiomatic assumptions mentioned, which cannot be refuted easily and, as long as they are not fully removed, any linguistic analysis of the Phaistos Disc can only be a preliminary one. I shall discuss in particular the question of how narrowly the historical context to which the Disc belongs can be determined (II.1), in addition I shall give a resumé of the previous attempts at dating the Disc and determining its place of origin (II.2). On the basis of a description of the Disc (II.3), I shall examine the type of characters used – as to whether they are ideographic or represent syllables – (II.4), I shall then proceed to sketch the evidence supporting a left-to-right or a right-to-left reading (II.5), and shall finally attempt to determine the function of the 'strokes' on the Disc (II.6).

II. On the Present State of Research[4]

(1) Archaeological and historical context

Often the place of an epigraphic discovery hints at its purpose and content. That does not hold true in this case: Pernier came across the Disc in building 101 at the north-eastern periphery of the palace in a small rectangular chamber in the basement measuring less than four square meters. The excavations also brought to light the fragment of a Linear A tablet, a few broken pieces of stone and clay pots, as well as cinders and remainders of coal, all from the same doorless room, which was only accessible from above.[5]

Neither these findings nor the few lesser objects which were unearthed from the second and larger room of 101 allow us to draw definite conclusions about the function of the building. The same applies to the seven pits in that second room which, despite certain similarities with supposed 'temple' repositories of the palaces of Knossos and Kato Zakro, cannot conclusively be classified as such, especially since in contrast to the latter, nothing indicates the storage of objects of worship. Merely the depository character of the house can be taken as certain; all further rubricating – whether as (secular) archives of the palace, as Pernier himself supposed, or as a repository for objects of worship – can only be of use as a working hypothesis.

Placing the Disc into an archeological context which would further characterize its meaning is therefore impossible. It is true that researchers on the subject have from the beginning deduced from the omnipresence of sacred elements in the Minoan palaces the existence of a priest-kingdom. Thus the assumption of a religious surrounding

[4] Yves Duhoux provides a detailed introduction into the up-to-date archaeological and epigraphic accomplishments (Duhoux 1977). Günter Neumann's contribution focuses on the problems of research which are concerned with the history of language (Neumann 1968). It is his special merit to have refuted earlier speculative attempts at decoding the Disc and to have pointed out methodically correct ways. Most recently Duhoux has published a cautious linguistic analysis of the Disc on the basis of a comparison with Linear A (Duhoux 1983).

[5] Pernier himself has given a description of the discovery (Pernier 1908, especially p. 261sq).

for the place of discovery is not unlikely. Nonetheless a second intermediate conclusion would be required as a prerequisite for basing the linguistic examination of the Disc on a cultic content. As temple archives frequently house non-sacred documents (such as contracts) this additional conclusion would be even more daring than the first. Notwithstanding these doubts, even on the part of competent scholars the opinion is held that the Disc text 'probably' has a religious content.[6]

Even the general association of the Disc with Crete and the Minoan culture has not always been undisputed. In the meantime isolated earlier attributions to North African cultures (Macalister) or to an origin in Asia Minor (Eduard Meyer, Evans) have been dismissed by scholarship. Aside from the 'Aegean character' of the Disc, which was noted earlier (Ipsen 1929, especially p. 4) though it is not easily verified, there are remarkable and hardly accidental similarities between some of the symbols to be found on the Disc and other relics of the Minoan culture.[7] Besides giving a survey of already known parallels, Neumann refers to a symbol on the bronze axe of Arkalolochori related to the "feather decorated head" and points out similarities between the Disc writing and the hieroglyphics on the stone altar of Mallia. I should like to add that the symbol of the "angle iron" on the Disc is to be found on the Arkalochori axe as well. Costis Davaras shows further parallels to later discoveries on Crete (Davaras 1981). Dieter Rumpel additionally notes another possibly quite important fact: The museum of Hagios Nikolaos houses the *"Kopf einer Adorantin aus dem mittelminoischen Gipfelheiligtum von Petsophas"* that shows the same oddly streaky hair-dress as is found in the pictogram "woman" on the Disc[8] which even when taken itself leave little doubt about the Cretan origin of the Disc. Moreover, a couple of striking resemblances between the writing on the Disc and Linear A are a strong argument for their affinity.[9] Taking all these aspects into consideration, Crete as the source of the

[6] Among the younger authorities, especially Duhoux supports this hypothesis by characterizing the Disc text as *"homogène, de type littéraire, à sujet probablement religieux"* (Duhoux 1983, p. 34).

[7] Cf. Neumann 1968, p. 28-30.

[8] Quotation and illustration in Davaras (undated), Ill. 41.

[9] Cf. the synopsis of eleven tokens in both writing systems by Duhoux (Duhoux 1983, p. 4). Duhoux states: *"Bien que différentes, les écritures du linéaire*

Disc can no longer be called into question.

(2) Dating

Concerning the dating of the Disc, which is difficult due to insufficient evidence, I should like to refer to the thorough resumé by Duhoux (Duhoux 1977, p. 6-14). From the archaeological context of both the place of discovery and the other objects found in house 101, Duhoux draws the well-balanced conclusion that the Disc was *deposited* (!) at the excavation site during the periods of MM II[10] to MM IIIb, about 1850 to 1600 B. C. As should be obvious, this is a *terminus ante quem* – the Disc could have an earlier origin, too. Although the time frame thus determined is fairly broad not even this point can be taken for certain: Kristian Jeppesen for example dates the Disc with some justification between MM IIb and the end of LM III, thus between about 1600 and 1100 B. C. (Jeppesen 1962, especially p. 161), yet Günter Neumann supplementarily points out that *"the Disc, because of its highly similar writing to the one on the axe of Arkalochori* [...] *could not be dated much earlier"*.[11] The simultaneous usage of a pictographic script together with Linear A as a fully developed current writing seems a peculiar archaism. An explanation could be found in the common opinion that the Disc text has a religious content: traditionalisms are a determinative trait of every religion.

(3) Description of the Disc and its Characters

The Disc consists of a round, not quite even terra-cotta disc, with a diameter of 15.8 to 16.5 cm and a thickness of 1.6 to 2.1 cm. The

A et du disque de Phaestos sont structurellement identiques. [...] *Mais ces exemples, isolés et de valeur très inégale, ne suffisent pas à identifier les deux systèmes graphiques."*

[10] Since Arthur J. Evans, the periods of the Minoan history have been divided up into Early-Minoan (EM I-III: about 2600 to 2000 B. C.), Middle-Minoan (MM I-III: about 2000 to 1600/1550 B. C.) and Late-Minoan (LM I-III: about 1600/1550 to 1200/1100 B. C.).

[11] Cf. Neumann 1968, p. 27. Spyridon Marinatos describes the rich discoveries made in the Arkalochori cave (Marinatos 1935); Concerning the axe itself which is dated by Marinatos into the period of transition from MM III to LM I, meaning about the 16^{th} century B. C., cf. ibid. col. 252-254.

material used is very fine clay which was fired with care. On either side of the disc a spiralled line functioning as a support for the characters leads from the centre to the rim. (Following Pernier, usually the side bearing a rosette in the centre is designated as side A.) The characters themselves are divided into 'groups' of two to seven symbols by vertical lines. Altogether there are 241[12] tokens consisting of 45 different types; side A contains 122 of the tokens mentioned in 31 groups, and side B 119 tokens in 30 groups. (For an illustration, see Fig. I of Dieter Rumpel's contribution.)

The symbols themselves are distinctly pictographic; the picture represented can mostly be distinguished with ease ("feather decorated head", "fish", "pigeon", "running man", etc.; cf. the complete catalogue in Table I of Rumpel's article), even though, despite the clarity of the image, the 'semantics' of some pictures remain hidden.

I shall not further discuss the question of the 'printing technique', which is not here addressed in detail either, since it is of secondary importance to a linguistic examination. It is probable that the symbols were pressed by hand into the wet clay with wooden stamps.

(4) The Type of Characters

Another significant and open question, the elimination of which must precede any linguistic approach to the text is the system of writing used.

Only the type of character-writing we are familiar with today can with some justification be rejected. Two arguments are usually adduced: the first regards the total number of types (45) in the Disc text as too high, but its conclusiveness suffers from the fact that alphabets are not necessarily restricted to the stock of types to be found in our Latin character system, which often is imprudently taken as a

[12] I am refering to the count by Duhoux (Duhoux 1977, p. 36). This count however has not remained uncontradicted, since some minor damage on the surface of the Disc renders a clear reading impossible: For side A, 122 to 124 tokens have been counted, and for side B 118 to 119. Neumann in particular, pleaded for 123 and 119, i. e. 242 in total (Neumann 1968, p. 27).

standard.[13] In comparison, the second argument is more convincing: since we are not familiar with any parallel to the degree of abstraction demanded by an alphabetic script at the assumed time of origin of the Disc, a Minoan 'solo attempt' is most improbable.

The distinct 'picture value' of the types leads to their classification as 'ideograms':[14] An ideographic writing system where each word is represented by a single picture can be considered plausible if one dates the Disc as Duhoux does, between the nineteenth and the seventeenth centuries B. C. The groups of tokens mentioned under (3) can be interpreted as 'sentences', or – more generally – as syntagms in this case. Nonetheless, two aspects contradict the assumption of a purely ideographic system: On the one hand the sum of different characters in the Disc text (45) is too limited, and on the other hand the use of stamps in this case seems extremely uneconomical, because a large number of different stamps would be needed in store for the equivalent number of different words of a vocabulary. The empiric evidence that no pure, non-phoneticised word-picture-writing is known for any civilized race (Jensen 1969, p. 44) seems to be another argument.

Nevertheless, following the research done by Gunther Ipsen, the

[13] The German language for instance has in addition to the 26 canonical letters, three 'umlauts' *ä, ö, ü*, and the *ß*, together with a number of substitutes for separate phonetic values such as *ch, ck, ng, sch, qu, ph* and several diphthongs (*ai, au, äu, ei, eu, ie, oi*, etc.), which can be interpreted monophthongally. A catalogue of more than 40 phonetic values which are always represented by one token could then be provided easily. The fact that the sound-inventory of the IPA (International Phonetic Association) distinguishes many more than 100 single phonetic values should only be noted in passing. In order to point out a concrete example: Amharic, an Ethiopian language, has 33 consonants and 7 vowels (Taylor 1883, p. 35).

[14] Neumann, however, pointed out that the argument of highly pictorial features of the characters which is frequently drawn on for supporting the existence of an ideographic system, is not necessarily solid. The picture writing of Byblos (dated about the 18^{th} century B. C.) could work as an example for a syllabary based on pictograms (Neumann 1968, p. 31). But in my opinion there is a difference: the pictograms of this writing are more stylized than those of the Disc, and their decoding has not yet reached a stage where the existence of ideographic elements can be definitely excluded. Furthermore, in the Byblos writing the total number of tokens with distinct pictorial features is much smaller than in the case of the Disc.

opinion has more and more gained dominance that the writing mode of the Disc is of the basic type *syllabary*.[15] Dieter Rumpel's contribution is based on the same premise as well. Since the pros and cons are hardly being discussed any longer, I cannot help feeling that this is due to the remote effect of the successful example given by Michael Ventris and John Chadwick, for whom the hypothesis of syllable-based characters became the key for the decoding of Linear B. But in this case the reasons for drawing such an analogy are absent: the writing on the Disc is by far older than Linear B, and, alos in contrast to Linear B, its language is *not* Greek, indeed it might not even belong to the Indo-European language family at all.[16]

There is yet another problem. The hypothesis that the basic type consists of a syllable-based script does not imply that it is a pure syllabary; its supporters are mostly cautious on this point.[17] This reserve is justified, since in *all* Cretan writing systems, Linear B included, ideographic elements are strongly represented. There is no reason to assume an exception in the case of the Disc text within a writing evolution of a thousand years. Nevertheless, many attempts to obtain linguistic information about the Disc are based on an *exclusive* stock of syllable

[15] Finally, Yves Duhoux, as one of the authorities on the Disc has pleaded for the syllabary assumption, without however discussing the arguments in favour of or against this hypothesis: "[...] *il semble* [!] *que l'on ait affaire à un syllabaire à syllabes ouvertes (da, de, di, do, du, etc.), séparant soigneusement les mots les uns des autres et ne notant pas les consonnes finales de mots.*" (Duhoux 1983, p. 35). More decidedly Irene Müller thought it possible to ascertain "*mit Sicherheit* [!] *eine Silbenschrift wie in Linear B und im Kyprischen*" (Müller 1983, p. 168); but she also neglects documenting the stated absence of any ambiguity.

[16] Concerning the question of the origin of the Minoan language cf. Georgiev 1981. Georgiev identifies the language of the Disc with that of Linear A and, as the result of his close to 40 years of research on this field, he considers the Minoan language as descended from the (Hetite-)Luwian, an Indo-European language originating in southern Anatolia. However, Alfred Heubeck demonstrates in the same volume with his contribution in regard to Linear A how much more differentiatedly and cautiously such classifications must be conducted (Heubeck 1981).

[17] Neumann, for example, pleaded for the syllabary theory as well, but in a differentiated conclusion, would consider it "*überraschend, wenn der Diskos nur Silbenzeichen und keine Träger von Wortbedeutungen gekannt haben sollte.*" (Neumann 1968, p. 31^{sq}).

equivalents, thus obviously restricting the chances of success.

Depending on the dating of the Disc, the defenders of the syllabary theory also see themselves confronted with the problem of the non-existence of a parallel for such an early pictographic script based on syllabograms.[18] The only way to avoid this difficulty is to date the Disc in the second half of the second millenium in accordance with Jeppesen. The interdependence of the questions concerning the writing system and the time of origin of the Disc once again makes it evident that the many unsolved riddles of the Disc form a complex of problems rending it difficult and in somes cases even impossible to deal with *one* question at a time.

Besides the arguments already adduced, there are, in my opinion, two statistical observations which weaken the hypothesis of a merely syllable-based script:

a) the first is the in part extremely heterogenous frequency of the distribution of the types on the Disc. Syllabograms should rather be expected to appear in a standard distribution over the text, since they are not subject to any connection with the content. Out of the 45 pictograms of the Disc, almost half, exactly 21, appear only on *one* side! The following tabular summary records the different tokens and their distribution:[19]

It is true that it is mainly a matter of rather infrequently represented types, but even among the remaining ones distinct deviations from the standard distribution attract our attention: the second most frequent type, for example 7, is to be found on side A only three times, whereas it appears 15 times on side B. With type 12, the relation is 15:2(!); with type 24 1:5; with type 26 5:1; and even the most frequent type, 2, with its relation of 14:5, shows a striking discrepancy with the expected average value. It is hardly conceivable that reasons inherent in the text might cause such an uneven distribution of certain syllables. Nonetheless I shall not here provide statistical figures, because the

[18] But cf. Neumann's information in n. 14.

[19] Cf. Rumpel's complete catalogue of the Disc tokens in table 1, based on Arthur J. Evans' research, together with a KWIC-Index, to be found in Duhoux 1977, p. 45-55.

Type	Side	Number	Type	Side	Number
3	A	2	21	A	2
4	A	1	22	B	5
5	B	1	28	A	2
9	B	2	30	B	1
10	A	4	31	A	5
11	A	1	36	B	4
15	B	1	41	A	2
16	B	2	42	B	1
17	A	1	43	B	1
19	A	3	44	A	1
20	B	2			

validity of the conclusions drawn on the basis of the observed frequency distributions is restricted due to a fundamental problem: in order to achieve reliable statistical statements the evaluated text corpus must be sufficiently large – which in the case of the Disc it not the case. Yet the observation adduced here should be taken into consideration as further possible evidence *against* the assumption of a pure syllabary.

b) The second observation concerns the length of the 'words'. With most of the languages of the Western world it can be verified *cum grano salis* that between 85 to 95% of the words used have one to three syllables.[20] Monosyllabic words (articles, pronouns, copulae, prepositions, etc.) prevail with a percentage of 35 to 60, followed by two- and three-

[20] To render a comparison possible, this specification is based upon an analysis of spoken syllables in the *tokens* of different text corpora, and *not* on a *type*-oriented examination. I am conscious that beyond the known and (according to the specific language) varying difficulties in giving a secure definition of a syllable (just to mention the occasionally considerable differences between spoken and written syllables in the Romance languages as well as in English), my rather general statement would have to be further differentiated according to language, author, style, level of language, the epoch in the history of language and perhaps of the literary genre, too, for a more thorough examination. In the interest of a concise presentation, I will neglect that, as well as the discussion of differences in the respective frequency graph, that e. g. in the English language rises much more rapidly from mono- to tri-syllabic words than in Spanish or German, where four-syllable words have a comparatively high share of the totality of tokens.

syllabic words.[21] Differing from this is e.g. Latin (or, more precisely: artistic Latin prose), and, to some extent, the Italian language, the di-syllabic words of which, followed by mono- and tri-syllabic ones, are at the head of the frequency lists.[22] The average values of 'syllables per word' range from 1.4 (English) to 2.4 (Latin).

In comparison, a cursory glance at the following table will show that the occurrences known from other languages do not correspond to the Disc in the least:[23]

	Number of tokens per group						
	2	3	4	5	6	7	total
Side A	5	9	8	4	2	3	31
Side B	1	7	14	8	–	–	30
total	6	16	22	12	2	3	61

The difference is twofold: the average number of 'types per group', with a value of 3.95, is almost twice as high as the average values for 'syllables per word' stated above. Furthermore, the development of the frequency distribution graph is completely different: instead of the usual cumulation observed at the beginning of the graph, in the area of mono- and di-syllabic words, this graph reaches its vertex only in the middle, with the groups of four-syllable words; 'groups' of one syllable are lacking altogether.

Doubtlessly it would be wrong to take our modern understanding of 'word' as the basis of results gained from this comparison, especially so since it is a highly artificial product which has emerged from a

[21] This characterization holds true *mutatis mutandis* for German, English, French, and Spanish as well as Ancient and Modern Greek, with especially the English language containing an extreme preponderance of monosyllabic tokens.

[22] In Latin, the main reasons for that are to be found in the rare occurrence of articles and personal pronouns, whereas in the Italian language (which also uses the 'person within the predicate') the cause lies in the distinct diphthongization and the frequent contraction of prepositions and articles.

[23] Here is an example for the understanding of the table: The third column demonstrates that the groups formed by four tokens (that is, according to the syllabary hypothesis, four-syllable words) are represented eight times on side A and fourteen times on side B.

lengthy process of standardization.[24] Even if the concept of 'word' underlying the Disc text contains – as is likely – further elements such as determinatives[25] as well as not precisely definable pro- and enclitica, the observed deviation can still not be explained sufficiently. This also seems to oppose the interpretation of the Disc symbols as syllabograms and the definition of the groups of types as words.[26]

There is yet another argument. The consistent separation of written words is the result of a process which took place rather late in the history of occidental writing, with regional disparities from the eighth century A. D. on up to the twelfth. Up to that time separation of the written text did not occur at all ('scriptura continua': all words were continually written together without spaces), or they were divided into linguistically or rhetorically constituted syntagms (sentences or parts of sentences), for which free spaces or other separators were used. Still nowadays in the Hindu devanāgarī writing system, word separation occurs only when at the word boundaries a vowel, *-ṃ* or *-ḥ* are to be found on one side, and a consonant on the other; otherwise the words are connected according to strictly euphonic rules (Sandhi). Examples of separations between words date all the way back to antiquity – I am thinking for instance of the earliest Greek papyri from the fourth century B. C. – but they result from the (necessary) pauses made by the scribe rather than from any concept of distinct words; therefore

[24] The attempt to distinguish with certainty single elements in a foreign spoken language must fail, just as it is futile to search for a compelling explanation for why the numerous compositа in German are to be understood as *one* word. In view of the various and seemingly universal difficulties of defining a word as such according to objective criteria, Frederick Bodmer felt moved to state *"daß Wörter das sind, was man im Wörterbuch findet"* (Bodmer [undated], p. 38).

[25] Determinatives are known from other ancient picture writings as indicators of word semantics. They served as reading supports and remained unvoiced.

[26] One restriction of the result must not be concealed: In my examination of analogies, no agglutinating languages (e. g. Finnish, Hungarian, Japanese) are represented, since no text corpora of them were at my disposal for a computer-based evaluation. Their average values probably would show much closer affinities to the Disc values than those gained from inflecting languages. (If they finally would agree with the Disc values, remains questionable, however.) Duhoux was the last one to plead cautiously plead for an agglutinating language in the case of the Disc (Duhoux 1983, p. 46-49].

they are never placed with consistency.[27] Stringent differentiation of word units is completely unknown in the earliest epigraphic tradition of writing. It occurs for instance neither in the early Greek inscriptions nor in the Egyptian hieroglyphics.[28] Seen from this perspective it would consequently be a surprise if the groups of signs on the Disc represented 'words'. Their classification as such – e. g. in order to base a linguistic examination on that assumption – would need thorough justification.

In general I consider the reduction of the writing type aspect to a strict 'either-or' question as a rather critical simplification. The 'phonetisation' of a written language – in this case: the transition from ideogram to syllabogram, or, to be more precise, to phonogram – is a process which generally takes place gradually rather than anonymously and discontinuous, as can be observed not only on the island of Crete. Ideograms, as well as a continuously growing stock of one and more consonant symbols, have been characteristic of Egyptian hieroglyphics, the best-known and most striking example of this, have been in use since

[27] For this information I owe thanks to the papyrologist of the University of Trier, Prof. John Shelton. Some Roman texts of the early empire form a true exception: The scribes used points to mark the intervals between the single words; but this custom soon fell into oblivion. Cf. in general on the development of the Latin writing: Bischoff 1979, especially p. 214-219. It should also be mentioned that in the manuscripts originating in late antiquity and the early Middle Ages, there occasionally occur vertical strokes. They function as word separators and are thus reminiscent of the Disc. In fact, however, they turn out to be reading aids added *later*.

[28] Irene Müller on the contrary calls to mind the word separations in the texts of Ugaritic (Müller 1968, p. 169). I cannot follow her there: I do not know of any example in Ugaritic cuneiform writing – in which *scriptura continua* was used as in the other contemporary writings – for such a clear and consistent application of word separators as on the Disc. Excluded from my argumentation are solely non-continuous texts such as vocabularies in which separations are indispensable for the understanding of the text. Besides, the analogy suffers from the fact that Ugaritic writing is a *character* (or rather, consonant) script. The urge to constitute 'word units' there is different from that in ideographic or syllable-based writing. More distinct in comparison are the word separators in later texts of the Hetiti-hieroglyphic writing (approximately 10^{th} to 8^{th} century B. C.), which are very reminiscent of the combination of *spiritus lenis* and acute in Greek. Even there, however, a regular and stringent use is out of the question.

the beginning of the third millenium. Another instance is the present Japanese writing system, which rests on a fusion of Chinese ideograms ('Kanji') as radicals with specifically Japanese syllable tokens ('Kana') for morphology, particles, postpositions, etc. The high plausibility of a comparable hybrid form is unquestionable, especially for the other Cretan pictographical systems (of Knossos and Mallia), even though the quantitative proportion of ideographic to syllabographic forms in them remains unspecified: The *communis opinio* of the relevant research, however, insists on the generally *ideographic* basic character of these scripts and regards as controversial only the size of the stock of syllabic elements.[29]

On the background of this evidence, I am convinced that the Disc pictograms represent an essentially ideographic writing; a probably high but for the time being not exactly determinable quota of syllable tokens or phonograms can be deduced from the relatively small stock of tokens and from the usage of stamps. Hence I would see the groups formed by these tokens as syntagms rather than as words. Unfortunately there is no ground to be gained from this in approaching the language: since clues for further differentiation of ideographic and syllable tokens are missing and an interpretation of the groups as syntagms leaves a much larger scope of interpretation for the rubrication of its elements, the linguistic decoding of the Disc is also rendered more complicated by the very same functional ambiguity of its characters.

A final comment on the total size of the stock of characters of the Disc writing: now and then attempts have been made to gain clarity on the issue through extrapolation.[30] I can only caution against any such undertaking: especially when the the possibility of a combination of ideographic and syllabic script is taken into account, and it is recalled that the characteristics of the frequency distribution of a vocabulary (few words occur frequently, many appear quite rarely) allow a reliable representative conclusion concerning the infrequently appearing elements of the anonymous whole only if one proceeds from a very large random sample, then any numerical specification can be disqualified as

[29] Cf. Neumann 1968, p. 32, which includes recommendations for further reading.

[30] Ipsen, for instance, estimated the number of tokens as *"mehr als fünfzig, aber gewiß nicht erheblich über 60"* (Ipsen 1929, 8).

unscientific speculation.

(5) The Writing and Reading Direction on the Disc

Besides the aspect of the writing system, the possibly most important obstacle to the linguistic decoding of the Disc is the problem of its reading direction. Be it right-to-left (that is, from the outside to the inside) or, on the contrary, left-to-right (that is from the inside to the outside), the reading direction will determine whether the search for text constituents, such as determinatives, radicals, prefixes and suffixes, and for morphological clues will be crowned with success. In the discussion which has gone on far more than eighty years now, numerous shrewd but in some cases unfortunately contradictory arguments in favour of either of the two opinions have been formulated. Nonetheless research even on this matter is far from solution.

At this point, since it is neither possible nor advisable to give a complete listing and evaluation of all the aspects of the problem assembled up to the present,[31] I shall confine myself to a survey.

As early as 1908, Pernier imagined that he had clues for a left-to-right reading (Pernier 1908, p. 274), an opinion which was shared by other authorities, for example by Evans. The controversy on the subject began about a year later, when Alessandro della Seta, in a detailed essay, presented a complete catalogue of mainly technical arguments for a right-to-left writing direction (Della Seta 1909). The essential point of his argumentation was the observation that when several tokens are interfere with each other, the one on the right-hand side is slightly overlapped by the left one. The conclusion seemed obvious: the left token, being situated nearer to the centre, must have been stamped later – thus it appeared to be proven that the writing direction ran from the outside to the inside. Many scholars, among them Evans, were convinced. Ever since Duhoux's comprehensive up-to-date resumé of the contemporary state of research promoting della Seta's argument as the only irrefutable one was published,[32] the right-to-left writing direction has been more and more considered as a canonical fact.

[31] Duhoux gives a resumé of the 'argumentation budget' (Duhoux 1977, p. 25-35).

[32] Duhoux rejected the remaining evidence as subjective or as a possible argument for the opposite theory: *"Ils sont indépendants de toute vue a priori et*

Nevertheless, a small community of researchers clung to the assumption of a left-to-right reading direction.[33] While they accepted the opposite writing direction as a possibility, for several reasons, they assumed a mismatch between the reading and the writing direction. Their arguments ranged from the barely conceivable hypothesis that the Disc served merely as a matrix for the actual writing of the documents to the assumption that an 'illiterate copyist' could have stamped the text that way for manual-technical reasons (Grumach). The latter argument in particular should not be rejected too easily in spite of its unfavourable reception: the scribe must have made use of a layout in order to create such a meticulously elaborated document with such a precise text arrangement. This premise can be considered as a connecting link to Grumach's hypothesis.

This, however, is only a hint that the reading direction might not coincide with the writing direction. The defenders of the left-to-right reading theory saw a *positive* proof only in the 'more meaningful' result ostensibly obtained when one reads from the inside to the outside.[34] But as long as there is no reliable preliminary linguistic knowledge of the Disc to permit precise deductions, this argument must be dismissed. On the one hand, the formulation mentioned is circular in that it introduces into the examination as a working hypothesis precisely the one aspect that actually should be found out, and on the other hand I cannot help noting that the research based on a right-to-left reading so far has not hindered the achievement of 'efficient' and respectable results.[35]

Only recently, an attempt has been made to infer the reading direction from a detail not inherent in the text, one to which little attention

de toute interprétation subjective. Ce sont eux, et eux seuls, qui permettent l'établissement indiscutable de la direction de l'écriture." (Duhoux 1977, p. 24-27].

[33] Its most important representative is Ernst Grumach (in particular Grumach 1962 and 1965).

[34] Cf. Grumach 1962 and 1965. Ernst Schertel even talked about a *"senseless chaos"* resulting from right-to-left reading (Schertel 1948, especially p. 336 & 349.

[35] Special importance has to be attached to Duhoux's linguistic analyses – even more so since they were done against the background of a comparison with Linear A (Duhoux 1983).

had been paid before (Hoschek 1981). On each side of the Disc, the spirals terminate at their *outer* ends with a vertical stroke bearing several globular bulges. Rudolf Hoschek interpreted these bars (that call to mind an abacus) as numerical tokens such as are known for instance from early Sumeric texts. He came to the conclusion that these tokens play the role of a sort of pagination for a few Disc sides. This pagination could only make sense in the present case – he counted four globular forms on side A and five on side B – if the reading was left-to-right.[36]

Unfortunately, this very original idea is traceable to a mistake in the research method. The finding of four points on side A goes back to Evans (Evans 1921, p. 651 & 662), who was misled by a faulty facsimile; after he had originally correctly determined the number to be five (Evans 1909, p.274). As the photographic enlargements in Duhoux's work[37] demonstrate, the number of five globular elements on side A is also a well-established fact. Consequently Hoschek's hypothesis must be rejected.

The most recent attempt at resolving the riddle has been made by Hans-Joachim Haecker, who made a remarkable attempt to weaken della Seta's 'overlapping argument' (Haecker 1986). Relying on the observation that the tokens involved in corrections are always stamped into the clay at a different depth, he carried out the following experiment:

> "[...] habe ich einen stempelähnlichen Gegenstand zweimal in Knetmasse (später auch in Ton) gedrückt, einmal tief und ein zweites Mal schwächer, und zwar so, daß der zweite Eindruck nur den Rand des ersten überschnitt. Das Ergebnis: nicht der erste, sondern der zweite – schwächere – war deformiert, wirkte wie angeschnitten."

The result proves nothing less than the necessity of turning the original

[36] It was Hoschek's idea that, in the case of right-to-left reading, the pagination at the upper rim is problematic: 'Our' disc would then contain the sides 4 and 5, the hypothetical disc before that would have sides 2 and 3, leaving for its predecessor only *one* side, no. 1. This rather forced postulation of an empty Disc side at the very beginning of the text is not necessary in the case of the left-to-right reading direction: the 'pagination' in this case could have the function of an *indicator* for the following side – thus the distribution of text on all Disc sides would be secured.

[37] Cf. Duhoux 1977, p. 82 & 85, Fig. 30 & 39; cf. as well p. 19 for the mentioned 'tradition of error'.

deduction of right-to-left script into its opposite! Is it possible that the apparent evidence of della Seta's observation prevented its verification? My cursory investigation of the literature has in any case revealed nothing indicating such attempts. Therefore I found myself impelled to make a preliminary imitation of the experiment, the results of which on the whole tend to confirm Haecker's hypothesis. Such examinations, however, have an impact only when they are conducted systematically – which in this context would require exact copies of the original stamps with precisely measured imprinting depths, widths of overlapping, and, furthermore, the presumably authentic mixture of clay soil.

Through Haecker's experiment, the possibly most important argument for the right-to-left hypothesis was seriously called into question for the first time. Notwithstanding its a success, the experiment cannot be regarded as *proof* of the left-to-right reading direction: arguments proposed by those who want to disconnect the reading and the writing direction will remain valid even in this case.

There is one final aspect to complete the apory: at the presumable time of the origin of the Disc, no common writing usage had yet been formed that could serve as an indicator. In early epigraphy, independent of the cultural area and epoch, writing from left to right and from right to left, from the top to the bottom and from the bottom to the top and (although more rarely), 'furrow writing' ('bustrophedon')[38] occur. Now and again all these variants appear in close spatial and temporal proximity. The observation that the tokens depicting human beings or animals in the Egyptian hieroglyphics usually face the beginnings of lines does not constitute a conclusive argument by analogy either, since there are cases in which the relation is reversed.[39]

[38] 'Bustrophedon' [gr.]. means the writing 'as an ox-plough': one line runs from left to right, the next from right to left, etc. A further variation, but one unknown in the occidental history of culture, is the writing of the Easter Islands: each line is reversed in relation to the preceding one – thus the text constantly has to be turned upside down while reading.

[39] This applies for the Meroïtic script for instance, a character writing developed on the Egyptian example (Jensen 1969, p. 71^{sq}). Besides, the case exists that a script runs in *both* directions, cf. Jensen's table about the Sinai script, especially the characters Aleph, Nahasch I, Samekh, Ṣadae, and Resch [ibid., p. 257].

(7) The strokes

In addition to the pictograms, the Disc contains other kinds of signs. There are altogether seventeen strokes to be found,[40] carved by the scribe at the bottom line of tokens when they were already pressed into the clay. They are always affixed to the element among a group of tokens situated closest to the centre of the Disc, and generally point diagonally to the bottom right side.

What they stand for is just as unclear as the function of the other writing constituents. They have been interpreted as diacritic symbols (1), as indicators of prosody (2) and morphology (3), as signals for omitted vowels (4), as sentence or paragraph limits (5), as metrical caesura or breaks within a verse (6), or as independent (affix-) symbols. Regardless of this variety it might still be possible to cut the Gordian knot: in my opinion, possibilities 1, 2 and 4 can be eliminated at once, since such functions are not conditionally related to the position of the character within a word or sentence. Thus they cannot serve as an explanation for the regular occurrence of the stroke solely at the end (or, according to the reading direction, at the beginning) of a group of symbols. The same applies to 3 and 7, in case one interprets the groups of tokens as syntagms, as I do. If the word theory holds good, possibility 7 is not convincing either, because of the frequent occurrence of the stroke connected with tokens that themselves are interpreted as affixes.[41] Finally, the interpretation of these strokes as metrical and strophe marks must in my opinion be dismissed as well, because all the text paragraphs thus formed are too irregular to constitute a rhythmical unit: the smallest of these paragraphs mentioned covers just one group (4 times), whereas the largest one contains 12 groups (once). Besides, metrical caesura that *continually* coincide with a the end of a group – be it word or sentence – would be extremely unusual.

[40] In the relevant literature, the number of strokes varies greatly, because small fissures on the surface of the Disc were taken to be strokes and, on the other hand, poorly executed strokes were not recognized as such. Duhoux convincingly pointed out, however, that the exact number of strokes must be seventeen (Duhoux 1977, p. 36-39).

[41] Cf. e. g. Ipsen's (1929), Duhoux's (1983), and Rumpel's results in their contributions.

Consequently only the classification of the strokes as sentence or paragraph limiters remains. This interpretation, differing from the other discarded possibilities, seems reasonable; I do not see anything contradictory here; indeed there are several aspects in favour of it.[42]

In any case it is obvious that the scribe connected an important function to the placement of the strokes. A linguistic examination of the text, taking syntactical features into account as well, should therefore aim at reconciling the methodical starting point with the allocation of a function to the strokes. To disregard them might prove to be a methodical mistake.

III. Conclusion

My task of *advocatus diaboli* described at the beginning is herewith fulfilled. I have pointed out that any research on the Disc can be performed only on an extraordinarily small basis of secured knowledge.

[42] Hans-Joachim Haecker has called attention to remarkable repetitions of groups of tokens and their parts that, with *left-to-right* reading, appear at the beginning (and only there!) of the syntactical units (Haecker 1986, p. 94-96). The following table by Haecker is reproduced with one modification, an additional stroke discovered by Duhoux; the numbers designate groups of tokens or parts of such groups occuring several times, whereas the asterisks represent groups that do not recur:

Side A	Side B
1 2 3 1	8 10 * *
4 2 * * *	11 *
5	8 * 10
6 7	11
5 * 3	8 *
5	11 * * * * * * * * * * *
6 7 *	* * *
* 4 * * * * * *	* * *
8 *	—
9	—

These repetitions can be taken – together with the striking (anaphoric?) rhythm of the group setting – as indicator that the reading direction might actually run from the inside to the outside. The strokes then could mark the beginnings of sentences or paragraphs.

As for further material, there is nothing available except a complex of deceptive and threadbare presumptions, possibilities and indicators. Among the multitude of problems there are some especially weighty ones:

- We have no information on the historical context of the Disc, the impact of which could be discovered in the text and made use of for the decoding.
- The Disc is a unique object: linguistic examinations remain restricted to a text which is by far too limited to be used as a statistical corpus.
- Nothing reliable can be said even about the most elementary propaedeutica of an approach to the contents of the text – the nature and the direction of the writing.

Thus it is hardly astonishing that the 'decoding' up to now has produced rather deplorable results although they have in part been heavily glossed over.[43] All these results were determined by the claim of having been achieved *more geometrico* – nonetheless they have never agreed in even the most crucial points: in the highly complex deduction chains even previses differing only slightly which produce totally different results. Finally, it need hardly be mentioned that the 'translations' do not agree with each other in the least either.

This result is disappointing. One might now conclude with inexorable severity that no further scientific search for the solution of the riddle is worth while. That would be a mistake, though. I am convinced that there exists a philological and historical competence which cannot be reduced to the mere atoms of positive knowledge. In the history of scholarship we find various cases where, with the help of this competence, proceeding even from hopelessly mutilated information, some at least partial reading could be gained. In some cases the outcome has been confirmed afterwards. An aporetic situation such as the present one, which has so often been unsuccessfully investigated calls for a difficult balancing act, however: the scholar must – as long as no further related discoveries, perhaps even the appearance of a bilingual text comparable to the 'Rosetta Stone', cause a substantial widening

[43] Cf. the critical examination of several such attempts in Neumann 1968, 32-36.

of the material basis – *"soberly see the limits of his possibilities, if he does not want to experience that no one except himself believes in the accuracy of his hypotheses."* (Neumann 1968, p. 42[sq]).

References

Bischoff, B. (1979), *Paläographie des römischen Altertums und des abendländischen Mittelalters.* Berlin: Schmidt (= Grundlagen der Germanistik, 24).

Bodmer, F. (undated), *Die Sprachen der Welt. Geschichte – Grammatik – Wortschatz in vergleichender Darstellung.* 5th ed., Köln, Berlin: Kiepenheuer & Witsch (undated).

Davaras, C. (undated), *Das Museum von Hagios Nikolaos.* Athens (undated).

Davaras, C. (1981), "Zur Herkunft des Diskos von Phaistos". In: *Kadmos*, 20, 101-105.

Della Seta, A. (1909), "Il disco di Phaistos". In: *Rendiconti della Reale Accademia dei Lincei: Classe di Scienze Morali [...]*, Ser. 5, 18, 297-367.

Duhoux, Y. (1977), *Le disque de Phaestos. Archéologie, Épigraphie, Édition critique*, Index. – Louvain: Édition Peeters.

Duhoux, Y. (1983), "Les langues du linéaire A et du disque de Phaestos." In: *Minos (Salamanca)*, 18, 33-68.

Evans, [Sir] A. J. (1909), *Scripta Minoa. The written documents of Minoan Crete [...].* Vol. 1, Oxford: Clarendon.

Evans, [Sir] A. J. (1921), *The palace of Minos. A comparative account of the successive stages of the early Cretan civilization.* Vol. 1, London: Macmillan.

Georgiev, V. I. (1981), "L'état actuel des études des inscriptions minoënnes". In: Heubeck, A., Neumann, G. (Hrsg.), *Res Mycenaeae. Akten des VII. Int. Mykenologischen Colloquiums [...].* – Göttingen: Vandenhoeck & Ruprecht, 126-130.

Grumach, E. (1962), "Die Korrekturen des Diskus von Phaistos". In: *Kadmos*, 1, 16-26.

Grumach, E. (1963), *Bibliographie der kretisch-mykenischen Epigraphik (nach dem Stande vom 31. Dezember 1961).* München, Berlin: Beck (additionally the supplement I, published in 1967 which covers 1962 to 1965).

Grumach, E. (1965), "Der Ägäische Schriftkreis". In: *Studium Generale*, 18, 742-756.

Haecker, Hans-Joachim: "Neue Überlegungen zu Schriftrichtung und Textstruktur des Diskos von Phaistos". In: *Kadmos*, 25, 89-96.

Heubeck, A. (1981), "Überlegungen zur Sprache von Linear A. In: Heubeck, A., Neumann, G. (Hrsg.), *Res Mycenaeae. Akten des VII. Int. Mykenologischen Colloquiums [...].* – Göttingen: Vandenhoeck & Ruprecht, 155-169.

Hoschek, R. (1981), "Zur Schriftrichtung beim Diskos von Phaistos". In: *Kadmos*, 20, 85-92.

Ipsen, G. (1929), "Der Diskus von Phaistos. Ein Versuch zur Entzifferung". In: *Indogermanische Forschungen*, 47, 1-41.

Jensen, H. (1969), *Die Schrift in Vergangenheit und Gegenwart.* 3rd ed. Berlin: VEB Deutscher Verlag der Wissenschaften.

Jeppesen, K. (1962), "En Gammelkretisk Gåde". In: *Kuml. Årbog for Jysk Arkaeologisk Selskab*, 157-190.

Marinatos, S. (1935), "Ausgrabungen und Funde auf Kreta 1934-1935. In: *Archäologischer Anzeiger*, 244-259.

Müller, I. (1983), "Ein paar grundsätzliche Bemerkungen zum Diskos von Phaestos". In: *Kadmos*, 22.

Neumann, G. (1968), "Zum Forschungsstand beim Diskos von Phaistos". In: *Kadmos*, 7, 27-44.

Pernier, L. (1908), "Il Disco di Phaestos con caratteri pittografici." In: *Ausonia*, 3, 255-302.

Schertel, E. (1948), "Der Diskos von Phaistos. Wege zu seiner Entzifferung". In: *Würzburger Jahrbücher für die Altertumswissenschaft*, 3, 334-363

Taylor, I. (1883), *The alphabet. An account of the origin and development of letters.* Vol. 1, London: Paul.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 175–177

Kybernetische Aspekte des synergetischen Modells von R.Köhler. Diskussion.

Jaroslaw Maj
Bochum

In diesem Aufsatz soll folgende Auffassung R.Köhlers aus dem Aufsatz "Zur Charakteristik dynamischer Modelle. Anmerkungen zu einem Beitrag von R.Hammerl und J.Maj" (Köhler,1989) diskutiert werden:

> "Da ich dieses Modell nun ausdrücklich als dynamisches eingeführt hatte, differenzieren entweder unsere Meinungen darüber, was es bedeutet, die Zeit in einem Modellansatz zu berücksichtigen oder aber darüber, worin in einem systemtheoretischen Modell die Dynamik besteht."

Wird aber die Bezeichnung "systemtheoretisches Modell" verwendet, dann sind daraus folgende Konseqenzen abzuleiten: In der Systemtheorie werden die Begriffe dynamisches und statisches System exakt definiert. Als ein dynamisches System (Philippow 1977, 271 ff.) wird ein System bezüglich seines dynamischen Verhaltens bezeichnet. Dynamisches Verhalten wiederum heißt Verhalten in der Bewegung. Das Verhalten in der Bewegung eines Systems (Modells) wird durch folgendes System von Gleichungen beschrieben (Philippow 1977, 213 ff.; diese Gleichungen gelten für ein analoges lineares und zeitinvariantes System - ein solches System untersucht Köhler):

$$\frac{dz(t)}{dt} = \hat{A}z(t) + \hat{B}x(t) \tag{1}$$

$$y(t) = \hat{C}z(t) + \hat{D}x(t), \tag{2}$$

wobei gilt:

t – Zeit,
$z(t)$ – zeitabhängige Zustandsgröße,
$x(t)$ – zeitabhängige Eingangsgröße (Eingabe),
$y(t)$ – zeitabhängige Ausgangsgröße (Ausgabe),
$\hat{A}, \hat{B}, \hat{C}, \hat{D}$ – Koeffizientenmatrizen.

Gleichung (1) wird Überführungs-oder Bewegungsgleichung genannt. Für ein eindimensionales System, d.h. für ein System, in welchem keine Bewegung (d.h. keine zeitliche Änderung der Zustandsgröße und somit der Ausgangsgröße) stattfindet, gilt:

$$\frac{dz(t)}{dt} = 0$$

d.h.

$$a \cdot z(t) + b \cdot x(t) = 0;$$

dann ist

$$z(t) = \frac{b}{a}x(t).$$

Nach Gleichung (2) gilt somit

$$y(t) = (d - \frac{b}{a})x(t). \qquad (3)$$

bzw.

$$\frac{y(t)}{x(t)} = (d - \frac{b}{a}) = \text{const.},$$

wo a, b, c und d Systemparameter sind.

Wie aus Gleichung (3) folgt, ist die Relation der Variablen y und x (und damit auch die Variable z) in Abhängigkeit von der Zeit konstant und stellt genau den von R.Köhler beschriebenen Sachverhalt dar. Auch für den Fall, daß Köhler Veränderungen der Parameter in einem sich vollziehenden Selbstregulationsprozeß zuläßt (die man z.B. als Veränderungen in Abhängigkeit von der Zeit beschreiben könnte),

so sagt das Köhlersche Modell jedoch nicht voraus, in welchem Maße diese Veränderungen auftreten.

Somit ist das von R.Köhler beschriebene Modell aus **systemtheoretischer Sicht** ein statisches Modell, unabhängig davon, ob es in "funktionaler Hinsicht" als dynamisch bezeichnet werden kann oder nicht (vgl. Köhler 1990, 42).

Literatur

Hammerl, R., Maj, J. (1989), "Beitrag zu Köhlers Modell der sprachlichen Selbsregulation." *Glottometrika 10*, 1-31.
Köhler, R. (1986), *Zur linqustischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Köhler, R. (1989), "Zur Charakteristik dynamischer Modelle. Anmerkungen zu einem Beitrag von R.Hammerl und J.Maj." *Glottometrika 11*, 19-40.
Philippow, E. (1977), *Taschenbuch der Elektrotechnik.* Band 2. Grundlagen der Informationstechnik. Berlin: Verlag Technik.

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 179–187

Elemente der synergetischen Linguistik[1]

Reinhard Köhler
Bochum/Trier

Einführung

Bei der Synergetik handelt es sich um eine spezifische Ausprägung systemtheoretischer Modellbildung, die sich durch die Beschäftigung mit der spontanen Entstehung und Entwicklung von Strukturen auszeichnet. Die Vertreter der synergetischen Linguistik haben in den letzten Jahren gezeigt, daß sich die funktionalanalytischen Modelle und Erklärungsansätze, wie sie innerhalb der quantitativen Linguistik verfolgt werden (vgl. Altmann 1981), fruchtbar in den interdisziplinären Rahmen der Synergetik einfügen lassen. Der synergetische Ansatz bietet geeignete Konzepte und Modellvorstellungen dafür an, Phänomene der Selbstregulation und Selbstorganisation, wie sie in der quantitativen Linguistik untersucht werden (das bekannteste und älteste Beispiel ist das sog. Zipfsche Gesetz und das Prinzip der geringsten Anstrengung), als Resultate eines Wechselspiels zwischen Zufall und Notwendigkeit zu erklären. Wie andere selbstorganisierende Systeme ist die Sprache durch das Vorhandensein von kooperierenden und konkurrierenden Prozessen gekennzeichnet, die zusammen mit den von außen auf die Sprache wirkenden (psychologischen, biologischen, physikalischen, soziologischen etc.) Kräften die Dynamik des Systems ausmachen. Eine wichtige Rolle spielen das Versklavungsprinzip und die Ordnungsparameter; letztere sind makroskopische Größen, die das Verhalten der mikroskopischen Mechanismen bestimmen, auf deren Ebene aber selbst nicht repräsentiert sind.

[1] Dieser Beitrag entstand im Zusammenhang mit dem Bochumer Projekt "Sprachliche Synergetik", das seit 1986 von der Stiftung Volkswagenwerk unterstützt wird. Dieses Projekt verbindet unter der Leitung von G.Altmann zahlreiche Wissenschaftler mehrerer Disziplinen in internationaler Zusammenarbeit bei der Erforschung theoretischer Aspekte der synergetischen Linguistik und ihrer empirischen Überprüfbarkeit anhand von Daten vieler und typologisch verschiedener Sprachen.

Die synergetische Modellierung basiert auf einer prozeßorientierten Vorgehensweise. Bei der Modellierung geht man von den bekanntesten oder vermuteten Mechanismen und Abläufen des Untersuchungsgegenstandes aus und formuliert sie mit Hilfe geeigneter mathematischer Entsprechungen (z.B. Differentialgleichungen oder stochastischer Prozesse). Aus den Beziehungen zwischen den Prozessen und den übergeordneten Ordnungsparametern ergibt sich das Verhalten des Systems. Die Möglichkeit der Ausbildung neuer Strukturen (Sprachwandel, Sprachrevolution) hängt wesentlich mit der Existenz von Fluktuationen zusammen, die den Evolutionsmotor bilden. Die möglichen Systemzustände ("Moden"), die sich aufgrund der Relation im System ergeben können, werden durch Randbedingungen und andere Ordnungsparameter begrenzt; nur die passenden Moden können sich im Wettbewerb untereinander behaupten.

Axiome

Das wichtigste strukturelle Axiom der synergetischen Linguistik besagt, daß sprachliche Systeme Selbstorganisations- und Selbstregulationsmechanismen besitzen, die für eine Anpassung an die Systemumwelt sorgen. Die Umwelt einer Sprache besteht aus den sozialen und kulturellen Systemen, die sie verwenden, aus den Menschen,[2] die sie benutzen, mit ihren Gehirnen, Gehör- und Artikulationsapparaten, den Kommunikations- und anderen sozialen Bedürfnissen, ihren Sprachverarbeitungs- und Spracherwerbssystemen. Zur Umwelt gehören auch die physikalischen Medien (Übertragungskanäle) der Sprache mit ihren spezifischen Eigenschaften.

Auf dem Axiom der Selbstorganisation beruht die Erklärungskraft des synergetischen Ansatzes. Die Entstehung sprachlicher Phänomene und ihre jeweilige Ausprägung können so **funktional** in bezug auf entsprechende Anforderungen aus der Systemumwelt (oder in bezug auf inner-

[2] Im Gegensatz dazu betrachtet Altmann (vgl. z.B. Altmann 1987,233) die Sprachträger, also die Menschen, und die Texte als Subsysteme des Sprachsystems. Er faßt also die Grenze des Systems Sprache weiter als es hier geschieht. Der Unterschied hat aber lediglich zur Konsequenz, daß auch der Unterschied zwischen außerhalb des Systems befindlichen Anforderungen und den internen Systemgrößen entsprechend getroffen werden muß.

sprachliche Ordnungsparameter (s. unten) mit Hilfe der entsprechenden Anpassungsprozesse erklärt werden.

Eine weitere Klasse von Axiomen bilden die **"Systembedürfnisse"**. Dieser Ausdruck ist eigentlich schlecht gewählt; er stammt aus dem soziologischen Funktionalismus ("need") und bezeichnet ein Bedürfnis, das vom System "bedient" werden muß. Es handelt sich also, wie oben gesagt, um **Anforderungen** aus der Systemumgebung, die das System erfüllen muß, um seiner Aufgabe gerecht zu werden – und nicht etwa um Bedürfnisse des Systems

Systembedürfnisse

Die bisher in synergetischen Untersuchungen betrachteten Systembedürfnisse lassen sich in drei Gruppen einteilen:

1. *Sprachkonstituierende* Systembedürfnisse

Hierzu gehören

- das Kodierungsbedürfnis (**Kod**) und
- das Spezifikationsbedürfnis (**Spz**), zu deren Bedienung die Sprache Ausdrücke als Kodierungsmittel zur Verfügung stellt, mit den Teilaspekten der Kodierung:
  - Darstellungsbedürfnis (**Darst**),
  - Appellbedürfnis (**App**) und
  - Ausdrucksbedürfnis (**Ausdr**).

  Für alle Teilaspekte müssen Sprachen, um ihre Funktion erfüllen zu können, entsprechende Ausdrucksmittel bereitstellen. So bieten alle natürlichen Sprachen zur Bedienung des Ausdrucksbedürfnisses (gemeint ist der Selbstausdruck des Sprechers) eine mehr oder weniger freie Wahl zwischen – in bezug auf die reine Darstellung – äquivalenten Ausdrucksmitteln, so daß **Stil** als Möglichkeit entsteht. Der gleiche Zusammenhang ist auch eine der Quellen von **Synonymie**.
- das Anwendungsbedürfnis (**Anw**). Für in einer Sprache vorhandene Ausdrucksmittel besteht ein jeweils spezifisches Anwendungsbedürfnis, das der Relevanz der durch das jeweilige Mittel kodierten Bedeutung für die Sprachgemeinschaft zu einer gegebenen Zeit entspricht. So ist – im Augenblick – das Bedürfnis

nach Anwendung des Worts *Katalysator* höher als das nach Anwendung von *Raubritter*. Das Anwendungsbedürfnis wirkt sich unmittelbar auf die Systemgröße **Frequenz** aus.

2. *Sprachformende* Systembedürfnisse
   - Das Bedürfnis nach Sicherheit der Informationsübertragung (**Red**), das zu der Erscheinung der **Redundanz** führt.
   - Das Ökonomiebedürfnis (**Ök**), das dem Zipfschen Prinzip der geringsten Anstrengung entspricht, mit seinen Teilaspekten
     - Minimierung des Produktionsaufwands (**minP**),
     - Minimierung des Kodierungsaufwands (**minK**),
     - Minimierung des Dekodierungsaufwands (**minD**),
     - Minimierung des Gedächtnisaufwands (**minG**) bzw.
     - Minimierung des Inventarumfangs (**minI**),
     - Kontextunabhängigkeit (**KÖ**),
     - Kontextspezifizität (**KS**) und
     - das Bedürfnis nach Invarianz der Relation Ausdruck-Bedeutung (**Invz**).

3. Systembedürfnisse der *allgemeinen Steuerung*

Auf der Ebene der Steuerungseffektivität (vgl. Köhler 1986, 159f.), der höchsten Abstraktionsebene der synergetischen Linguistik, gibt es zwei konkurrierende Bedürfnisse:

- das Anpassungsbedürfnis (**Anp**) und
- das Stabilitätsbedürfnis (**Stab**),

von denen das erstere fördernd auf die Variabilität der Systemeigenschaften wirkt, um die Flexibilität zu erhöhen, das zweite bremsend und moderierend, um einem Zerfall des Systems entgegenzuwirken.

Die Systembedürfnisse werden für den jeweils untersuchten Zusammenhang postuliert und sind aus dem Modell als solchem nicht ableitbar. Sie sind dennoch theoretisch und empirisch überprüfbar – nur eben nicht *innerhalb* des modellierten Systems. Die Bedürfnisse, die Systemgrößen und ihre Relation untereinander bilden die **Struktur** des Systems.

Funktionale Äquivalente

In den natürlichen Sprachen gibt es zur Bedienung eines Bedürfnisses gewöhnlich mehr als ein Mittel. So entsprechen dem Kodierungsbedürfnis (der Anforderung an die Sprache, Mittel zur Kodierung von Bedeutungen darzustellen) vier prinzipiell mögliche Techniken: die lexikalische, die syntaktische, die morphologische und die prosodische Methode. Die meisten Sprachen verwenden alle vier oder wenigstens drei dieser Methoden, wenn auch in unterschiedlichem Ausmaß (diese Tatsache wurde auch zur typologischen Klassifizierung herangezogen).

Zwei oder mehr Mittel, die ein gegebenes Bedürfnis bedienen können, nennt man **funktional äquivalent** in bezug auf dieses Bedürfnis. Die Existenz funktionaler Äquivalente muß bei der Modellierung sorgfältig überprüft werden. Sie spielt eine wichtige Rolle bei der **funktionalen Erklärung** (vgl. Altmann 1981; Köhler 1986, 26ff.; Köhler 1990b)

Kooperierende und konkurrierende Prozesse

Bei der synergetischen Modellierung spielen kooperierende bzw. konkurrierende Prozesse eine zentrale Rolle. Diese Prozesse verkörpern die Selbstorganisations- bzw. Selbstregulationsmechanismen des sprachlichen (Sub-)Systems, indem sie die Systemgrößen, auf die sie wirken (z.B. die Länge lexikalischer Einheiten), den Anforderungen (Systembedürfnissen) anpassen. Die Prozesse sind unterschiedlicher Art, aber immer stochastisch. Sie spielen in der Sprachdynamik eine der biologischen Evolution durch Mutation und Selektion analoge Rolle: zufällig entstehende Varianten (z.B. die stets leicht voneinander abweichenden Aussprachevarianten von Lauten in der konkreten Sprechsituation) haben eine Überlebenswahrscheinlichkeit, die davon abhängt, wie gut ihre Eigenschaften (z.B. Artikulationsaufwand, Wahrnehmungsaufwand) den Anforderungen (z.B. minP, minD) entgegenkommen. Gut passende "Mutationen" haben die Chance, sich durchzusetzen. Beispiele für Prozesse dieser Art sind (vgl. Köhler 1987, 253):

Phonologische Unifikation,
Phonologische Diversifikation,
Phonologische Restriktion,

Lexikalisierung,
Lexikalische Unifikation,
Lexikalische Diversifikation,
Lexikalische Reduktion,
Spezifizierung,
Kontext-Globalisierung,
Kontext-Zentralisierung,
Anwendung,
Kürzung.

Ordnungsparameter

Die Wirkungsweise der synergetischen Mechanismen kann sehr verschiedenartig sein; allen gemeinsam ist, daß sie mehr oder weniger direkt einer Steuerung durch andere Systemgrößen oder durch nicht zum System gehörende Parameter unterliegen. Diese Größen, von denen also die betrachteten Variablen und ihre Veränderungen abhängig sind, werden in der Synergetik Ordnungsparameter genannt.

Sie sind durch ihre Rolle in dem jeweils betrachteten Zusammenhang als Ordnungsparameter charakterisiert und nicht durch inhaltliche Kennzeichnungen. Insbesondere sind Ordnungsparameter solche Größen, die – im betrachteten Zusammenhang – nicht zu der Mikroebene, auf der die synergetischen Prozesse ablaufen, zu rechnen sind, sondern zu einer Makroebene der Analyse. In einigen Fällen fallen die Ordnungsparameter direkt mit Anforderungen ("Systembedürfnissen") zusammen. So sind etwa in dem von Altmann entwickelten Modell zur anstrengungsbedingten Lautveränderung (Job, Altmann 1985) die Anforderungen **minP** (hier: Minimierung der Artikulationsanstrengung) und **minD** (Minimierung des Dekodierungsaufwands), die ja die Prozesse des Lautwandels steuern, auch als Ordnungsparameter anzusehen. In anderen Fällen sind Ordnungsparameter selbst Systemgrößen: So ist die **Lexikongröße** einer der Ordnungsparameter für die Mikroprozesse, die für die Wortlänge verantwortlich sind. Die Lexikongröße ist aber kein Systembedürfnis; sie ist vielmehr selbst ein Resultat von Prozessen, die das **Kodierungsbedürfnis** bedienen.

Wie G.Altmann betont, ist es durchaus möglich, daß eine Größe von Prozessen betroffen ist, für die sie selbst die Rolle eines Ordnungsparameters spielt. Ein Beispiel hierfür ist der Verbreitungsprozeß neuer Formen in einer Sprache (**"Piotrowski-Gesetz"**), bei dem die Veränderung der Proportion von der erreichten Proportion abhängt. Dies stellt aber keinen Widerspruch zu dem Ebenen-Kriterium dar! Die zu einer Zeit gegebene Proportion ist *in bezug auf die lokalen Veränderungsprozesse* als globale, nahezu konstante Größe anzusehen. Diese Eigenschaft wird auch mit der Bezeichnung "Versklavungsprinzip" betont.

Modelltypen

Modellierung ist stets Vereinfachung. Ein Modell konzentriert sich auf bestimmte Eigenschaften des modellierten Wirklichkeitsausschnitts und vernachlässigt andere. Ein Modell erfaßt die gesamte Komplexität der Realität – es müßte sonst auch mindestens genauso komplex sein wie diese und verlöre dadurch jeden Sinn. Daher sind, natürlich auch innerhalb des synergetischen Forschungsansatzes, immer mehrere verschiedene Modelle denkbar, die unterschiedliche Aspekte der Untersuchungsobjekte zum Gegenstand haben. Einige wichtige Typen von Modellen sollen hier kurz charakterisiert werden.

1. Stochastische Prozesse

Die explizite Modellierung von Mechanismen der Sprachdynamik kann mit Hilfe stochastischer Prozesse durchgeführt werden. Dabei sind die Wahrscheinlichkeiten der möglichen Ausgänge abhängig von Systemgrößen oder von der Zeit. So läßt sich die Länge lexikalischer Einheiten als eine Variable modellieren, die über einen stochastischen Prozeß von der Frequenz dieser Einheit abhängt: Zu jedem Zeitpunkt t besteht eine Wahrscheinlichkeit dafür, daß zu einer gegebenen Einheit mit der Länge L und der Frequenz F eine neue Variante mit der Länge L' (kürzer oder länger als L) zum Zeitpunkt t+1 entsteht. Diese Wahrscheinlichkeit hängt von F und L zu t ab. Die Frequenz der neuen Variante ist beim erstmaligen Auftreten natürlich 1; ihre weitere Entwicklung unterliegt einem Geburts- und Todesprozeß – die Variante kann sich schließlich durchsetzen, oder aber nicht. Die Wahrscheinlichkeit ihrer Ausbreitung steigt mit schon erreichter Frequenz. In dieser Situation existieren zwei oder mehr verschieden lange Varianten der betrachteten lexikalischen Einheit, die im Hinblick auf ihre Anwendung

miteinander in Konkurrenz stehen. Die Anwendungswahrscheinlichkeit jeder Variante hängt von der Länge und der zum Zeitpunkt t erreichten Frequenz (Verbreitung) ab.

2. Funktionale Abhängigkeiten

Die Struktur eines synergetischen Modells besteht aus den Systemgrößen und den Relationen zwischen ihnen. Die Relationen verkörpern die dynamischen Abhängigkeiten, die aufgrund der Mechanismen zwischen den Systemelementen bestehen. Somit kann aus der **Struktur** des Systems bzw. seiner Teile die **Funktion** abgeleitet werden.

Die einzelnen funktionalen Abhängigkeiten werden aus Differentialgleichungen gewonnen, die linguistischen Hypothesen über die Mechanismen entsprechen. Dabei wird von den zeitlichen Abläufen abstrahiert. Dem stochastischen Charakter der Prozesse wird dadurch Rechnung getragen, daß die Werte der funktionalen Abhängigkeiten als Mittelwerte aufgefaßt werden. Sie geben an, welche Werte der abhängigen Variablen aufgrund der Systemstruktur optimal bei gegebenen Werten der unabhängigen Variablen sind. Enthält das Modell beispielsweise die Abhängigkeit der Wortlänge von der Frequenz in Form einer Funktion L=f(F), so wird durch sie jeder Worthäufigkeit die Länge zugeordnet, die ihr am besten entspricht (bekanntlich ist die optimale Zuordnung die, bei der die häufigsten Wörter die kürzesten sind).

3. Wahrscheinlichkeitsverteilungen

Das Aufstellen direkter Wahrscheinlichkeitsmodelle ist für viele Erscheinungen ebenfalls möglich. In einem solchen Modell gibt ein Verteilungsgesetz an, auf welche Weise die Wahrscheinlichkeit des Auftretens über die verschiedenen Ausprägungen einer Variable verteilt ist (z.B. welchen Anteil an einem Text/einem Wörterbuch Wörter mit der Länge 1,2,3... haben). Hierzu gehören auch die Rangverteilungen, die u.a. im Bereich der Diversifikationsprozesse (Rothe 1990) eine wichtige Rolle spielen.

Für diesen Modelltyp hat Altmann (1990) einen allgemeinen Ansatz entwickelt, aus dem ein ganzes System von Verteilungsfamilien folgt. Abgesehen von dem mathematischen Wert dieser Arbeit erlaubt der

Ansatz, ausgehend von linguistischen Parametern (Systemgrößen und Systembedürfnissen) je nach Konstellation die verschiedensten und bisher meist nicht interpretierbaren Verteilungen *abzuleiten*.

Literatur

Altmann, G. (1981), "Zur Funktionalanalyse in der Linguistik". In: Esser, J., Hübler, A. (Hrsg.), *Forms and Functions*. Tübingen, 25-32.

Altmann, B. (1983), "Das Piotrovskij-Gesetz und seine Verallgemeinerungen". In: Best, K.-H., Kohlhase, (Hrsg.), *Exakte Sprachwandelforschung*. Göttingen, 59-90.

Altmann, G. (1987), "The Levels of Linguistic Investigation". In: *Theoretical Linguistics* 14, 2/3, 227-239.

Altmann, G. (1990), "Modelling diversification phenomena in language". In: Rothe, U. (Hrsg.), *Diversification Processes in Language: Grammar* (erscheint).

Altmann, G., von Buttlar, H., Rott, W., Strauß, U. (1983), "A Law of Language Change". In: Brainerd, B. (Hrsg), *Historical Linguistics*. Bochum, Brockmeyer, 104-115.

Job, U., Altmann, G. (1985), "Ein Modell für anstrengungsbedingte Lautveränderung". In: *Folia Linguistica Historica* 6, 401-407.

Köhler, R. (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brochmeyer.

Köhler, R. (1987), "Systems Theoretical Linguistics". In: *Theoretical Linguistics* 14, 2/3, 241-257.

Köhler, R. (1990a), "Zur Characteristik dynamischer Modelle. Anmerkungen zu einem Beitrag von R.Hammerl und J.Maj". *Glottometrika* 11. 41-46.

Köhler, R. (1990b), "Linguistische Analyseebenen. Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation". *Glottometrika* 11, 1-18.

Rothe, U. (1990, Hrsg), *Diversification Processes in Language: Grammar*. (erscheint)

Hammerl, R. (Ed.):
Glottometrika 12.
Bochum: Brockmeyer 1990, 189–199

Rezension

M.V. Arapov, Kvantitativnaja lingvistika. Moskva, Nauka, 1988, 184 S.

Rolf Hammerl

Die rezensierte Monographie von M.V. Arapov, der schon eine Reihe sehr wesentlicher Arbeiten zur quantitativen Linguistik veröffentlichte (in deutscher Sprache ist im Brockmeyer-Verlag in Bochum im Jahre 1983 die in Zusammenarbeit mit M.M.Cherc geschriebene Monographie "Mathematische Methoden in der historischen Linguistik" erschienen), ist Ergebnis langjähriger Untersuchungen und zeichnet sich besonders durch Auswahl von sehr aktuellen Forschungsproblemen aus, durch die breite Anwendung mathematischer und statistischer Verfahren, durch das Bestreben nach empirischer Überprüfung der abgeleiteten Modelle, durch klare und reich illustrierte Erklärungen. Diese Arbeit stellt eine "Schatzkammer" für alle Forscher dar (nicht nur für Linguisten), die sich mit der Anwendung exakter Forschungsmethoden bei der Untersuchung sprachlicher Erscheinungen befassen.

Es ist unmöglich, alle wichtigen Gedanken und Methoden der rezensierten Arbeit hier zu nennen; da bisher aber noch keine Übersetzung dieser Arbeit ins Deutsche vorliegt (die wir übrigens für notwendig halten), werden wir unten den Inhalt der einzelnen Kapitel möglichst übersichtlich und mit einer nicht zu großen Straffung wiederegeben. Eine Bevorzugung und besondere Wertung spezieller Probleme werden wir bewußt vermeiden, da diesbezüglich nur das gesamte Buch in Frage kommen würde.

Es soll besonders vermerkt werden, daß diese Arbeit am Anfang der achziger Jahre entstand, so daß der Autor eine Vielzahl von Publikationen zur quantitativen Linguistik aus den letzten Jahren nicht berücksichtigen konnte (z.B. die ähnliche Forschungsprobleme betreffende Arbeit von R. Köhler "Zur synergetischen Linguistik: Struktur

und Dynamik der Lexik", Bochum, Brockmeyer, 1986). Ein Vergleich der Methoden und Ergebnisse der Untersuchungen von Arapov auf der einen Seite und der Mitarbeiter des Projektes "Sprachliche Synergetik", welches an der Ruhr-Universität Bochum realisiert wird auf der anderen Seite, erscheint deshalb von besonderer Bedeutung. Diesem Problem wollen wir uns in einer gesonderten Arbeit zuwenden.

M.V. Arapov verfolgt die Absicht, an begrenztem Sprachmaterial zu zeigen, wie man aus quantitativen sprachlichen Untersuchungen Informationen über qualitative Aspekte der Sprache gewinnen kann. Als Untersuchungsmaterial dient die Lexik natürlicher Sprachen, wobei Wörter nicht nur als Einheiten eines Wörterbuches, sondern auch von Texten untersucht werden. Dieses Buch soll die Kommunikation zwischen Linguisten, die mit traditionellen Methoden arbeiten, und Wissenschaftlern erleichtern, die sich mit Fragen der Mensch-Maschine-Kommunikation befassen.

Der überwiegende Teil des Buches wurde vorher in verschiedenen Zeitschriften veröffentlicht, nur das 5. und 6. Kapitel wurde speziell für dieses Buch geschrieben. Dem Leser dieses Buches sollten die Grundlagen der Wahrscheinlichkeitstheorie bekannt sein; der Autor selbst war bestrebt, keine komplizierten statistischen Methoden anzuwenden. Diesbezüglich sei uns die Bemerkung erlaubt, daß die Lektüre dieses Buches sogar für den mathematisch vorgebildeten Linguisten nicht immer einfach sein wird.

Der Hauptakzent wird bei der Besprechung einzelner Probleme nicht so sehr auf deren erschöpfende Behandlung gelegt (das ist wohl auch beim gegenwärtigen Entwicklungsstand der quantitativen Linguistik gar nicht anders möglich), sondern darauf, im Rahmen eines einheitlichen Vorgehens logische Verbindungen zwischen den untersuchten sprachlichen Charakteristika herzustellen. Dies ist jedoch nur dann möglich, wenn der Autor mit bestimmten vereinfachten Modellen der komplizierten sprachlichen Realität arbeitet. Hierin liegt keineswegs ein Nachteil der Untersuchungen von Arapov, sogar dann nicht, wenn sich diese Vereinfachungen zu einem späteren Zeitpunkt als unzulässig erweisen, sondern es handelt sich um eine notwendige Bedingung für die Erreichung eines Erkenntnisfortschrittes bezüglich der analysierten Problematik. Die Arbeit Arapovs besteht aus 8 Kapiteln, die in logischer Aufeinanderfolge angeordnet wurden; der Besprechung dieser Kapitel

wenden wir uns nun zu.

1. Kapitel 1: Die Bedeutung quantitativer Daten für die Erforschung der Sprache

Arapov hebt nach einer Diskussion erkenntnistheoretischer Ansätze für eine systemhafte Untersuchung der natürlichen Sprache zwei Tatsachen hervor: Erstens untersucht jeder Linguist als Praktiker die Sprache immer unter qualitativem und gleichzeitig auch quantitativem Aspekt, und zweitens verfügen die Linguisten (in der Regel) nicht über die theoretischen Grundlagen für eine umfassende Interpretation der Ergebnisse entsprechender Untersuchungen. Dies ist wohl auch einer der Gründe, daß Erkenntnisse und Methoden der quantitativen Sprachwissenschaft von einer immer noch relativ kleinen Gruppe von Wissenschaftlern angewandt werden.

Arapov unterscheidet streng zwischen einer statistischen Linguistik und einer quantitativen Linguistik, die bei der Anwendung beliebiger statistischer Verfahren in der sprachwissenschaftlichen Forschung nicht stehenbleibt. Für das Verständnis der Untersuchungsmethodologie Arapovs ist das 3. Unterkapitel von besonderer Bedeutung, da sowohl methodologische Grundpositionen als auch wesentliche Begriffe, die im weiteren Text Anwendung finden, erläutert werden. Arapov vertritt die Meinung, daß die aus empirischen quantitativen Untersuchungen gewonnenen "Zahlen" selbst keine grundlegende Rolle spielen, sondern die durch diese Zahlen festgelegten Ordnungsrelationen, der "Platz" der jeweiligen Untersuchungseinheit in einer solchen Ordnung, d.h. deren Relationswert. In der quantitativen Linguistik seien drei miteinander verbundene Fragen zu lösen:

a) Wie sind die Wertskalen aufgebaut, d.h., welche Eigenschaften besitzt die jeweilige Ordnungsrelation?
b) In welchen quantitativen Wechselbeziehungen zeigt sich dieser "Wert"?
c) Mit welchen anderen qualitativen Eigenschaften der sprachlichen Objekte ist dieser "Wert" verbunden?

In Antwort auf diese Fragen führt Arapov aus, daß er den Wortschatz V als Menge der Wörter natürlicher Sprachen untersucht und eine einfache lineare Ordnung in V. Im Wortschatz V kann somit auch eine

einheitliche Numerierung dessen Elemente durchgeführt werden (z.B. als Rang im Häufigkeitswörterbuch). Als "Wert" der Elemente der Ordnungsskala wird in der Arbeit von Arapov stets der Rang berücksichtigt (so wie er aus Häufigkeitslisten bekannt ist); untersucht werden soll dann der Zusammenhang zwischen Rang und (mittlerer) Häufigkeit der Elemente, zwischen Rang und (mittlerer) Länge, (mittlerer) Bedeutungszahl und (mittlerem) Alter.

2. Kapitel 2: Häufigkeit als Charakteristikum der "Gebräuchlichkeit" (russ. upotrebitel'nost') der Wörter im Text

Ziel dieser Untersuchungen ist es, die aus Textanalysen bekannten starken Differenzen der Gebräuchlichkeit der Textelemente (Wörter, Wortformen) zu erklären, ohne sich der Hypothese zu bedienen, daß die Textelemente eine stabile Auftrittswahrscheinlichkeit besitzen. Es ist das große Verdienst Arapovs, als erster (soweit uns bekannt ist) aus der Analyse von sogenannten freien Klassifikationsexperimenten, welche ein einfaches (Text)modell realisieren, das Zipfsche Gesetz abgeleitet zu haben. Der Autor diskutiert in diesem Zusammenhang das Problem des Wertebereiches für die Gültigkeit dieses Gesetzes (auch in Konfrontation mit den Untersuchungen Orlovs zum Zipfschen Gesetz in sogenannten "vollendeten" Texten). Arapov zeigt am Beispiel von Strukturwörtern, daß das Postulat der "Stilometrik" nach Konstanz der Auftrittshäufigkeit dieser Wörter nicht mit der Verteilung derselben nach dem Zipfschen Gesetz vereinbar ist. Im Zusammenhang damit findet man im Text auch interessante Überlegungen über die Relation zwischen der syntaktischen Gliederung der Texte und dem Zipfschen Gesetz.

An anderer Stelle untersucht der Autor die Tatsache, daß man ein und dieselbe Klassifikation auf verschiedene Weisen erhalten kann, und zeigt, wie man dieses Variationsprinzip für die Erklärung der Verteilung der Auftrittshäufigkeit verschiedener Wörter im Text (Text als sehr einfache Klassifikationsstruktur) anwenden kann. Für besonders bedeutsam schätzen wir auch den Hinweis ein, daß man eine Erklärung für die bekannten Probleme bei der Anpassung des Zipfschen Modells an empirische Daten nicht über die Einführung immer neuer Parameter

suchen sollte. Das Zipfsche Gesetz, welches z.B. aus einem vereinfachten Textmodell hergeleitet werden konnte (d.h. unter bestimmten vereinfachten Annahmen über die Textstruktur), gilt eben gerade unter diesen Voraussetzungen für die entsprechenden Textmodelle und nicht ohne weiteres für die komplizierte Struktur realer Texte.

3. Kapitel 3: Veränderungen der Gebräuchlichkeit von Wörtern in der Synchronie

In diesem Kapitel werden Differenzen in den Häufigkeitsverteilungen der Einheiten von Häufigkeitswörterbüchern untersucht, die ja bekanntlich auf bestimmten Textkorpora basieren. Es wird die Hypothese überprüft, daß bestimmte Unterschiede in der Thematik und im Genre der zugrundeliegenden Texte Differenzen im Vokabular der Häufigkeitswörterbücher verursachen und daß diese proportional den Differenzen der Logarithmen der Ränge ein und desselben Wortes in den verglichenen Wörterbüchern sind. Somit müßte man auch, wenn der Rang eines Wortes in einem Häufigkeitswörterbuch bekannt ist, den Rang dieses Wortes im entsprechenden anderen Häufigkeitswörterbuch voraussagen können. Die Güte dieser Prognose ist aber wiederum an die Erfüllung bestimmter Bedingungen gebunden und wird um so schlechter, je größer der Rang des Wortes ist.

Arapov entwickelt ein mathematisches Modell (hier u.a. unter der Annahme, daß die Texte, die der Erstellung von Häufigkeitswörterbüchern zugrundeliegen, einen sehr großen Umfang haben), welches einem quantitativen Vergleich von Häufigkeitswörterbüchern (über den Vergleich der häufigsten Wörter dieser Wörterbücher) dient. Dieses Modell wird in umfangreichen empirischen Untersuchungen überprüft. Im dritten Unterkapitel geht Arapov einer sehr interessanten Fragestellung nach, die auch für unsere Untersuchungen im Rahmen des Projektes "Sprachliche Synergetik" von besonderer Bedeutung ist. Es wird danach gefragt, in welcher Anzahl von individuellen Texten, welche dem für die Erstellung von Häufigkeitswörterbüchern untersuchten Textkorpus angehören, das jeweilige Wort mindestens einmal auftritt. Köhler (1986) nannte diese Eigenschaft Polytextie. Das von Arapov abgeleitete Modell über den Zusammenhang zwischen dem Rang und dieser Polytextie hat jedoch nur beschränkten Anwendungsbereich; dieses Modell

gilt nämlich nur unter der Annahme, daß – wenn für jeden individuellen Text ein Häufigkeitswörterbuch vorliegt - das Vokabular dieser Häufigkeitswörterbücher untereinander gleichermaßen ähnlich ist, daß die Zahl der unterschiedlichen Wörter in diesen Texten gleich ist und daß es sich um solche Texte handelt, die im Orlovschen Sinne als "vollendet" gelten.

4. Kapitel 4: Historische Veränderungen der Gebräuchlichkeit von Wörtern (Gebräuchlichkeit und Alter der Wörter)

In diesem Kapitel wird die Relation zwischen dem Rang der Wörter und deren Alter untersucht; es sei hier noch einmal erwähnt, daß dieser Problematik auch das von Arapov zusammen mit Cherc geschriebenes Buch "Mathematische Methoden in der historischen Linguistik" (Bochum, Brockmeyer 1983) gewidmet ist. In der rezensierten Monographie versucht Arapov, einige Resultate des in der früheren Arbeit untersuchten Modells mittels der in den vorangegangenen Kapiteln eingeführten Begriffe umzuformulieren. Resultat ist die Ableitung eines neuen Verfahrens zur Einschätzung des Tempos von Veränderungen im Vokabular. Im Unterschied zu den Analysen in den Kapiteln 1 bis 3 werden hier keine Charakteristika von Einzelwörtern untersucht, sondern von ganzen Wortmengen.

Untersucht werden wiederum Häufigkeitswörterbücher, von Texten mit "ähnlichem" Genre und "ähnlicher" Thematik umfassen. Die Relationen zwischen zwei Häufigkeitswörterbüchern einer Sprache, die zu unterschiedlichen Zeiten erarbeitet wurden bzw. das Vokabular unterschiedlicher zeitlicher Epochen betreffen, können prinzipiell mit denselben Mitteln untersucht werden wie die Relationen zwischen zwei Wörterbücher, die ein und denselben Zeitpunkt betreffen (letztere Relationen wurden im Kapitel 3 analysiert). Die zeitlichen Veränderungen des Vokabulars werden ausschließlich als Sterbeprozeß beschrieben, nicht aber als Geburts- und Sterbeprozeß (wie z.B. Altmann (1983) im Artikel "Das Piotrowski-Gesetz und seine Verallgemeinerungen" (in: Best, K.-H., Kohlhase, J. (eds.), Exakte Sprachwandelforschung, Göttingen, Herodot, 59-90)), und werden mathematisch als Poisson-Prozeß modelliert. Resultat dieser Überlegungen ist eine Wahrscheinlichkeitsverteilung, welche angibt, wie groß die Wahrscheinlichkeit dafür ist, daß nach einer gewissen Zeit ein Wort mit einem bestimmten Rang "abstirbt"; der Parameter in der entsprechenden Funktion soll das Tempo der Veränderungen des Vokabulars messen.

Arapov stellt mehrere Methoden zur Schätzung dieses Parameters vor. Die vom Autor gestellte Hypothese, daß dieser Parameter für verschiedene zeitliche Perioden in der Entwicklung einer Sprache konstant ist, kann in empirischen Untersuchungen zur russischen Sprache nicht bestätigt werden. Dieser Problematik ist eine umfangreiche Diskussion gewidmet.

Im dritten Unterkapitel wird ein besonders kompliziertes, aber auch hoch interessantes Problem aufgegriffen: Es wird die Hypothese gestellt, daß Wörter mit "ähnlicher" Bedeutung auch einen "ähnlichen" Rang besitzen (z.B. in Häufigkeitswörterbüchern zweier verschiedener Sprachen). Beim Vergleich von zwei Wörterbüchern aus zwei verschiedenen Sprachen hat man es mit mehrdeutigen Relationen (in jeder der beiden Richtungen) zu tun. Arapov stellt zu dem genannten Problem erste theoretische und experimentelle Untersuchungen an, ohne dieses Problem jedoch umfassend lösen zu können (was ja auch nicht das Ziel dieser Untersuchungen war).

5. Kapitel 5: Wortlänge und Gebräuchlichkeit

In den Kapiteln 5 und 6 werden Relationen zwischen Eigenschaften lexikalischer Einheiten untersucht, die auch im lexikalischen Basismodell von R. Köhler (op. cit.) aus einem allgemeinen Modellansatz abgeleitet und an deutschem Sprachmaterial überprüft wurden. Gegenwärtig wird dieses Modell auch an umfangreichem polnischen Sprachmaterial überprüft. Im Unterschied zu Arapov wird als Kriterium der Gebräuchlichkeit nicht der Rang der Wörter berücksichtigt, sondern die Auftrittshäufigkeit der sprachlichen Einheiten. Vor allem aus diesen Gründen sind die theoretischen und empirischen Untersuchungen Arapovs gegenwärtig von höchster Aktualität. Im Kapitel 6 begründet Arapov die Wahl des Ranges anstelle der absoluten Frequenz damit, daß der Rang nicht direkt vom Stichprobenumfang abhängt wie die Frequenz. Es ist aber im Modell von Köhler gleichgültig, ob nun

die absolute Häufigkeit oder die relative Häufigkeit (die ja genauso stark vom Stichprobenumfang und der Stichprobenzusammensetzung abhängt wie der Rang) berücksichtigt wird, da sich lediglich der Wert einer Konstanten ändern wird, der Funktionstyp selbst wird davon aber nicht beeinflußt. Arapov leitet aus theoretischen Überlegungen ein Modell ab, welches einen linearen Zusammenhang zwischen der mittleren Wortlänge in Silben und dem Logarithmus des Ranges dieses Wortes betrifft. Dieser Funktionstyp wird später auch angewandt, wenn die Wortlänge nicht in Silben, sondern in Buchstaben gemessen wird.

Der besondere Wert der Analysen Arapovs besteht hier auch darin, daß zusätzlich der Zusammenhang zwischen dem Rang und der Dispersion der Längen aller Wörter mit dem entsprechenden Rang untersucht wird. Dieser Zusammenhang sagt etwas über die Güte der Voraussage der mittleren Wortlänge für einen bestimmten Rang voraus.

Der Autor wendet sich auch besonders der Erklärung des "Menzerathschen Gesetzes" im Rahmen des abgeleiteten Modells zu. Die von Arapov genannten zwei Ursachenkomplexe (Einfluß von Anfangs- und Endsilben in Wörtern unterschiedlicher Länge auf die mittlere Silbenlänge; Einfluß der relativ großen Zahl von Homonymen in einem Wörterbuch auf die mittlere Silbenlänge bei Berücksichtigung aller Wörter des Wörterbuches), welche den "Menzerathschen Effekt" hervorrufen sollen, sind wohl nicht ganz von der Hand zu weisen (zumindest wenn man sich nur auf die Messung der Wortlänge in Silben und der mittleren Länge dieser Silben in Buchstaben konzentriert), der Gültigkeitsbereich dieses Gesetzes ist jedoch wesentlich größer (und betrifft z.B. auch den Zusammenhang zwischen Satzlänge und mittlerer Clauselänge); für die Erklärung dieses Gesetzes liegen mehrere aussagekräftige Modelle vor (vgl. Altmann, G., Schwibbe, M., Kaumanns, W., Köhler, R., Wilde, J.: "Das Menzerathsche Gesetz in informationsverarbeitenden Systemen." Stuttgart, Olms 1988).

6. Kapitel 6: Polysemie der Wörter und deren Gebräuchlichkeit

In diesem Kapitel wird der Zusammenhang zwischen der mittleren Bedeutungszahl von Wörtern und deren Rang im Häufigkeitswörterbuch untersucht. Zunächst wird das Problem diskutiert, auf welche Weise

die Polysemie eines Wortes bestimmt werden kann. Anschließend wird eine Grenzwertinterpretation des aus dem Potenzgesetz von Tuldava und dem Zipfschen Gesetz resultierenden Zusammenhanges zwischen der Bedeutungszahl und dem Rang vorgenommen. Davon ausgehend führt Arapov eine Funktion zur Beschreibung des Zusammenhanges zwischen der mittleren Bedeutungszahl und dem Rang der Wörter ein (linearer Zusammenhang zwischen dem Logarithmus des Ranges und der mittleren Bedeutungszahl), ohne diesen aus bestimmten Annahmen mathematisch herzuleiten (wie es in den vorangegangenen Kapiteln in der Regel der Fall war). Dafür wird aber leitet Arapov aus der Annahme, daß die Wahrscheinlichkeit der Wörter mit k Bedeutungen innerhalb aller Wörter mit einem bestimmten Rang r der einsverschobenen negativen Binomialverteilung folgen, den Zusammenhang zwischen dem Rang und der Wahrscheinlichkeit der Wörter mit k Bedeutungen unter allen Wörtern mit diesem Rang ab. Summiert man dann z.B. alle diese Wahrscheinlichkeiten für eine konkrete Zahl von k=1,2,3,.., Bedeutungen über alle Ränge r, so erhält man die Wahrscheinlichkeitsverteilung für die Zahl der Wörter mit k Bedeutungen im untersuchten Häufigkeitswörterbuch. Einen ähnlichen Zusammenhang - nämlich die Wahrscheinlichkeitsverteilung der Wörter mit k Bedeutungen in einem einsprachigen Bedeutungswörterbuch - hat Krylov untersucht. Dieser Zusammenhang wird jedoch von Arapov nicht diskutiert. Die Abhängigkeit zwischen der mittleren Bedeutungszahl und dem Rang und auch die Bedeutungsverteilung der Wörter innerhalb bestimmter Rangbereiche wird an einigen Stichproben zur russischen Sprache überprüft.

7. Kapitel 7: Produktivität von Wortklassen

In diesem Kapitel werden Eigenschaften ganzer paradigmatischer Wortklassen (z.B. die Klasse der Wörter mit einem bestimmten Suffix) untersucht. Es wird eine auf mengentheoretische Überlegungen basierende Diskkussion verschiedener Möglichkeiten einer quantitativen Bestimmung der Produktivität von Wortklassen gegenüber der Nichtproduktivität geführt. Die Sprache wird als ein offenes dynamisches System angesehen, in welchem das Auftreten bestimmter Kombinationen sprachlicher Einheiten im Vergleich zu anderen Kombinationen mit

einer bestimmten Wahrscheinlichkeit vorausgesagt werden kann. Arapov leitet zwei Definitionen der Produktivität von Wortklassen ab (und gleichzeitig Methoden deren Berechnung), welche beide auf Wörterbuchanalysen beruhen. Die Anwendung der erarbeiteten Methoden setzt die Existenz entsprechender Wörterbücher (Häufigkeitswörterbücher, historische Wörterbücher, Bedeutungswörterbücher) voraus; die Erfüllung dieser Bedingung kann in konkreten Fällen Schwierigkeiten bereiten.

8. Kapitel 8: Gleichartigkeit (odnorodnost') und Regularität (reguljarnost') der Relationen zwischen den Wörterbucheinheiten

Im letzten Kapitel werden Wörter bestimmter binärerer grammatischer Kategorien vom Typ Singular/Plural, Nominativ/Genitiv untersucht. Es handelt sich hier nicht um eine Darstellung der Relationen zwischen Vertretern dieser Kategorien mittels Proportionen und der Erklärung des Prinzips der Zusammenstellung entsprechender Paare, sondern um eine quantitative Theorie der Proportionen. Diese Analysen beruhen auf der Hypothese, daß für den Fall, wenn die Wörter dieser Paare durch "gleichartige" (und "produktive") semantische Relationen verbunden sind, die Differenzen zwischen den quantitativen Eigenschaften (z.B. Rang) dieser Paare bestimmten Verteilungsgesetzen folgen. Um entsprechende Paare von solchen "gleichartigen" (und "produktiven") binären Kategorien zu finden, sind Analysen von Häufigkeitswörterbüchern (welche z.B. die Häufigkeiten von Wortformen aufführen) notwendig. Die Anwendung des von Arapov vorgestellten Modells erlaubt exakte Entscheidungen darüber zu treffen, ob die jeweilige Kategorie regulär (produktiv) ist oder nicht. Die Anwendung des Modells auf die polnische Sprache liefert ein sehr interessantes Ergebnis: Die binäre Kategorie der Adjektive und der entsprechenden Adverbien, die von vielen Lexikographen als reguläre Kategorie behandelt wird, stellt im Sinne des Modells von Arapov eine nichtreguläre Kategorie dar.

Da auch die Anwendung des hier besprochenen Modells auf Analysen bestehender Häufigkeitswörterbücher beruht, deren "Schwächen"

gegenüber realen Texten bekannt sind, sollten die Ergebnisse der Anwendung dieses Modells stets mit entsprechender Vorsicht interpretiert werden.

9. Nachwort

Eine Grunderkenntnis aus den in der rezensierten Monographie dargestellten Untersuchungen sieht deren Autor darin, daß die Forderungen nach erschöpfendem und gleichzeitig absolut zuverlässigem Wissen über die Sprache nur schwer zu vereinbaren sind. Je genauer man ein Charakteristikum der Wörter erfassen will, desto größer die Unbestimmtheit bei der Erfassung anderer Charakteristika (je größer z.B. die Sicherheit, daß ein Wort wenig Bedeutungen besitzt und von geringem "Alter" ist, desto ungenauer die Prognose der Auftrittswahrscheinlichkeit dieses Wortes).