QUANTITATIVE LINGUISTICS
Vol. 20

GLOTTOMETRIKA 5

edited
by
R. Köhler
J. Boy

Studienverlag Dr. N. Brockmeyer
Bochum 1983

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Glottometrika. – Bochum: Studienverlag Brockmeyer
5. Ed. by R. Köhler; J. Boy. – 1983.
(Quantitative linguistics; Vol. 20)
ISBN 3-88339-307-X

NE: Köhler, Reinhard [Hrsg.]; GT

ISBN 3-88339-307-X

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

INHALT

ALTMANN, G., KIND, B.
Ein semantisches Gesetz 1

KLAVIŅA, S.
Linguostatistischer Vergleich von Funktionalstilen der lettischen Sprache 14

ARAPOV, M.V.
Regularity and Homogeneity of Morphological and Word-Forming Patterns 45

BOY, J.
Die notwendige Asymmetrie binärer Stammbäume 71

ROTHE, U.
Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen 101

HEUPS, G.
Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen 113

KÖHLER, R.
Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse 134

WINKLER, P.
Markierungen paraphonetischer Information: Kurventypen, Kombinationen und Strukturen 168

TAMBOVTSEV, Yu.A.
Linguo-statistical Studies of Siberian Languages in the USSR 203

ZÖRNIG, P., ALTMANN, G.
The Repeat Rate of Phoneme Frequencies and the Zipf-Mandelbrot Law 205

BOOKS RECEIVED 212

CURRENT BIBLOGRAPHY 214

ANNOUNCEMENT OF A PROJECT 228

Ein semantisches Gesetz

G.Altmann, B.Kind, Bochum

1. Klassifikationen (wissenschaftliche, sprachliche, praktische), soweit sie nicht automatisch aus einer Theorie folgen (vgl. HEMPEL 1965), haben den Zweck, in einer Daten- oder Gegenstandsmenge eine Ordnung zu schaffen, die einer leichteren Orientierung im gegebenen Gebiet dient. Durch die Klassifikation wird besonders das Gedächtnis entlastet. Die natürliche Sprache kann man unter anderem auch als ein Klassifikationssystem betrachten, denn in ihr werden durch Nomina bezeichnete konkrete und abstrakte Dinge in Klassen geordnet und ihre Zugehörigkeit zu der Klasse wird durch einen Oberbegriff (genus proximum) festgehalten. Einzelne Sprachen weisen zwar unterschiedliche Klasseneinteilungen der Dinge auf, aber es ist anzunehmen, daß die quantitative Klassenbildung in allen Sprachen gleich ist. Sobald eine Klasse, die unter einen Oberbegriff A subsumiert ist, zu groß wird und eine allzugroße Belastung für das Gedächtnis darstellt, wird sie in Teilklassen mit neuen Oberbegriffen $A_1, A_2, \ldots$ aufgeteilt, deren genus proximum A ist. Mit anderen Worten, in große Klassen wird eine begriffliche Zwischenstufe eingeführt. Dies geschieht sicherlich nicht in dem Augenblick, wenn die Klasse einen ganz bestimmten Umfang erreicht, aber trotzdem nach einem Gesetz, das vom menschlichen Gedächtnis und anderen menschlichen Bedürfnissen diktiert wird. Die Größe der durch einen Oberbegriff zusammengehaltenen Klasse ist also nicht festgelegt, weist aber eine klare Tendenz auf, die sich quantitativ erfassen läßt. Die Bildung neuer Begriffe verläuft in beiden Richtungen, d.h. es werden zu einem Oberbegriff immer neue Einzelbegriffe gebildet (dies ist das übliche Anwachsen des Lexikons); falls ihre Zahl aber zu groß wird, dann werden neue "Zwischenoberbegriffe" gebildet.

Diesen Prozeß der Klassenbildung könnte man also als einen Verzweigungsprozeß, der von den allgemeinsten Begriffen ausgeht, modellieren, aber auch "von unten" her, als eine Zuordnung der spe-

zifischen Begriffe zu immer allgemeineren Begriffen. Wir werden hier diesen zweiten Weg verfolgen.

2. Die Klassenzugehörigkeit eines Begriffs, d.h. seine Subsumierung unter einen Oberbegriff wird anhand eines einsprachigen Erklärungswörterbuches festgestellt. In einem derartigen Wörterbuch wird die Erklärung durch genus proximum und differentia specifica gegeben. Der gegebene Begriff ist von 1. Ordnung, der Begriff, der das genus proximum darstellt, von 2. Ordnung. Ein Begriff 2. Ordnung wird wiederum durch einen Begriff 3. Ordnung definiert usw.

Der übliche Definitionsweg verläuft von spezifischen Begriffen mit großer Intension zu immer allgemeineren Begriffen mit großer Extension. MARTIN (1974), dessen Daten den Anlaß zu dieser Untersuchung gaben, hat im Französischen beispielsweise die folgende Folge von Oberbegriffen gefunden:

pistolet - arme - instrument - outil - objet - chose.

Hier ist "pistolet" ein spezifischer Begriff 1. Ordnung, "arme" ein generischer Begriff 2. Ordnung,..., "chose" ist ein Begriff 6. Ordnung. Bei der Findung derartiger Folgen von Oberbegriffen im Lexikon kann man an mehrere von MARTIN (1974:63) beschriebene Probleme stoßen:

(a) Zirkuläre Definition, wie z.B.
réveil - pendule - appareil - machine - appareil.

In solchen Fällen muß man das letzte Wort streichen (hier appareil) und die Folge nach eigener Kompetenz vervollständigen, z.B. wäre hier möglich

... - machine - instrument - outil - objet - chose

anzusetzen.

(b) Nichtnominale Definition, z.B. mit Hilfe von "das was", "etwas was", "ce que" usw. In solchen Fällen ist der letzte Begriff vor "das was" entweder als der oberste anzusehen oder man soll die Folge nach eigener Kompetenz vervollständigen.

(c) Metonymische Definition mit Hilfe von "Teil", "Glied", "Stück", "Menge", "Kollektiv" usw. Diese können als letzte Glieder der Begriffskette betrachtet werden.

(d) Da Wörter mehrere Bedeutungen haben können, gibt es für ein Wort möglicherweise mehrere Definitionsfolgen. In dem Falle kann man die Untersuchung folgendermaßen durchführen:

(i) Wenn man Wörter als Träger von Begriffen untersucht, dann soll man nur die erste Bedeutung in Betracht ziehen;

(ii) werden alle Bedeutungen des Wortes separat untersucht, so gibt es eine Definitionsfolge für jede Bedeutung, wobei mehrere auch zusammenfallen können. Diese zweite Untersuchungsart bringt kaum etwas neues; es vermehrt sich lediglich die Zahl der Begriffe erster Ordnung, während die der höheren Ordnungen fast unverändert bleibt. Wie in Tabelle 2 von MARTIN (1974:70) ersichtlich, ergeben 1723 Wörter 1. Ordnung 2392 Bedeutungen, d.h. um 669 Bedeutungen mehr als Wörter; während bei der 2. Ordnung der Unterschied nur 21, bei der dritten 1 beträgt, die anderen sind ohne Unterschied. Daran sieht man, daß unterschiedliche Bedeutungen eines Wortes in überwältigenden Mehrheit unter einen und denselben Oberbegriff subsumiert werden. Da sich dadurch hauptsächlich die Ebene der Begriffe 1. Ordnung vermehrt, kann dies zu einer Verzerrung der Klassenbildungstendenz führen. Wir werden daher als Einheiten die lautliche Form der Wörter betrachten.

3. Die einzigen uns bekannten Daten stammen von MARTIN (1974), der 1723 Substantive in G. GOUGENHEIM, Dictionaire fondamental de la langue française (Paris, Didier 1958) untersucht hat. Seine Daten sind in der Tabelle 1 aufgeführt.

Tabelle 1. Zahl der Wörter auf einzelnen Abstraktionsebenen nach MARTIN (1974)

Ebene x	Zahl der Wörter y_x
1	1723
2	348
3	108
4	39
5	13
6	3

Es wurden also 1723 Wörter aus dem Wörterbuch ermittelt. Sie wurden unter 348 Oberbegriffe (Wörter) 2. Ordnung subsumiert, diese wiederum unter 108 Wörter 3. Ordnung usw.

Um diesen monotonen Verlauf mathematisch zu beschreiben, gehen wir von zwei Annahmen aus:

(1) Auf jeder höheren Ebene (x+1) verringert sich die Zahl der Oberbegriffe proportional zu der unmittelbar niedrigeren Ebene, d.h. $y_{x+1} \sim y_x$. Wir setzen eine im Durchschnitt gleichmäßige Auslastung eines Oberbegriffs voraus. Diese Annahme drückt eben die Klassifikationstendenz der Sprache aus.

(2) Die Zahl der Begriffe auf einer Ebene ist gleichzeitig auch der Ordnung der Ebene proportional, d.h. die Sprache besitzt Begriffe x-ter Ordnung in gewissen festen Proportionen, eine Wucherung oder Einschränkung der Zahl der Begriffe auf Ebenen $x \geq 2$ läßt die Sprache nicht zu. Formal ausgedrückt: $y_x \sim x$ für $x \geq 2$.

Setzen wir diese zwei Annahmen zusammen, so erhalten wir

$$y_{x+1} \sim (x + 1)y_x \qquad (1)$$

oder

$$y_{x+1} = a(x + 1)y_x \qquad (2)$$

wo a den Proportionalitätskoeffizienten darstellt. Die Differenzengleichung (2) können wir schrittweise lösen. Es ist

$$y_2 = 2ay_1$$
$$y_3 = 3ay_2 = 2(3)a^2y_1$$
$$y_4 = 4ay_3 = 2(3)4a^3y_1$$
$$\ldots\ldots\ldots\ldots\ldots$$

so daß wir allgemein

$$y_x = y_1 x! a^{x-1} \qquad (3)$$

erhalten.

Um zu überprüfen, ob diese Formel korrekt ist, müssen wir die Konstanten aus den Daten abschätzen.

(I) Die Konstante y_1 ist einfach die Zahl der Begriffe (Wörter) 1. Ordnung in der Stichprobe.

Die Konstante a kann man z.B. folgendermaßen bestimmen:

(a) Wegen $y_2 = 2y_1 a$ erhalten wir

$$\hat{a} = \frac{y_2}{2y_1}, \qquad (4)$$

wobei y_1 und y_2 die Stichprobenwerte sind. In unserem Fall bekämen wir

$$\hat{a} = \frac{348}{2(1723)} = 0.100987 \approx 0.1.$$

Mit Hilfe von $y_1 = 1723$ und $a = 0.100987$ bzw. $a = 0.1$ erhalten wir die Resultate in der Tabelle 2.

Tabelle 2. Beobachtete und berechnete Wortzahlen auf einzelnen Ebenen

x	beobachtet y_x	berechnet $\hat{y}_x$ $\hat{a} = 0.100987$	$\hat{a} = 0.1$
1	1723	1723.00	1723.00
2	348	348.00	344.60
3	108	105.43	103.38
4	39	42.59	41.35
5	13	21.50	20.68
6	3	13.03	12.40

Ohne Test vergleichen wir die Fehlerquote beider Anpassungen mit $SSE = \sum_x (y_x - \hat{y}_x)^2$. So ergibt sich für $\hat{a} = 0.100987$, SSE = 192.34 und für $\hat{a} = 0.1$, SSE = 185.77. Der abgerundete Wert liefert, wie man sieht, sogar eine etwas bessere Anpassung.

(b) Logarithmiert man beide Seiten von (3), so erhält man

$$\ln y_x = \ln y_1 + \ln (x!) + (x-1)\ln a.$$

Wendet man an diese Gleichung die Methode der kleinsten Quadrate an, so erhält man aus

$$\frac{\partial}{\partial \ln a}\left\{\sum_{x=1}^{n}[\ln y_x - \ln y_1 - \ln (x!) - (x-1)\ln a]^2\right\} = 0$$

schließlich

$$\ln \hat{a} = \frac{\sum_x (x-1)\ln y_x - \ln y_1 \sum_x (x-1) - \sum_x (x-1)\ln(x!)}{\sum_x (x-1)^2} \qquad (5)$$

In unserem Fall erhalten wir

$$\ln \hat{a} = \frac{41.9600 - 7.4518(15) - 65.8571}{55} = -2.4668$$

und

$$\hat{a} = 0.084856.$$

Mit diesem a erhalten wir die theoretischen Werte wie in Tabelle 3 angegeben.

Tabelle 3. Beobachtete und berechnete Wortzahlen auf einzelnen Ebenen

x	beobachtet y_x	berechnet $\hat{y}_x$
1	1723	1723.00
2	348	292.41
3	108	74.44
4	39	25.27
5	13	10.72
6	3	6.46

Diese Anpassung ist besser bei größeren Werten von x, im allgemeinen aber schlechter als die erste. Mit $\hat{a} = 0.084856$ ergibt sich die Fehlerquote von SSE = 4422.20, eine recht große Abweichung, die zeigt, daß diese Methode weniger geeignet ist.

(c) Eine schrittweise Verbesserung der Anpassung kann man folgendermaßen erhalten. Man entwickelt die Funktion $y_{x,a}$ mit dem Parameter a in eine Taylorreihe in der Umgebung von a_o und bekommt

$$y_{x,a} = y_{x,a_o} + \delta a \left.\frac{\partial y_{x,a}}{\partial a}\right|_{a=a_o} \qquad (6)$$

wo $\delta a = a - a_o$ ist. Zieht man auf beiden Seiten von (6) die Meßwerte y_x ab, so erhält man

$$y_{x,a} - y_x = y_{x,a_o} - y_x + \delta a \left.\frac{\partial y_{x,a}}{\partial a}\right|_{a=a_o}.$$

Auf der linken Seite stehen jetzt explizite die "Meßfehler" $\varepsilon_x = y_{x,a} - y_x$, die wir minimisieren wollen. Bezeichnen wir $y_x - y_{x,a_o} = L_x$ und $\left.\frac{\partial y_{x,a}}{\partial a}\right|_{a=a_o} = A_x$, so erhalten wir

$$\varepsilon_x = A_x \delta a - L_x.$$

Um nun δa zu finden, setzen wir die Ableitung (nach δa) der Summe der quadratischen Meßfehler gleich Null, d.h.

$$\frac{\partial}{\partial(\delta a)} \sum_x (L_x - A_x \delta a)^2 = 0 \qquad (7)$$

und erhalten

$$\delta a = \frac{\sum_x L_x A_x}{\sum_x A_x^2} = \frac{\sum_x \{(y_x - y_1 x! a_o^{x-1})[y_1 x!(x-1)a_o^{x-2}]\}}{\sum_x [y_1 x!(x-1)a_o^{x-2}]^2}. \qquad (8)$$

Daraus folgt dann

$$a = a_o + \delta a.$$

Setzt man in unserem Fall $a_o = 0.1$ ein, so erhält man nach (8)

$$\delta a = \frac{6168.4503}{18757617.31} = 0.000329$$

und daraus

$$a = 0.1 + 0.000329 = 0.100329.$$

Berechnet man mit diesem a die $\hat{y}_x$ Werte, so ergeben sich die Resultate in der Tabelle 4. Eine weitere Iteration ergibt $\delta a =$ $= -0.000008$, aber keine Verbesserung der Anpassung. Die Fehlerquote ergibt jetzt SSE = 183.85.

Tabelle 4.

x	$\hat{y}_x$
1	1723.00
2	345.73
3	104.06
4	41.76
5	20.95
6	12.61

(II) Möchte man die Kurve ohne Rücksicht auf y_1 gestalten, so besteht die Möglichkeit, die Kurve (3) als $y_x = cx!a^{x-1}$ zu betrachten und auch den Wert von c zu schätzen.

(a) Logarithmiert man diese Gleichung, so erhält man

$$\ln y_x = \ln c + \ln x! + (x-1)\ln a.$$

Mit Hilfe der Methode der kleinsten Quadrate erhält man dann

$$\ln \hat{a} = \frac{\Sigma(x-1)[\Sigma \ln x! - \Sigma \ln y_x] - n\,\Sigma(x-1)(\ln x! - \ln y_x)}{n\,\Sigma(x-1)^2 - [\Sigma(x-1)]^2}$$

$$\ln \hat{c} = \frac{\Sigma(x-1)[\Sigma(x-1)(\ln x! - \ln y_x)] - \Sigma(x-1)^2[\Sigma \ln x! - \Sigma \ln y_x]}{n\,\Sigma(x-1)^2 - [\Sigma(x-1)]^2}$$

In unserem Fall erhält man

$$\ln \hat{a} = \frac{15(17.0297 - 25.3133) - 6(23.8970)}{6(55) - 15^2} = -2.5489$$

$$\hat{a} = 0.078166$$

$$\ln \hat{c} = \frac{15(23.8970) - 55(17.0297 - 25.3133)}{6(55) - 15^2} = 7.7529$$

$$\hat{c} = 2328.2815$$

Mit diesen â und ĉ erhalten wir die Werte in der Tabelle 5, mit SSE = 367293.97, eine sehr schlechte Anpassung.

Tabelle 5.

x	$\hat{y}_x$
1	2328.28
2	363.98
3	85.35
4	26.69
5	10.43
6	4.89

(b) Benutzt man die Newtonsche iterative Methode für zwei Parameter, so gilt analog wie oben

$$y_{x,a,c} = a_{x,a_o,c_o} + \delta a \left.\frac{\partial y_{x,a,c}}{\partial a}\right|_{\substack{a=a_o \\ c=c_o}} + \delta c \left.\frac{\partial y_{x,a,c}}{\partial c}\right|_{\substack{a=a_o \\ c=c_o}}$$

oder $\quad a_{x,a,c} = y_{x,a_o,c_o} + \delta a\, A_x + \delta c\, C_x,$

und nach Abzug der Meßwerte a_x ergibt sich

$$\varepsilon_x = \delta a A_x + \delta c C_x - L_x.$$

Setzt man

$$\frac{\partial}{\partial(\delta a)} \sum_x (L_x - \delta a A_x - \delta c C_x)^2 = 0$$

$$\frac{\partial}{\partial(\delta c)} \sum_x (L_x - \delta a A_x - \delta c C_x)^2 = 0$$

so erhält man

$$\delta a \, \Sigma A_x^2 + \delta c \, \Sigma A_x C_x = \Sigma L_x A_x$$

$$\delta a \, \Sigma A_x C_x + \delta c \, \Sigma C_x^2 = \Sigma L_x C_x \, ,$$

woraus man δa und δc als

$$\begin{pmatrix} \delta a \\ \delta c \end{pmatrix} = \begin{pmatrix} \Sigma A_x^2 & \Sigma A_x C_x \\ \Sigma A_x C_x & \Sigma C_x^2 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma L_x A_x \\ \Sigma L_x C_x \end{pmatrix} \qquad (9)$$

berechnet. Hier ist

$$A_x = c_o x!(x-1)a_o^{x-2}$$

$$C_x = x! a_o^{x-1}$$

$$L_x = y_x - c_o x! a_o^{x-1}.$$

Die verbesserten Parameter erhält man als

$$a = a_o + \delta a$$

$$c = c_o + \delta c.$$

Man kann das Verfahren fortsetzen bis $\delta a_n = \delta c_n = 0$ werden. Setzt man die Anfangswerte als $a_o = 1723$ und $c_o = 0.1$, so erhält man $a = 0.100301$ und $c = 1723.4625$. Damit bekommt man die Resultate in der Tabelle 6, mit SSE = 183.49. Der Unterschied zu den Resultaten in Tab. 4 ist geringfügig.

Tabelle 6.

x	$\hat{y}_x$
1	1723.46
2	345.73
3	104.03
4	41.74
5	20.93
6	12.60

Zur Überprüfung reicht es die Signifikanz des Parameters a zu testen, der den Trend bestimmt. Verwenden wir dazu die Resultate des Verfahrens Ic, so haben wir

$$t_{n-2} = \frac{a\sqrt{\Sigma A_x^2}}{s} = \frac{0.100329\sqrt{18757617.31}}{\sqrt{183.85/4}} = 64.09,$$

einen Wert, der mit 4 Freiheitsgraden sehr hoch signifikant ist.

Verwenden wir Verfahren IIb, das fast identische Resultate liefert, so haben wir

$$t_{n-2} = \frac{a\sqrt{\Sigma A_x^2 \Sigma C_x^2 - (\Sigma A_x C_x)^2}}{s\sqrt{C_x^2}} =$$

$$= \frac{0.100301\sqrt{18757617.31(1.0444) - (857.4199)^2}}{\sqrt{183.49/4}\ \sqrt{1.0444}} =$$

$$= \frac{435.5338}{6.9217} = 62.92,$$

was wiederum mit dem obigen Resultat übereinstimmt.

Hier ist n = 6 die Zahl der Klassen von x(x = 1,2,...,6). Da die Wahrscheinlichkeit, einen so hohen oder noch extremeren t-Wert zu bekommen gleich P ≈ 0.0000002 ist, können wir die Anpassung als bestätigt betrachten.

4. Mit der Ableitung dieses Gesetzes, das wir als Martins Gesetz der Abstraktionsebenen bezeichnen wollen, ist jedoch ein Problem verbunden, das hier besprochen werden soll.

Bei der Stichprobenerhebung geraten viele (bei Martin alle) Wörter höherer Ordnungen in die Mengen der Wörter niedrigerer Ordnungen. Dies läßt sich nicht vermeiden, denn ein Wort höherer Ordnung kann mit gleicher Wahrscheinlichkeit wie alle anderen erhoben werden (wodurch es zu den Wörtern der ersten Ordnung gezählt wird), oder ein Wort z.B. sechster Ordnung kann zum Wort zweiter Ordnung werden, wenn ein nur durch es definiertes Wort direkt in der Stichprobe erscheint. Anders gesagt, es gibt Wörter, die gleichzeitig in mehreren y_x enthalten sind. Ein Beispiel aus MARTIN (1974) macht es ersichtlich:

Ordnung 6: chose, espace, façon

Ordnung 5: chose, espace, façon, fait, liquide, maison, manière, matière, morceau, nombre, objet, place, terrain

Ordnung 4 enthält alle 13 Wörter der Ordnung 5 und noch andere 26 Wörter usw.

Die Zählung kann aber auch so durchgeführt werden, daß jeder generische Begriff nur einmal gezählt wird und zwar nur in der höchsten Ordnung, in der er vorkommt. In dem Falle müssen wir die Zahlen in der Tabelle 1 so modifizieren, daß wir von jedem y_x das nächste y_{x+1} subtrahieren. So bekommen wir die Zahlen in der Tabelle 7.

Es stellt sich die Frage, ob man mit Formel (3) auch diesen Trend erfassen kann. Mit Hilfe von (4) erhalten wir â = 0.087273 und die Resultate in der dritten Spalte der Tabelle 7. Mit der Schätzung (5) erhalten wir â = 0.084626 und die Resultate in der vierten Spalte der Tabelle 7. Beide Anpassungen sind sehr gut, die erste ergibt $F_{1,4}$ = 319.69 mit P = 0.00006, die zweite $F_{1,4}$ =

Tabelle 7.

x	y_x	$\hat{y}_x$	$\hat{y}_x$
1	1375	1375.00	1375.00
2	240	240.00	232.72
3	69	62.84	59.08
4	26	21.94	20.00
5	10	9.57	8.46
6	3	5.01	4.30

= 385.48 mit P = 0.00004. Man würde intuitiv die erste Anpassung für besser halten; der bessere F-Test der zweiten Anpassung wird vor allem durch die Werte in x = 6 verursacht. Noch bessere Anpassungen ließen sich iterativ erreichen.

LITERATUR

HEMPEL, C.G., Aspects of scientific explanation. New York, The Free Press 1965, 155-171

MARTIN, R., Syntaxe de la définition lexicographique: étude quantitative des définissants dans le "Dictionnaire fondamental de la langue française". In: David, J., Martin, R. (Hrsg.), Statistique et linguistique. Paris, Klincksieck 1974, 61-71

LINGUOSTATISTISCHER VERGLEICH VON FUNKTIONALSTILEN DER LETTISCHEN SPRACHE

S. Kļaviņa, Riga

0.1 PROBLEM UND SEINE AUSGANGSPOSTULATE

Die gesamte Entwicklung der funktionellen Stilistik zeichnet sich durch das Bestreben aus, ihren zentralen Begriff, den des Funktionalstils,genau zu bestimmen. Es ist dabei zweckmäßig, die funktionell-stilistische Differenzierung der Rede als eine Art soziolinguistischen Variierens zu betrachten. Daher ergibt sich die Bestimmung des Funktionalstils als eine traditionsgemäße Gesamtheit sprachlicher Mittel, die in einem bestimmten Kommunikationsbereich gebraucht werden, und die Tendenzen der Wahl und Kombinierung verschiedener sprachlicher Mittel, derer man sich in einem bestimmten Kommunikationsbereich oder einer typischen Sprachsituation bedient.

Eine solche Auffassung vom Funktionalstil liegt der bekanntesten und traditionellsten Einteilung der Funktionalstile der Literatursprache zugrunde, bei der man den Umgangssprachstil, den Amtssprachstil, den publizistischen, den wissenschaftlichen und den künstlerischen Sprachstil der schönen Literatur unterscheidet. An dieses Klassifikationsschema hält sich auch die Stilkunde der baltischen Sprachen (des Lettischen und des Litauischen).

Die linguistische Spezifik jedes Funktionalstils ist jedoch ein recht unbestimmter Begriff, der Stil selbst aber tritt als eine vage Menge (engl. "fuzzy set") auf (ZADEH 1973; NALIMOV 1974:53; PIOTROVSKIJ ET AL. 1977:359-361), wodurch wiederum die Bestimmung und die Klassifizierung der Stile in hohem Maße erschwert werden. Die vorliegenden Beschreibungen der Funktionalstile beruhen bald

auf den Funktionen des Stils im Kommunikationsbereich, bald auf der Übereinstimmung der Rede mit der literarischen Form, bald auf der Wahl der Ausdrucksmittel usw. Im Zusammenhang damit tritt das Problem hervor, diejenigen linguistischen Merkmale aufzufinden und formell zu beschreiben, mit deren Hilfe die Zugehörigkeit eines Textes zu einem bestimmten Funktionalstil festgestellt wird.

Die Erfahrung zeigt, daß man vage Mengen mit Hilfe stochastischer Methoden analysieren kann. Das System der Stilmerkmale wird aufgrund der Häufigkeit sprachlicher Erscheinungen in Texten aufgestellt.

Die Hauptmasse jedes Textes machen allgemeinsprachliche stilirrelevante Einheiten aus, die in allen Funktionalstilen vorkommen. Sie bilden den Hintergrund des Stils und haben einen suprastilistischen Charakter. Die statistischen Charakteristiken ihres Gebrauchs weisen in Texten verschiedener Stile keine signifikanten Abweichungen auf.

Ein Funktionalstil wird durch die Summe stilprägender Mittel bestimmt. Die statistischen Charakteristika ihres Gebrauchs sind in Texten verschiedener Funktionalstile variabel, d.h. es werden signifikante Abweichungen statistischer Parameter beobachtet.

In den Verwendungshäufigkeiten von sprachlichen Elementen kommt die sprachliche Systemhaftigkeit zum Ausdruck.

Gleich G.HERDAN (1964:155-157), L.DOLEŽEL(1969:10-11), B.N. GOLOVIN (1968:39) und anderen Vertretern der statistischen Stilistik läßt sich der Stil als eine Wahrscheinlichkeitserscheinung interpretieren, und die Sprachstile kann man als sprachliche Varianten betrachten, die bestimmten typischen Situationen der Sprachkommunikation entsprechen und sich voneinander durch signifikante Unterschiede der Häufigkeiten sprachlicher Entitäten abheben.

0.2 ZIEL UND AUFGABEN DER UNTERSUCHUNG

Das Ziel der vorliegenden Arbeit ist es, die linguostatistisch relevanten Merkmale von drei Funktionalstilen der lettischen Schriftsprache (des publizistischen, wissenschaftlich-technischen

und künstlerischen Sprachstils) zu bestimmen und ihre Wechselbeziehungen im System der Kommunikationsmittel festzustellen. Das allgemeine Ziel der Untersuchung bestimmt auch ihre einzelnen Aufgaben:

erstens, die Ermittlung statistischer Charakteristika für jeden Funktionalstil, vor allem mit Hinblick auf die Verteilung der Häufigkeit lexikalischer und morphologischer Kategorien;

zweitens, die Feststellung differentieller und integraler Merkmale in diesen Charakteristika;

drittens, die Ermittlung der Wechselbeziehungen zwischen den Funktionalstilen im Bereich der Lexik und der Morphologie.

0.3 FORSCHUNGSMATERIAL

Für die Untersuchung ist das Material aus vier Häufigkeitswörterbüchern verwendet worden:

1) Häufigkeitswörterbuch der Publizistik (P1), nach der von der Gruppe "Sprachstatistik" ausgearbeiteten Methodik verfaßt (Gruppenleiter Prof. R.G.Piotrovskij), -- s. KĻAVIŅA 1968;
2) Häufigkeitswörterbuch des wissenschaftlich-technischen Sprachstils (WT), -- s. Latviešu valodas biežuma vārdnīca[1] 1966 und 1969;
3) Häufigkeitswörterbuch der Publizistik (P2), -- s. Latviešu valodas biežuma vārdnīca 1969;
4) Häufigkeitswörterbuch der Belletristik (B), -- s. Latviešu valodas biežuma vārdnīca 1972.

Die quantitativen Charakteristika der Wörterbücher sind in der Tabelle 1 angegeben.

Außerdem ist für die Ermittlung der statistischen Charakteristik der Affixe das rückläufige Wörterbuch der lettischen Sprache benutzt worden (SOIDA & KĻAVIŅA 1970).

Bei der stochastischen Modellierung wird der Text als Folge unabhängiger, diskreter Elemente und Realisation eines stationären Zufallsprozesses betrachtet. Solch vereinfachende Annahmen führen zu einem gewissen Schwund semantischer und syntaktischer Information, geben aber die Möglichkeit, die Verfahren der Wahrscheinlichkeitsrechnung und der mathematischen Statistik bei der Untersuchung anzuwenden.

Tabelle 1. Charakteristika der Häufigkeitswörterbücher P1, P2, WT und B

Charakteristik \ Häufigkeitswörterbuch	P1	P2	WT	B
Stichprobe (in Wörtern)	200000	300000	292000	300000
Vokabular (in Vokabeln)	16789	21069	13319	19696
Vokabeln, deren Häufigkeit die Konfidenz 95,5% hat	824	1175	1249	893
Statistik der Wortarten	vorhanden	vorhanden	vorhanden	nicht vorhanden
Statistik der morphologischen Kategorien	vorhanden	vorhanden	vorhanden	nicht vorhanden
Statistik der Affixe	vorhanden	für Verben u. Adjektive	für Verben u. Adjektive	nicht vorhanden

1. DER LEXIKALISCHE VERGLEICH VON FUNKTIONALSTILEN

1.0 Die Unterschiede der Stile, die im Zusammenhang mit den denotativ-signifikativen Unterschieden auch durch die Kommunikationssituation bedingt sind, offenbaren sich vor allem in der Lexik. Deswegen waren einige lexikalisch-statistische Aufgaben zu lösen:

erstens, die allgemeinbräuchliche Lexik mit Hilfe statistischer Kriterien festzustellen;

zweitens, diejenigen Vokabeln (types) statistisch zu ermitteln, die als differentielle, diagnostizierende Merkmale eines jeden Funktionalstils auftreten;

drittens, die Wechselbeziehungen zwischen den Stilen zu charakterisieren.

1.1 Zur Lösung der ersten und zweiten Aufgabe sind die Konfidenzintervalle der Vokabelhäufigkeit einem statistischen Vergleich unterzogen worden (s. PIOTROVSKIJ ET AL. 1977:269-277, 283-286).[2]

Die untere (p_u) und die obere (p_o) Vertrauensgrenze, zwischen denen bei einem gegebenen Sicherheitsgrad die unbekannte Wahrscheinlichkeit der Vokabel liegt, sind unter Verwendung folgender Formeln berechnet:

$$p_u = \frac{F + \frac{1}{2}Z\alpha^2 - Z\alpha\sqrt{F + \frac{1}{4}Z\alpha^2}}{N}$$

$$p_o = \frac{F + \frac{1}{2}Z\alpha^2 + Z\alpha\sqrt{F + \frac{1}{4}Z\alpha^2}}{N}$$

wo F -- die absolute Häufigkeit der Vokabel,
Zα-- der Koeffizient des gegebenen Konfidenzniveaus,
N -- der Stichprobenumfang ist.

Diese Formel ist eine vereinfachte Transformation der Formel des absoluten Fehlers von VAN DER WAERDEN 1960:45,(eine eingehende Ausführung der Transformation s. ALEKSEEV 1975:46-47).

Die Übereinstimmung zwischen den Konfidenzintervallen der Vokabelhäufigkeit in verschiedenen Häufigkeitswörterbüchern kann man gewiß nur in Vokabularteilen feststellen, deren Häufigkeit die Konfidenz 95,5% hat. Die berechnete untere Grenze dieser Zone bei 95,5%igem Konfidenzniveau ist p_u = 0,000147 (die untere Vertrauensgrenze der geringsten Häufigkeit auf dem Niveau 95,5%). Diese Größe wurde für die obere Häufigkeitsschwelle gewählt; die untere Häufigkeitsschwelle ist p_o = 0,000045 (die obere 95,5%ige Vertrauensgrenze der absoluten Häufigkeit F_i = 3).

Als allgemeingebräuchlich werden die Vokabeln betrachtet, deren obere Vertrauensgrenze der Häufigkeit in allen 4 Häufigkeitswörterbüchern die angenommene obere Häufigkeitsschwelle ($p_o \geq 0{,}000147$) übersteigt oder mit ihr gleich ist. Zu dieser Gruppe gehören sowohl Hilfswörter als auch eine Reihe von Begriffswörtern, z.B. *gads* 'das Jahr', *darbs* 'die Arbeit', *tauta* 'das Volk', *valsts* 'der Staat', *laiks* 'die Zeit', *cilvēks* 'der Mensch', *jauns* 'neu, jung', *liels* 'groß', *labs* 'gut', *pirmais* 'der erste', *viens* 'eins', *būt* 'sein', *varēt* 'können', *dot* 'geben', *sacīt* 'sagen', *notikt* 'geschehen' u.a.

Die Vokabeln, die nur in einem der Häufigkeitswörterbücher[3] $p_o \geq 0{,}000147$ haben, in den anderen dagegen die obere Vertrauensgrenze ihrer Häufigkeit die angenommene untere Häufigkeitsschwelle ($p_o \leq 0{,}000045$) nicht übersteigt, gelten als typisch, für einen Funktionalstil spezifisch.

Von solchem Sichtpunkt aus ist für den p u b l i z i s t i s c h e n S p r a c h s t i l folgendes charakteristisch:

1) gesellschaftspolitische Lexik, z.B. *prezidents* 'der Präsident', *plenums* 'das Plenum', *karaspēks* 'das Heer', *agentūra* 'die Agentur', *arodbiedrība* 'die Gewerkschaft', *sociālisms* 'der Sozialismus', *deputāts* 'der Deputierte' u.a.;

2) Substantive mit lokaler und temporaler Bedeutung, z.B. *apgabals* 'das Gebiet', *galvaspilsēta* 'die Hauptstadt', *gadadiena* 'der Jahrestag', *vēlēšanas* 'die Wahlen', *meistarsacīkstes* 'die Meisterschaftskämpfe' u.a.;

3) Strukturwörter, die Mitteilungen formen und zahlreiche stabile Wortverbindungen ohne phraseologischen Charakter bilden, z.B. *uzsvērt* 'betonen', *publicēt* 'veröffentlichen', *izcīnīt* 'erringen' u.a.; ihr Gebrauch ist oft eine Folge der Automatisation und Demantisierung der sprachlichen Mittel.

Typisch und für den w i s s e n s c h a f t l i c h - t e c h n i s c h e n S p r a c h s t i l spezifisch sind Vokabeln, die allgemeinwissenschaftliche (z.B. *lenķis* 'der Winkel', *frekvence* 'die Frequenz', *formula* 'die Formel', *vertikāls* 'vertikals', *horizontāls* 'horizontal', *izmērīt* 'ausmessen') und konkrete wissenschaftlich-technische (z.B. *spriegums* 'die elektrische Spannung', *sakausējums* 'die Legierung', *cirsma* 'der Holzschlag', *karjers* 'die Sandgrube', *vārpsta* 'die Welle', *sajūgs* 'die Kupplung') Begriffe bezeichnen.

Das statistische Kriterium ermöglicht es, die den b e l l e t r i s t i s c h e n S p r a c h s t i l prägenden Vokabeln zu er-

mitteln, darunter

1. Alltagslexik, z.B. *pusdiena* 'der Mittag', *cepure* 'die Mütze', *gulta* 'das Bett', *saimnice* 'die Wirtin', *ēst* 'essen' u.a.;
2. Bezeichnungen der Naturobjekte, z.B. *leja* 'das Tal', *pali* 'das Hochwasser', *strauts* 'der Bach', *salna* 'der Frost', *putekļi* 'der Staub' u.a.;
3. Gefühlsbezeichnungen, z.B. *sāpes* 'der Schmerz', *bēdas* 'der Kummer', *asara* 'die Träne', *smiekli* 'das Lachen', *just* 'fühlen' u.a.;
4. Benennungen der Verwandschaftsstufen, z.B. *māsa* 'die Schwester', *tēvs* 'der Vater', *mamma* 'die Mutter' u.a.;
5. Bezeichnungen der Körperteile, z.B. *lūpa* 'die Lippe', *piere* 'die Stirn', *vaigs* 'die Wange', *auss* 'das Ohr' u.a.;
6. Begriffe aus der Pflanzenwelt, z.B. *bērzs* 'die Birke', *priede* 'die Kiefer', *egle* 'die Fichte', *ziedēt* 'blühen', *smaržot* 'duften' u.a.;
7. Begriffe aus der Tierwelt, z.B. *vilks* 'der Wolf', *dzērve* 'der Kranich', *strazds* 'der Star' u.a.;
8. Farbenbezeichnungen, z.B. *pelēks* 'grau', *brūns* 'braun' u.a.;
9. temporale Adverbien und Adverbien mit einschätzender Bedeutung z.B. *kādreiz* 'einmal', *sonakt* 'heute nacht', *briesmīgi* 'schrecklich' u.a.;
10. Begrüßungswörter, z.B. *labdien!* 'guten Tag!', *sveiki!* 'lebe wohl!' u.a.

1.2 Die obenerwähnten Häufigkeitswörterbücher werden paarweise miteinander verglichen; die Ergebnisse des Vergleiches auf der Ebene des Vokabulars unter Verwendung des Jaccardschen Koeffizienten veranschaulicht die Tabelle 2 , die Vergleichsergebnisse auf der Ebene des Textes unter Verwendung des Pearsonschen Korrelationskoeffizienten die Tabelle 3.

Als Maß der lexikalischen Verknüpfung wird der aus der Biometrie bekannte Jaccardsche Koeffizient verwendet, der nach folgender Formel berechnet wird:

$$R = \frac{C}{V_a + V_b - C} ,$$

wo C -- die Zahl der gemeinsamen Vokabeln in beiden Wörterbüchern,

V_a -- die Zahl der Vokabeln im ersten Wörterbuch,

V_b -- die Zahl der Vokabeln im zweiten Wörterbuch ist.

Für das Messen der lexikalischen Verknüpfung können auch andere Verknüpfungsindizes verwendet werden (s. TULDAVA 1974:35-42; MULLER 1972:256-257; SOKAL, SNEATH 1963:129-139). Die Schwäche aller dieser Indizes liegt darin, daß sie vom Textumfang abhängig sind.

Da die Jaccardschen Koeffizienten für die Wörterbuchpaare aus den Stichprobenwerten berechnet wurden, ergibt sich daraus die Frage nach den Konfidenzintervallen dieser Koeffizienten für die Grundgesamtheit.

Die Koeffizienten werden als Stichprobenanteile betrachtet, unter der Voraussetzung, daß ihre Verteilung einer Normalverteilung entspricht; ihre Standardabweichung wird nach der Formel ermittelt:

$$\sigma_R = \sqrt{\frac{R\ (1 - R)}{V_a + V_b - C}}$$

Bei 95,5% Sicherheit bestimmt die Konfidenzgrenze des Jaccardschen Koeffizienten der Ausdruck $R \pm 2\ \sigma_R$. Die Konfidenzintervalle der Koeffizienten sind in der Tabelle 2 eingetragen.

Tabelle 2. Jaccardsche Koeffizienten (Indizes der lexikalischen Ähnlichkeit auf der Ebene des Vokabulars) und ihre Konfidenzintervalle (95,5%)

	P2	WT	B
P1	0,440 0,433 ÷ 0,447	0,262 0,265 ÷ 0,268	0,295 0,289 ÷ 0,301
P2		0,265 0,259 ÷ 0,271	0,356 0,350 ÷ 0,362
WT			0,187 0,182 ÷ 0,192

Der Vergleich der Konfidenzintervalle des Jaccardschen Koeffizienten (s. Tabelle 2) zeigt, daß sich nur die Intervalle der Koeffizienten von P_1 : WT und P_2 : WT überschneiden, die übrigen Intervalle überschneiden sich nicht, folglich sind ihre Unterschiede signifikant. Aus dem Vergleich läßt sich schließen, daß der wissenschaftlich-technische Sprachstil seinem Vokabular nach im Vergleich mit dem belletristischen und dem publizistischen Sprachstil separat steht. Dabei kann man von lexikalischer Ähnlichkeit des publizistischen und des belletristischen Sprachstils sprechen, die in diesem Sinne in Opposition zum wissenschaftlich-technischen Sprachstil stehen. Diese Verhältnisse sind hauptsächlich durch denotativ-signifikative Faktoren bedingt. Den Inhalt des wissenschaftlich-technischen Sprachstils macht die Produktions- und Forschungstätigkeit der Gesellschaft aus, den Inhalt der Belletristik und der Publizistik bilden jedoch die verschiedenartigen Gebiete der menschlichen Tätigkeit; außerdem sind der belletristische und der publizistische Sprachstil durch eine gemeinsame kommunikative Funktion verbunden, und zwar die Funktion der Einwirkung, obgleich die Wirkungsmittel beider Stilbereiche unterschiedlich sind.

Die Messung der lexikalischen Ähnlichkeit auf der Textebene, d.h. mit der Beachtung der Häufigkeiten der Vokabeln, wird mit Hilfe des Pearsonschen Korrelationskoeffizienten durchgeführt.

Der Pearsonsche Korrelationskoeffizient wird nach der Formel berechnet:

$$r = \frac{\sum_{i=1}^{n} a_i b_i - \frac{\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i}{n}}{\sqrt{\sum_{i=1}^{n} a_i^2 - \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{n}} \sqrt{\sum_{i=1}^{n} b_i^2 - \frac{\left(\sum_{i=1}^{n} b_i\right)^2}{n}}}$$

wo a_i -- der Wert der Vokabelhäufigkeit in P1,
b_i -- der Wert der Vokabelhäufigkeit in P2 (oder in WT, oder in B),
n -- die Zahl der Korrelanten (Vokabeln) ist.

Tabelle 3. Pearsonsche Korrelationskoeffizienten (Indizes der lexikalischen Ähnlichkeit auf der Textebene) und ihre Konfidenzintervalle (95,5%)

	P2	WT	B
P1	0,979 0,979 ÷ o,981	0,862 0,854 ÷ o,864	0,801 o,793 ÷ o,808

Die Stichprobenverteilung des Korrelationskoeffizienten unterscheidet sich signifikant von der Normal- und Student-Verteilung. Der Unterschied wächst umso schneller, je kleiner der Stichprobenumfang n wird. Dies kommt besonders zum Ausdruck bei $|r| > 0{,}50$. Die Schiefe der Verteilung des Korrelationskoeffizienten macht die Prüfung seiner Signifikanz und die Berechnung des Konfidenzintervalles nach dem t - Test unmöglich. Es wird daher nötig, die Verteilung der r - Werte durch eine Transformation zu "normalisieren". Dazu dient die Fischersche z - Transformation:

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r} \quad .$$

Man kann der Tafel für beliebige Werte von r (z.B. CLAUß, EBNER 1967:351, Tafel 10) die entsprechenden z - Werte unmittelbar entnehmen. Der Stichprobenfehler für z hängt nur von der Stichprobengröße n ab und ist nach der Formel

$$S_z = \frac{1}{n - 3}$$

zu ermitteln

Die Signifikanz z, dadurch auch r, lassen sich durch den t - Test prüfen.

$$t = \frac{z}{S_z}$$

Der berechnete und der theoretische Wert von t werden bei Irrtumswahrscheinlichkeit $\alpha = 5\%$ und der Zahl der Freiheitsgrade $\nu = n - 2$ (in konkreten Fällen ist es ∞) verglichen. Der empirische t - Wert ist für alle drei Häufigkeitswörterbücherpaare beträchtlich größer als der theoretische Wert $t_{0,05;\infty} = 1{,}96$, als

auch $t_{0,01;\infty}$ = 2,58 (s. Tabelle 4), folglich sind die Korrelationen dieser Häufigkeitswörterbücher signifikant.

Die Vertrauensgrenzen für z werden bei α = 5% nach der Beziehung $z = \pm 1{,}96 \cdot S_z$ bestimmt. Die Rücktransformation von z in r (nach der Tafel) läßt die Vertrauensgrenzen für den Pearsonschen Korrelationskoeffizienten der Grundgesamtheit bestimmen (s. Tabelle 3 und 4).

Tabelle 4. Prüfung der Pearsonschen Korrelationskoeffizienten (Lösungsweg)

	r	z	n	s_z	t	$z_u \div z_o$	$r_u \div r_o$
P1:P2	0,979	2,27	8542	0,0108	210	2,25÷2,29	0,978÷0,980
P1:WT	0,862	1,2933	10646	0,0096	134	1,27÷1,31	0,854÷0,864
P1:B	0,801	1,0986	10433	0,0097	113	1,08÷1,12	0,793÷0,808

Die Korrelationskoeffizienten werden unter Verwendung der Fischerschen z - Transformation und des t - Testes verglichen (s. Tabelle 5):

$$t = \frac{z_1 - z_2}{s_{z_1 - z_2}} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Tabelle 5. Prüfung der Differenz zwischen den Korrelationskoeffizienten

$r_1 - r_2$	t
$r_{P1:P2} - r_{P1:WT}$	72,88
$r_{P1:P2} - r_{P1:B}$	84,79
$r_{P1:WT} - r_{P1:B}$	14,13

In allen Fällen erhält man $t_{emp} > t_{0,05;\infty} = 1{,}96$ und

$t_{emp} > t_{0,01;\infty} = 2{,}58$. Das zeugt von der Signifikanz der Differenz zwischen den Korrelationskoeffizienten.

Der Pearsonsche Korrelationskoeffizient zeigt auf der Textebene eine merkliche Annäherung publizistischer und wissenschaftlicher Texte (vgl. $r_{P1:WT}$ mit $r_{P1:B}$). Der belletristische Stil bleibt gewissermaßen abseits stehen. Dieses Korrelationsverhältnis der Stile zeugt von einer bedeutenden Einwirkung der Unterschiede des Kommunikationstyps, der kommunikativen Funktion und der Kommunikationsbedingungen auf die lexikalisch-statistische Struktur. Der wissenschaftlich-technische Sprachstil ist ein eigentlich kommunikativer, offizieller Stil, der die Funktion der Mitteilung ausübt. Im publizistischen Sprachstil aber koexistieren Genres mit meinungsbildender Einwirkung. Der belletristische Sprachstil hat die Funktion der ästhetischen Einwirkung und die kommunikative Funktion inne, in der Figurensprache und in der Autorensprache finden wir eine Synthese der interpersonalen und der Massenkommunikation. Folglich ergeben sich bei der Einschätzung der Ähnlichkeit von lexikalisch-statistischer Struktur der Texte andere Resultate als bei der Einschätzung der lexikalischen Ähnlichkeit allein.

2. DER MORPHOLOGISCHE VERGLEICH DER STILE

2.0 Die morphologischen Eigentümlichkeiten der Funktionalstile sind vorläufig am wenigsten erforscht, da der grammatische Bau der Sprache eine lange Zeit als stilneutral betrachtet wurde. Die stilstatistischen Untersuchungen haben jedoch beträchtliche Differenzen im Gebrauch mancher grammatischer Kategorien aufgedeckt. In diesem Zusammenhang ist von uns das Vorkommen der grammatischen Kategorien in drei Funktionalstilen der lettischen Sprache statistisch untersucht worden. Demgemäß hat man im Rahmen der allgemeinen Forschungsaufgabe folgende Sonderaufgaben zu lösen:

erstens, differentielle morphologisch-statistische Merkmale der Stile zu ermitteln,

zweitens, die Wechselbeziehungen zwischen den Stilen im Be-

reich der Morphologie festzustellen.

2.1 Die morphologischen Unterschiede der Funktionalstile sind unter Verwendung des Vergleiches der 95,5%igen Konfidenzintervalle der morphologischen Kategorien eingeschätzt worden. Ein signifikanter Unterschied liegt vor, wenn sich zwei Konfidenzintervalle nicht überschneiden.

Tabelle 6. Konfidenzintervalle (95,5%) für die Häufigkeiten der Wortarten (prozentual von der Gesamtzahl der Wörter)

Wortarten	P1 N = 200000	P2 N = 300000	WT N = 292000	B N = 300000
Substantiv	48,92÷49,36	43,62÷43,98	46,22÷46,58	27,71÷28,03
Adjektiv	7,26÷ 7,50	7,01÷ 7,19	8,60÷ 8,80	5,36÷ 5,52
Numerale	1,41÷ 1,51	1,65÷ 1,75	1,26÷ 1,34	1,16÷ 1,24
Pronomen	6,87÷ 7,09	8,39÷ 8,61	5,61÷ 5,79	14,50÷14,76
Verb	15,67÷15,99	17,26÷17,54	17,46÷17,74	23,03÷23,33
Adverb	5,42÷ 5,64	6,31÷ 6,49	4,62÷ 4,78	9,44÷ 9,64
Partikel	2,17÷ 2,31	1,85÷ 1,95	1,06÷ 1,14	4,50÷ 4,66
Präposition	4,97÷ 5,17	5,12÷ 5,28	6,91÷ 7,09	5,13÷ 5,29
Konjunktion	6,23÷ 6,45	5,91÷ 6,09	6,41÷ 6,59	7,08÷ 7,26
Interjektion	0,023÷0,037	0,042÷0,058	0,002÷0,006	0,51÷ 0,57

2.1.1 Die Intervallwerte zeigen, daß die Funktionalstile beträchtliche quantitative Differenzen im Vorkommen d e r W o r t a r t e n auf der Textebene aufweisen.
Ein differentielles Merkmal der Funktionalstile sind die Verhältnisse der Zahl, besonders der Häufigkeit von Verben und Substantiven. Im Vergleich mit dem belletristischen Sprachstil (B) ist für den publizistischen (P1, P2) und den wissenschaftlich-technischen (WT) Sprachstil ein zahlenmäßiges Übergewicht der Substantive über die Verben kennzeichnend (s.Tabelle 6).
Unseres Erachtens ist der aktive Gebrauch der Substantive, darunter der deverbalen Substantive, eine Folge des geschwächten verbalen Charakters der Wortfügungen und der Tendenz, Prozesse

als substantivierte Ereignisse darzustellen. In der Normalisierungstendenz kommt offensichtlich der abstrakt-verallgemeinernde Charakter und eine gewisse Statik dieser Funktionalstile (P1, P2, WT) zum Ausdruck.[4] Dem konkret-bildlichen Charakter der Belletristik, der Emotionalität des Inhalts und der Dynamik der Erzählung entspricht aber der häufige Gebrauch von Verben, hauptsächlich als Prädikat in der Personalform. Folglich ist die Häufigkeit der Verben und der mit ihnen syntaktisch verbundenen Wortarten (Pronomen, Adverbien, Partikeln) in den belletristischen Texten (B) höher als in den anderen Stilbereichen.
Den wissenschaftlich-technischen Sprachstil (WT) kennzeichnet seine Tendenz zur Präzision und Eindeutigkeit. Dieser Zug äußert sich im aktiven Gebrauch der Präpositionen, der Wortart, die die syntaktische Verbindung zwischen den Wortformen und die syntaktischen Funktionen der Satzglieder verdeutlicht. Der Häufigkeit der Präpositionen nach stehen die wissenschaftlich-technischen Texte unter den von uns erforschten an der ersten Stelle, sie unterscheiden sich wesentlich dem Anteil der Wörter nach von den belletristischen und publizistischen Texten.[5] Die Tendenz zur Exaktheit im wissenschaftlich-technischen Sprachstil kommt auch durch aktiven Gebrauch der Adjektive zum Ausdruck. In wissenschaftlich-technischen Texten ist ihre Häufigkeit höher als in den anderen Bereichen.[6]
Ein differentielles Merkmal des belletristischen Sprachstils ist die Zahl der Interjektionen, sie übertrifft signifikant den geringen Anteil dieser Wortart in den anderen Sprachstilen.

2.1.2 Die statistischen Korrelationsverhältnisse zwischen Funktionalstilen in Bezug auf die Häufigkeit der Wortarten sind unter Verwendung der Pearsonschen Korrelationskoeffizienten untersucht worden .
Die Werte der Pearsonschen Korrelationskoeffizienten werden in der Tabelle 7 angegeben.

Tabelle 7. Pearsonsche Korrelationskoeffizienten und ihre Konfidenzintervalle (95,5%) für den Gebrauch der Wortarten (n = 10)

	P2	WT	B
P1	0,997 0,986 ÷ 0,999	0,996 0,982 ÷ 0,999	0,857 0,493 ÷ 0,965
P2		0,995 0,978 ÷ 0,999	0,890 0,592 ÷ 0,974
WT			0,860 0,501 ÷ 0,966

Tabelle 8. Prüfung der Differenz zwischen Korrelationskoeffizienten ($n_1 = n_2 = 10$)

r_1-r_2	t	r_1-r_2	t	r_1-r_2	t
$r_{P1:P2}-r_{P1:WT}$	0,187	$r_{P1:WT}-r_{P2:WT}$	0,187	$r_{P1:B}-r_{WT:B}$	0,019
$r_{P1:P2}-r_{P2:WT}$	0,374	$r_{P1:WT}-r_{P1:B}$	3,405	$r_{P2:WT}-r_{P1:B}$	3.218
$r_{P1:P2}-r_{P1:B}$	3,592	$r_{P1:WT}-r_{P2:B}$	3,143	$r_{P2:WT}-r_{P2:B}$	2,956
$r_{P1:P2}-r_{P2:B}$	3,330	$r_{P1:WT}-r_{WT:B}$	3,386	$r_{P2:WT}-r_{WT:B}$	3,199
$r_{P1:P2}-r_{WT:B}$	3,573	$r_{P1:B}-r_{P2:B}$	0,262	$r_{P2:B}-r_{WT:B}$	0,243

$t_{0,05;10} = 2,228$, $t_{0,01;10} = 3,169$

Aus dem Vergleich der berechneten r - Werte mit den kritischen r - Werten des Pearsonschen Koeffizienten ergibt sich eine signifikante Korrelation zwischen allen Paaren der Funktionalstile dem Gebrauch der Wortarten nach bei $\alpha = 0,01$, weil $r > r_{0,01;10} =$

0,765. Andererseits zeigt der Vergleich der Werte der Korrelationskoeffizienten (s. Tabelle 8) unter Verwendung der Fischerschen Transformation z und des t - Tests, daß die Werte der Korrelationskoeffizienten $r_{P1:WT}$ und $r_{P2:WT}$ signifikant höher als die Werte $r_{P1:B}$, $r_{P2:B}$ und $r_{WT:B}$ sind.
Folglich ist die Ähnlichkeit im Gebrauch der Wortarten zwischen den publizistischen und wissenschaftlich-technischen Texten stärker als zwischen den publizistischen und belletristischen Texten.

2.2 Die Häufigkeit d e r K a s u s f o r m e n der Substantive erweist sich als ein differentielles Merkmal der Stile. Die statistischen Daten ihres Gebrauchs zeigt die Tabelle 9, die Rangordnung von Kasus ihrer Häufigkeit nach zeigt die Tabelle 10.

Tabelle 9. Konfidenzintervalle (95,5%) für die Häufigkeiten der Kasusformen der Substantive (prozentual von der Wortzahl der Substantive)

Zahl	Kasus[7]	P1 N=96506	P2 N=131341	WT N=136056	BD[8] N=16776
Singular	N	22,75÷23,25	20,58÷21,2	18,97÷19,42	36,53÷38,47
	G	41,61÷42,39	39,53÷40,07	44,08÷44,72	20,38÷22,02
	D	4,48÷ 4,72	5,38÷ 5,62	5,66÷ 5,94	4,08÷ 4,92
	A	13,8 ÷14,2	17,49÷17,91	13,99÷14,41	18,61÷20,19
	I	2,21÷ 2,39	2,71÷ 2,89	4,08÷ 4,32	1,63÷ 2,17
	L	13,8 ÷14,2	13,12÷13,48	12,0 ÷12,4	12,62÷13,98
	V	0,08÷ 0,12	0,08÷ 0,12	-	2,47÷ 3,13
Plural	N	21,95÷22,45	26,06÷26,54	17,98÷18,42	32,08÷34,92
	G	39,11÷39,69	34,63÷35,17	36,33÷36,87	17,63÷19,97
	D	12,7 ÷13,1	13,41÷13,79	15,6 ÷16,0	15,19÷17,41
	A	13,01÷13,39	13,12÷13,48	15,7 ÷16,1	11,68÷13,68
	I	3,0 ÷ 3,2	3,4 ÷ 3,6	4,09÷ 4,31	3,69÷ 4,91
	L	8,64÷ 8,96	8,16÷ 8,44	9,13÷ 9,47	12,18÷14,22
	V	0,37÷ 0,43	0,08÷ 0,12	-	0,78÷ 1,4

Tabelle 10. Rangordnung der Kasushäufigkeiten

Rang	Singular				Plural			
	P1	P2	WT	B_D	P1	P2	WT	B_D
1	G	G	G	N	G	G	G	N
2	N	N	N	G	N	N	N	G
3	A;L	A	A	A	A;D	A;D	A;D	D
4		L	L	L				L;A
5	D	D	D	D	L	L	L	
6	I	I	I	V	I	I	I	I
7	V	V	-	I	V	V	-	V

Es sei bemerkt, daß sich die Funktionalstile dem Gebrauch von Kasusformen des Nominativs und Genitivs nach in künstlerische (B) und nicht künstlerische (P1, P2, WT) einteilen lassen. Diese Schlußfolgerung stimmt mit der von E.G. RIESEL 1975:9 vorgeschlagenen Grobklassifikation aller Systeme der Funktionalstile überein.
Die Ursache der verschiedenen Kasushäufigkeiten in einzelnen Funktionalstilen sind unseres Erachtens nach die grammatischen Beziehungen der Wortformen.[9] Der Genitiv ist adnominal, der Nominativ ist adverbal. Der nominale Charakter der wissenschaftlich-technischen und der publizistischen Texte bewirkt offenbar das Dominieren vom Genitiv in WT, P1 und P2. Der verbale Charakter der belletristischen Texte bedingt seinerseits den intensiven Gebrauch des Nominativs in der gebundenen Rede (B_D).

2.3 Einige differentielle Merkmale zeigt die Statistik von grammatischen Formen der Verben.

a) Das quantitative Verhältnis d e r P e r s o n a l f o r m e n u n d d e r N o m i n a l f o r m e n (s. Tabelle 11).

Tabelle 11. Konfidenzintervalle (95,5%) für die Häufigkeiten der Personalformen und der Nominalformen (prozentual von der Wortzahl der Verben)

Formen	P1 N=31163	P2 N=52109	WT N=51429	B_D N=11542
Personale	71,63÷72,77	71,24÷71,96	63,9 ÷64,82	80,85÷82,41
Nominale	26,75÷28,85	27,75÷29,07	34,72÷36,48	17,59÷19,15
darunter: Infinitiv	11,21÷11,99	12,75÷13,25	7,06÷ 7,54	8,79÷ 9.95
Partizip[10]	12,37÷ 13,23	12,74÷13,26	20,84÷21,56	5,49÷ 6,43
Adverbial-partizip	3,17÷ 3,63	3,26÷ 3,56	6,82÷ 7,38	2,7 ÷ 3,38

Obwohl die Personalformen in allen Funktionalstilen dominieren, fällt der größte Anteil davon auf die belletristischen Texte. Die Nominalformen kennzeichnen ihrerseits die wissenschaftlich-technischen Texte. In allen Stilen werden die Partizipien bevorzugt, dabei liegen die Publizistik und die gebundene Rede dem Partizipiengebrauch nach wesentlich hinter der Wissenschaft und der Technik. Dem Gebrauch der Adverbialpartizipien nach übertreffen die wissenschaftlichen und die technischen Texte beträchtlich die Publizistik und die gebundene Rede. Die Gebrauchsaktivität des Infinitivs ist in den wissenschaftlich-technischen Texten bedeutend niedriger als in der Publizistik.
Folglich sind die wissenschaftlich-technischen Texte durch eine Überzahl der Nominalformen gekennzeichnet.[11] Diese Erscheinung wird durch die sogenannte syntaktische Kondensation bestimmt, worunter man im gegebenen Falle das Ersetzen der Nebensätze, hauptsächlich der Adverbialsätze, durch Partizipialkonstruktionen versteht. Diese Tendenz wird im Bereich der Wissenschaft durch die Tendenz zur Dichte und Kürze der Aussage bewirkt.

b) Die Verteilung der Personalformen u n t e r d e n M o d i (s. Tabelle 12)

Tabelle 12. Konfidenzintervalle (95,5%) für die Häufigkeiten der Modi (prozentual von der Wortzahl der gebrauchten Personalformen)

Modus	P1 N=22437	P2 N=37310	WT N=33120	B_D N=9463
Indikativ	88,09÷88,91	89,69÷90,71	89,57÷90,23	86,7 ÷88,02
Konjunktiv	5,41÷ 5,99	3,4 ÷ 3,8	3,89÷ 4,31	2,59÷ 3,27
Debitiv	3,65÷ 4,15	4,47÷ 4,97	5,85÷ 6,35	1,53÷ 2,05
Imperativ	0,87÷ 1,13	0,99÷ 1,21	0,009÷0,011	7,22÷ 8,3
Modus relativus	0,77÷ 1,03	0,33÷ 0,47	0,019÷0,021	0,08÷ 0,24

Es ist aus der Tabelle ersichtlich, daß etwa 90% der Personalformen der Verben in allen Texten im Indikativ erscheinen. Die Zahl der Konjunktiv- und Debitivformen ist in den publizistischen und den wissenschaftlich-technischen Texten gering, der Anteil des Imperativs schwankt in der Publizistik etwa um 1%, in den wissenschaftlichen und technischen Texten ist er außerordentlich gering.
Anders ist die Verteilung der Personalformen unter den Modi in der gebundenen Rede. Neben dem offensichtlichem Dominieren des Indikativs ist hier der Gebrauch des Imperativs beträchtlich. Der Anteil dieses Modus übertrifft signifikant die entsprechenden Werte in den anderen Funktionalstilen. Dem Gebrauch des Debitivs und des Konjunktivs nach liegt die gebundene Rede aber hinter dem wissenschaftlich-technischen und dem publizistischen Sprachstil. Es gibt mehr Formen des Modus relativus in der gebundenen Rede als in den wissenschaftlich-technischen, weniger aber als in den publizistischen Texten.

c) Die Häufigkeit d e s G e n u s . Verb: In allen Funktionalstilen wird der Vorrang zweifelsohne dem Aktiv eingeräumt (s. Tabelle 13). Der Gebrauch des Passivs weist in den wissenschaftlich-technischen Texten eine signifikant höhere Häufigkeit auf als in der Publizistik und Belletristik (in der gebundenen Rede).

Tabelle 13. Konfidenzintervalle (95,5%) für die Häufigkeiten der Aktiv- und Passivformen (prozentual von der Wortzahl der Personalformen)

Genus	P1 N=22437	P2 N=37310	WT N=33120	B_D N=9463
Aktiv	93,56÷94,24	91,53÷92,07	87,34÷88,06	97,28÷97,90
Passiv	5,76÷ 6,44	7,93÷ 8,47	11,94÷12,66	2,10÷ 2,72

d) Die Häufigkeit d e r Z e i t f o r m e n ist nur auf der Basis des Indikativs betrachtet worden, weil etwa 90% von den Personalformen des Verbs im Indikativ gebraucht werden. Die Häufigkeit der Zeitformen ist in allen Funktionalstilen stark unterschiedlich (s. Tabelle 14).

Tabelle 14. Konfidenzintervalle (95,5%) für die Häufigkeiten der Zeitformen des Indikativs (prozentual von der Wortzahl im Indikativ)

Zeitebene	P1 N=18756	P2 N=33654	WT N=29742	B_D N=9006
Gegenwart	66,06÷67,94	65,93÷66,87	92,34÷92,86	73,18÷74,92
Vergangenheit	22,95÷24,65	26,36÷27,24	4,43÷ 4,85	14,25÷15,67
Zukunft	9,4 ÷10,6	6,45÷ 6,95	2,53÷ 2,85	10,37÷11,61

In allen drei Stilen dominieren die Gegenwartsformen, die Vergangenheits- und Zukunftsformen kommen beträchtlich seltener vor. Jedoch ist das zahlenmäßige Übergewicht der Gegenwartsformen in den wissenschaftlichen Texten viel deutlicher als in der

Publizistik und in der gebundenen Rede.
Die Vergangenheitsformen werden relativ häufig in den publizistischen Texten gebraucht, die Zukunftsformen eher in der gebundenen Rede.
Somit ist das quantitative Verhältnis der Zeitformen für jeden Funktionalstil spezifisch, folglich ist es ein signifikantes differentielles Merkmal der Stile.
Dabei übertrifft die Zahl der Wörter in einfachen Zeitformen in allen Stilen wesentlich die Zahl der Wörter in den zusammengesetzten Zeitformen (s. Tabelle 15). Das Verhältnis des Anteils der einfachen und der zusammengesetzten Zeitformen ist in der Publizistik und in den wissenschaftlich-technischen Texten ähnlich, in der gebundenen Rede aber unterscheiden sich ihre Proportionen -- hier sind die synthetischen Formen in der Überzahl.

Tabelle 15. Konfidenzintervalle (95,5%) für die Häufigkeiten der einfachen und zusammengesetzten Verbalformen (prozentual von der Wortzahl der Personalformen)

Formen	P1 N=22437	P2 N=37310	WT N=33120	BD N=9463
Einfache	80,59÷81,61	77,66÷80,44	79,56÷80,44	92,81÷93,81
Zusammengesetzte	18,39÷19,41	21,55÷22,45	19,56÷20,44	6,19÷ 7,19

e) Die Verteilung der Verbalformen auf d i e P e r s o n e n ist auf der Basis des Indikativs berechnet (s. Tabelle 16).

Person,Zahl	P1 N=18756	P2 N=33654	WT N=29742	BD N=9006
1. Singular	0,67÷ 0,93	2,74÷ 3,06	0,003÷0,009	14,01÷15,43
1. Plural	3,44÷ 3,96	3,61÷ 3,99	1,94÷ 3,99	4,93÷ 5,83
2. Singular	0,22÷ 0,38	0,24÷ 0,36	-	6,74÷ 7,78
2. Plural	0,22÷ 0,38	0,43÷ 0,57	+ 0,04	1,35÷ 1,85
3.	94,49÷95,11	92,03÷92,57	96,91÷97,27	70,07÷71,89

In der Verteilung der Verbalformen auf die Personen offenbaren sich die charakteristischen Züge eines jeden Funktionalstils. Das Gemeinsame dabei ist, daß die Hauptmasse der Personalformen des Verbs in der 3. Person steht (in WT beinahe 100%). Nicht hoch, doch wesentlich genug ist die Häufigkeit der 1. Person des Plurals in den publizistischen und den wissenschaftlich-technischen Texten. Der angegebene Prozentsatz ist durch den Gebrauch des Plurals der Gemeinschaftlichkeit in der Publizistik und des Plurals der Bescheidenheit in den wissenschaftlichen Texten zu erklären.
Auch in den publizistischen Texten findet man in einer geringen Anzahl die 1. Person Singular sowie die 2. Person Singular und Plural.
In der wissenschaftlichen Literatur wird die 2. Person Singular überhaupt nicht gebraucht. Das Vorkommen der 2. Person Plural und der 1. Person Singular beobachtet man in Einzelfällen.
Wesentlich anders sind die quantitativen Verhältnisse und die Rangordnung der Personalformen in der gebundenen Rede. Die häufigste Form ist hier die 3. Person, ebenso wie in den anderen Funktionalstilen. Darauf folgt aber die 1. Person Singular, wodurch die dichterische Individualität zu ihrem poetischen Ausdruck kommt. Erhöhte Expressivität und Emotionalität, die Funktionen der Einwirkung und der Kommunikation sind auch mit der Gebrauchsintensivität der 2. Person Singular verbunden. Die Formen der 1. und 2. Person Plural kommen in der gebundenen Rede wesentlich seltener als die entsprechenden Singularformen vor.
Die Verteilung der Verbalformen unter die Personen ist folglich ein markantes stilprägendes Merkmal.

2.4 Die Verteilung d e r P e r s o n a l p r o n o m e n (s. Tabelle 17) entspricht der quantitativen Verteilung der Personalformen des Verbs und bestätigt auf solche Weise die stilprägende Funktion dieser Spracherscheinungen.

Tabelle 17. Konfidenzintervalle (95,5%) für die Häufigkeiten der Personalpronomen (prozentual von der Wortzahl der Personalpronomen)

Person		P1 N=3257	P2 N=3773	WT N=126	B N=19566	B_D N=3290
1.Sg.--es	(ich)	3,39÷ 4,99	16,51÷18,79	-	32,8 ÷33,8	39,41÷43,35
1.Pl.--mēs	(wir)	38,66÷42,58	18,86÷21,3	53,17÷70,65	11,87÷12,53	11,86÷14,56
2.Sg.--tu	(du)	0,28÷ 0,88	1,45÷ 2,25	-	21,31÷22,13	23,92÷27,42
2.Pl.—jūs	(ihr)	2,97÷ 4,49	5,43÷ 6,87	1,95÷10,73	10,07÷10,69	4,08÷ 5,82
3. —viņas	(er) (es)	48,88÷52,88	52,05÷55,05	23,37÷40,13	21,98÷22,82	13,38÷16,22
viņa	(sie)					
viņi	(sie)					
viņas						

Tabelle 18. Rangordnung der Personalpronomen in Texten

Rang	P1	P2	WT	B_D
1	viņš	viņš	mēs	es
2	mēs	mēs	viņš	viņš
3	es	es	jūs	tu
4	jūs	jūs	-	mēs
5	tu	tu	-	jūs

Die vorliegende Tabelle (s. Tabelle 18) zeigt, daß jeder Stil seine eigene Rangordnung der Personalpronomen hat, die Sprache der Wissenschaft wird außerdem durch die Gesamtzahl der gebrauchten Personalpronomen gekennzeichnet: hier fehlen überhaupt die Pronomen für die 1. und 2. Person Singular.
Unentbehrlich für den künstlerischen Sprachstil sind die Pronomen "*es*" (ich), "*tu*" (du); als typisch für die Sprache der Wissenschaft ist "*mēs*" (wir) zu betrachten. Im Vergleich zu den anderen Texten wurde der höchste Prozentsatz im Gebrauch der Pronomen der 3. Person in den publizistischen Texten festgestellt.

2.5 Eine Reihe von spezifischen Zügen der Funktionalstile findet man im Gebrauch der Funktionsgruppen d e r P r o n o m e n (s. Tabelle 19).

Tabelle 19. Konfidenzintervalle (95,5%) für die Häufigkeiten der Funktionsgruppen der Pronomen[12] (prozentual von der Wortzahl der Pronomen)

Funktionsgruppen	P1 N=13743	P2 N=25640	WT N=16613	B_D N=7479
Personal-pronomen	22,98÷24,42	14,28÷15,12	0,66÷ 0,94	43,35÷45,33
Possessiv-pronomen	4,34÷ 5,06	13,19÷14,01	2,54÷ 3,06	14,71÷16,15
Reflexiv-pronomen	0,38÷ 0,62	0,6 ÷ 0,8	0,13÷ 0,27	2,09÷ 2,71
Demonstrativ-pronomen	33,99÷35,61	32,44÷33,56	49,33÷50,87	15,13÷16,59
bestimmte Pronomen	13,21÷14,39	11,02÷11,78	9,34÷10,26	5,69÷ 6,65
unbestimmte Pronomen	9,88÷10,92	7,38÷ 8,02	6,02÷ 6,78	2,75÷ 3,45
Interrogativ-pronomen	0,47÷ 0,73	1,07÷ 1,33	0,13÷ 0,27	1,62÷ 2,16
Relativ-pronomen	10,17÷11,23	15,76÷16,64	28,59÷30,01	8,9 ÷10,08
Negativ-pronomen	0,65÷ 0,95	1,35÷ 1,65	0,3 ÷ 0,5	1,09÷ 1,55

Erstens ist ein intensiver Gebrauch der Personalpronomen für die gebundene Rede (im Vergleich zu den anderen Texten) kennzeichnend, was mit dem ausgeprägt verbalen Charakter und mit dem Vorherrschen der Personalformen des Verbs sowie mit der Tendenz zur Vermeidung von Wiederholungen der Substantive verbunden ist.
Zweitens weist die gebundene Rede von allen Texten die größte Häufigkeit sowohl des Reflexivpronomens als auch der Possessiv- und Interrogativpronomen auf.
Drittens ist der größte Anteil von Demonstrativ- und Relativpronomen in den wissenschaftlich-technischen Texten festzustellen. Offensichtlich wird diese Erscheinung durch die Tendenz des wissenschaftlichen Sprachstils zur Exaktheit und Monosemie bedint.
Viertens ist der Anteil der bestimmten Pronomen und der unbestimmten Pronomen in der Publizistik höher als in den anderen Texten.

3. ZUSAMMENFASSUNG

Die durchgeführte Untersuchung, die sich auf die Einheit der quantitativen und qualitativen Analysen stützt, ermöglicht einerseits, die differentiellen und lexikalisch-morphologischen Merkmale von drei Funktionalstilen der modernen lettischen Literatursprache zu ermitteln, und andererseits, die statistischen Korrelationsverhältnisse zwischen diesen Stilen aufzudecken.
Der Vergleich der Korrelationskoeffizienten und der quantitativen Einschätzungen der lexikalischen Annäherung der Stile (s. Tabelle 20) erlaubt zu behaupten, daß in dem stilistischen System der lettischen Schriftsprache der wissenschaftlich-technische Stil und der Stil der Belletristik polare Größen bilden. Der publizistische Stil hat einen Mittelwert und tritt als ein Bindeglied zwischen diesen beiden polaren Größen auf.
Die Korrelation im Gebrauch der Wortarten sowie in der lexikalisch-statistischen Struktur der Texte (s. Tabelle 20) ist zwischen den publizistischen und den wissenschaftlich-technischen Texten stärker aks zwischen den publizistischen und den belletristischen Texten. Die Funktionalstile bilden zwei gegenübergestellte Gruppen -- die "nichtkünstlerischen Stile" (der wissenschaftlich-technische und der publizistische Sprachstil) und der künstlerische Sprachstil der Belletristik.

Hingegen ist die Ähmlichkeit des Vokabulars zwischen den publizistischen und den belletristischen Texten stärker als zwischen den publizistischen und den wissenschaftlich-technischen Texten (s. Tabelle 20). Diese Differenz und Ähnlichkeit der Stile werden durch die Stufe der Unterschiede in ihren situativ-inhaltlichen Aspekten bestimmt, die die Wahl der Lexik determinieren.

Tabelle 20. Wechselbeziehungen zwischen Funktionalstilen

	Die lexikalische Verknüpfung (Jaccardsche Koeffizienten)		
	P2	WT	B
P1	0,433÷0,477	0,256÷0,268	0,289÷0,301
P2		0,259÷0,271	0,350÷0,362
WT			0,182÷0,192

(Fortsetzung)

	Die lexikalisch-statistische Struktur der Texte (Pearsonsche Koeffizienten)		
	P2	WT	B
P1	0,979÷0,981	0,854÷0,864	0,793÷0,808
P2		-	-
WT			-

(Fortsetzung)

	Gebrauch der Wortarten (Pearsonsche Koeffizienten)		
	P2	WT	B
P1	0,986÷0,999	0,982÷0,999	0,493÷0,965
P2		0,978÷0,999	0,592÷0,974
WT			0,501÷0,966

Folglich liegt die Publizistik auf der Sprachachse der Sprache der Belletristik näher, auf der Textachse aber liegt sie dem wissenschaftlich-technischen Stil näher.

Diese Ergebnisse sind nicht nur vom Standpunkt der Stilkunde der lettischen Gegenwartssprache von Interesse, sie können auch bei der Entwicklung neuer linguistischer Konzeptionen Verwendung finden, die mit der Theorie der vagen Mengen (fuzzy sets), der linguistischen Variablen, der linguistischen Algorithmen (fuzzy algorithms) verbunden sind.

ANMERKUNGEN

1. Das Häufigkeitswörterbuch der lettischen Sprache (Latviešu valodas biežuma vārdnīca) ist von dem Kollektiv des Laboratoriums für mathematische Linguistik an dem Institut für Sprache und Literatur der Akademie der Wissenschaften der Lettischen SSR unter der Leitung von T.A. Jakubaitis verfaßt.
2. Die Feststellung von statistischen Parametern der Stile in der ukrainischen Gegenwartssprache unter Verwendung des Vergleichs der Konfidenzintervalle ist von einem Kollektiv unter der Leitung von W.I. Perebejnos ausgeführt worden (in der Sammlung Statisticni parametri stiliv. Kiiv, 1967)
3. P1 und P2 werden zusammen untersucht.
4. Der nominale Charakter der wissenschaftlich-technischen Texte ist von vielen Sprachforschern auf der Basis verschiedener Sprachen angemerkt worden. So nimmt z.B. M.RENSKY(1972:224-233), der belletristische und wissenschaftliche Texte des Englischen und des Tschechischen untersucht hat, an, daß der quantitative Zuwachs des Nomens in der wissenschaftlichen Prosa für alle indoeuropäischen charakteristisch ist.
5. Analoge Ergebnisse hat N.W.NEVEROVA (1970:96) auf der Basis des Russischen erhalten.
6. Vgl. die Häufigkeit der Adjektive in der deutschsprachigen wissenschaftlichen und schönen Literatur: in den wissenschaftlich-technischen Texten ist sie etwa um 30% höher als in der Belletristik (DZAVLALOV 1975:143).
7. Die lettische Sprache kennt 7 Kasusformen -- den Nominativ, den Genitiv, den Dativ, den Akkusativ, den Instrumental, den Lokativ, den Vokativ.
8. Da in dem Häufigkeitswörterbuch der Belletristik (B) die Statistik der morphologischen Kategorien nicht vorhanden ist, sind diese statistischen Charakteristiken auf der Basis von 6 Gedichtsammlungen zeitgenössischer lettischer Dichter (B_D) ermittelt; der Gesamtumfang beträgt 50764 Wörter.

9. Derselben Meinung ist V.A.Nikonov (1959:59), der die Häufigkeit der Kasus im Russischen untersucht hat.
10. Die Partizipien, die Bestandteile der zusammengesetzten Zeitformen sind. gehören zu den Personalformen des Verbs.
11. Der häufige Gebrauch von Nominalformen als eine der Charakteristiken des wissenschaftlichen Sprachstils wurde auch von S.I.KAUFMAN (1961:105) festgestellt.
12. Die statistische Untersuchung des Vorkommens der Pronomen ist auf der Grundlage der Klassifikation durchgeführt worden, die gleich J.ENDZELIN (1922:372-406) von der modernen Theorie und Praxis des Lettischen gebraucht wird.

LITERATUR

ALEKSEEV, P.M., Statističeskaja leksikografija. Leningrad, 1975.

CLAUß, G., EBNER, H., Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Berlin, 1967.

DOLEŽEL, L., A Framework for the Statistical Analysis of Style. In: Statistics and Style. New York, 1969, 10-22.

DŽAVLALOV,M., Sravnitel'naja količestvennaja i kačestvennaja charakteristika prilagatel'nych v nemeckoj naučnoj i chudožestvennoj literature. In: Jazyk naučnoj literatury. Moskva, 1975, 140-155.

ENDZELIN, J., Lettische Grammatik. Riga, 1922, 184-800.

GOLOVIN, B.N., O roli statistiki v opisanii jazykovych i rečevych stilej. In: Tezisy dokladov Mežvuzovskoj konferencii "Častotnye slovari i avtomatičeskaja pererabotka lingvističeskich tekstov". Minsk, 1968, 36-41.

GOLOVIN, B.N., Jazyk i statistika. Moskva, 1971.

HERDAN, G. Quantitative Linguistics. London, 1964.

KAUFMAN, S.I., Ob imennom charaktere techničeskogo stilja (na materiale amerikanskoj literatury). In: Voprosy jazykoznanija 5, 1961, 103-108.

KĻAVIŅA, S., Latviešu publicistikas valodas biežuma vārdnīca. In: P. Stučkas Latvijas Valsts universitātes Zinātniskie raksti, 86, sēj. Latviešu leksikas attīstība. Riga, 1968, 199.-216.lpp.

KOŽINA, M.N., O rečevoj sistemnosti naučnogo stilja sravnitel'no s nekotorymi drugimi. Perm', 1972.

Latviešu valodas biežuma vārdnīca. 1.-3. sej. Rīga, 1966-1972.
1. sej. Tehnika un rūpniecība. 1.d. 1966.
1. sej. Tehnika un rūpniecība. 2.d. 1968.
2. sej. Laikraksti un žurnāli. 1.d. 1969.
2. sej. Laikraksti un žurnāli. 2.d. 1969.
3. sej. Daiļliteratūra. 1.d. 1972.

MULLER, Ch., Einführung in die Sprachstatistik. Berlin, 1972.

NALIMOV, V.V., Verojatnostnaja model' jazyka. Moskva, 1974.

NEVEROVA,N.B., Količestvennye parametry stilej russkogo jazyka na materiale predlogov. In: Strukturno-matematičeskie metody modelirovanija jazyka. (Tezisy dokladov i soobščenij Vsesojuznoj naučnoj konferencii). C.P. Kiev, 1970, 95-96.

NIKONOV,V.A., Statistika padežej russkogo jazyka. In: Mašinnyj perevod i prikladnaja lingvistika 3/10, Moskva 1959, 45-65.

PIOTROVSKIJ,R.G., BEKTAEV,K.B., PIOTROVSKAJA, A.A., Matematičeskaja lingvistika. Moskva, 1977.

RENSKY,M., The Frenquency of Word Classes as a Function of Style and Linguistic Structure. In: The Prague School of Linguistics and Language Teaching. London, 1972, 224-233.

RIZEL',E.G., K voprosu ob ierarchii stilističeskich sistem i osnovnych tekstologičeskich edinic. In: Inostrannye jazyki v škole 6, 1975, 8-14.

SOIDA,E., KĻAVIŅA,S., Latviešu valodas onversā vārdnīca. Rīga, 1970.

SOKAL, R.R., SNEATH, P.H.A., Principles of numerical taxonomy. San Francisci, Freeman 1963.

Statistični parametry stiliv. Kiiv, 1967.

TULDAVA, Ju.A., Ob izmerenii leksiceskoj svjazi tekstov na urovne slovarja. In: Voprosy statističeskoj stilistiki. Kiev, 1974, 35-42.

ZADEH, L.A., The Concept of a Linguistic Variable and its Application to Approximate Reasoning. New York, 1973.

Regularity and Homogeneity of Morphological and Word-Forming Patterns

Michail V. Arapov, Moscow

1. INTRODUCTION

A great amount of information about the meaning and use of words in any language can be represented in form of sets of pairs of words[1], e.g.

a_1 ⟨boy, boys⟩, ⟨man, men⟩, ⟨bus, buses⟩, ⟨star, stars⟩, ...
a_2 ⟨father, father's⟩, ⟨lady, lady's⟩, ⟨dog, dog's⟩, ...
a_3 ⟨work, worked⟩, ⟨go, went⟩, ⟨come, came⟩, ⟨add, added⟩, ...
b_1 ⟨swift, swiftly⟩, ⟨soft, softly⟩, ⟨loud, loudly⟩, ...
b_2 ⟨cold, coldness⟩, ⟨sweet, sweetness⟩, ⟨ill, illness⟩, ...
c_1 ⟨correct, wrong⟩, ⟨handsome, ugly⟩, ⟨high, low⟩, ...
c_2 ⟨hate, enmity⟩, ⟨love, adoration⟩, ⟨attend, nurse⟩, ⟨pull, drag⟩, ...
c_3 ⟨feeling, keen⟩, ⟨heat, fierce⟩, ⟨enemy, sworn⟩, ⟨blow, hard⟩, ...

From the above examples (a - c) we can see that the elements of such pairs have some formal and semantic relationships. The nature of these relationships is studied in different branches of linguistics: grammar deals with the semantic aspects of the relationships for the pairs a (a_1 - a_3), semasiology does so for the relationships b_1 - c_3; the formal aspects of these relationships are covered by morphology (a_1 - a_3), the theory of word-formation (b_1 - b_2), and lexicology (c_1 - c_3).

In this paper our main interest lies not in the concrete relationships between words which constitute those pairs, but in more abstract properties of the sets of such pairs, i.e. their regularity and homogeneity.

2. PATTERNS

In order to be able to treat these concepts (regularity and homogeneity) more precisely, we must introduce a pattern. The pattern is regarded as present if two sets are defined: 1) a set of pairs $M_{xy} = \{\langle x, y\rangle\}$ like those of a - c, and 2) a set X of elements x from which the first components of the pairs $\langle x, y\rangle$ are chosen. M_{xy} is termed a graph of the pattern.

That is, the pattern is a pair (X, M_{xy}). It is, however, convenient in this paper to use the same notation both for the pattern and for its graph, if it does not lead to misunderstanding.

We now need to introduce the concept of the co-pattern. The co-pattern is a set of pairs $\langle y, x\rangle$ which is written in inverse order to the pairs of M_{xy}, and set Y whose elements y are those from which the first components of the pairs $\langle y, x\rangle$ (i.e. the second components of the pairs (x, y)) are chosen. The co-pattern to M_{xy} is denoted M'_{xy} or M_{yx}.

Hence, in order to define the pattern and co-pattern we need to describe

(1) the sets X and Y; correspondingly the first and second components of each set are chosen; e.g. the first components of the pattern a_1 are chosen from the array of the English nouns in the singular, and then the second from the nouns in the plural; as to the pattern b_1, its first components are adjectives and the second ones nouns which contain a suffix -ly etc.

(2) the semantic and formal relationships between the components of the pairs M_{xy} ; e.g. the second elements of the pattern a_1 add the idea of plurality to the first ones, an idea which can be expressed by means of certain inflections, alternations and accent; the second components of the pattern c_3 express an idea of the utmost degree of quality (or action) denoted by the first components of the pair, and do so by pure lexical means [the second component is a kind of intensifier, so called "lexical parameter Magn" of the first (see MEL'ČUK 1974: 89)].

In order to avoid mathematical complications, it is necessary to impose some additional restrictions (3 - 5) on the concept of the pattern.

(3) The classes X and Y must be productive ones (about the concept of productivity see ARAPOV 1974; ARAPOV/CHERC 1974; the main properties of the productive classes will be listed below, see Section 4).

(4) Each element of the two sets X and Y must have a definite frequency of usage[2].

(5) If $x \in X$ and the pairs $\langle x, y_1\rangle$ and $\langle x, y_2\rangle$ belong to M_{xy}, then $y_1 = y_2$; the same holds for $y \in Y$.

In other words, M_{xy} is a partial mapping of the set X onto the set Y. It is defined not everywhere on X, but only on one of its subsets X_M, which is then called the domain of definition of M_{xy}; within the limits of its domain of definition, M_{xy} is a one-to-one mapping (of the set X_M onto the set Y).

Having imposed the restrictions (3) - (5) (especially (5)!), we must take in account that from now on not all of the examples cited (a_1 - c_3) can be treated as patterns. E.g. c_3 is not a pattern, since "blow" may be combined not only with "hard", but with "great", "violent" etc.; on the other hand, "hard" may be put together with "problem", "work", "winter", "frost" etc.

But difficulties may occur with morphological patterns as well. E.g. some Russian nouns have two forms of the genitive singular which are not entirely equivalent. Thus, beside a pair ⟨čaj, 'tea', čaja 'of tea'⟩ we have ⟨čaj, čaju⟩ etc., that is the so-called second genitive.

There are, however, a lot of interesting morphological and word-forming patterns for which the conditions (3) - (5) are perfectly fullfilled. In general those patterns allow the use of the method proposed below, and what is more, they are quite fit for its verification, because they are relatively well studied.

3. REGULARITY

In order to understand more clearly the meaning of regularity, let us make a small mental experiment. Let us assume the following procedures might be carried out:

(i) It is possible to arrange all the elements of set X, one after another.

(ii) The same holds for all the pairs $\langle x, y \rangle \in M_{xy}$.

(iii) Analogously, it is possible to list the first components of all the pairs of graphs M_{xy}, i.e. all x such that $x \in X_M$ (X_M is, as noted above, the domain of definition of M_{xy}).

(1) { Let us now call the pattern M_{xy} regular if $X = X_M$, and irregular if not.

To put it another way, the pattern is regular if each element x from X has a counterpart y from Y which forms a pair of M_{xy} with it. The regularity of the co-pattern can be defined in the same way.

The criterion (1) of regularity can be slightly weakened, when we introduce the concept of a nearly regular pattern. Let us call a pattern (X, M_{xy}) nearly regular, if the difference $X \setminus X_M$ is a non-productive class.

The regularity or irregularity of a pattern obviously depends on the way we have defined the class X. The definition of the pattern implies only $X \supset X_M$, then any class for which this inclusion holds is suitable.

If X is the class of all English nouns, then the pattern a_2 is irregular, if X_M consists only of animal nouns. On the contrary, the intuitively irregular pattern c_1 would be regular, if the class X had been previously defined as the set of only those adjectives which have antonyms.

Such a trivial solution to the problem of regularity of particular patterns, however, cannot always be found. In case of pattern b_1, the class X consists of all English adjectives.

It is unlikely that this pattern would be a regular one: there are many adjectives which lack matching adverbs with a suffix -ly. But is the pattern b_1 nearly productive? The answer is far from being obvious: our intuition tells us nothing about the productivity of the class of the adjectives which have no matching adverbs with the suffix -ly.

We can, of course, rewrite the definition of the class X. Let X consist of qualitative adjectives. Yet there is not much sense in re-defining the class X in such a way; this leads to a sort of a vicious circle, because the very notion of an adjective as the qualifying one is based in turn on the existence of the matching adverb.

Let us now turn to the question whether our mental experiment is feasible. If the experiment could really be performed on all patterns (nor just on trivial ones, viz., with empty or finite graphs M_{xy}) then the constructions proposed here would be unnecessary. At best, they would lead to the disclosure of some additional interesting facts about the patterns, and nothing more.

It is very likely, however, that our mental experiment could work if performed on any interesting pattern. No linguist (and particularly no native speaker) can keep in mind infinitive sets X, Y, and M_{xy}. He usually applies criterion (1) to their finite subsets only, and then is truing to extrapolate the results obtained on a larger fragment of vocabulary or on the vocabulary as a whole.

Even ZALIZNJAK's exhaustive Russian Grammar Dictionary (1977) does not cover all the morphological phenomena in question. We believe, nevertheless, that by using this excellent compendium we can probably get a very good approximation to the reality of the language, yet it is impossible for us to express the degree of our certainty by means of a number.

Various data obtained by applying criterion (1) to a finite dictionary will lead to different conclusions about the regularity of a given pattern. If criterion (1) holds for this dictionary, the regularity of the patterns in question would still not be proved, because some $x \in X$ that have no partner

may by pure chance have not been included in the dictionary. But if (1) holds, the degree of our certainity as to the regularity of the pattern would increase.

On the other hand, if (1) does not hold, this would prove nothing. The point is that some x may be without partners, because the proper y might fall beyond the scope of the dictionary, although such a y does exist in the language. But in this case our certainity as to the regularity of the pattern would generally decrease.

The problem confronting us now is to understand, which properties of a pattern make our certainity increase (or decrease) and to what extent. That is, we intend to elucidate and formalize a mechanism of forecasting, hoping that it may resemble that of a speaker in a similar situation.

The mechanism in question enables us not only to estimate the probability of the pattern's regularity for the entire vocabulary provided that the pattern is regular over some finite dictionary (the base of forecast), but also to decide whether the pattern is nearly regular or not. If the pattern is an irregular one, we could measure the degree of such irregularity by means of the mechanism.

In order to obtain the forecast we need first of all to know the place of the base V in the lexical system of a language. As it turns out, the only necessary data are the ranks r(x) of all the words of the base (hence the restriction (4) on the pattern). The rank of x is the number of the words in a language which are used more frequently than x or as frequently as x.

It is convenient (but not necessary) that the base V should consist of N words. which are most frequently used. Having a base which consists of N of the most frequent words (i.e. the words with the lowest ranks) - call such a base V_N - we can easily move from one base to another simply by changing N.

Let us consider the fragment of the pattern M_{xy} which is defined on the base dictionary V_N. The fragment consists of all x, such that $x \in X$ and $r(x) \leq N$ (we call them X_N) and all the pairs $\langle x, y \rangle$, such that $(x) \leq N$, $(y) \leq N$ and $\langle x, y \rangle \in M_{xy}$.

We are interested in such properties of this fragment of

the pattern as can be formulated exclusively in terms of word ranks. Then, the pattern M_{xy} having been defined, we may forget about the words themselves. Now we make the following stipulation; take for granted that the symbols x, y, $\langle x, y \rangle$ are always interpreted as the symbols of ranks of the corresponding words, i.e. as numbers r(x), r(y), $\langle r(x), r(y) \rangle$, if the opposite is not clear from the context.

Now we could go on formulating those properties we need for a pattern. They will be postulated in Section 5; then in Section 6 we shall demonstrate that these postulates do not contradict our empirical data.

In order to complete this project, we need some information from the theory of class productivity.

4. PRODUCTIVITY

The theory of productivity is based on the assumption that the behaviour of a class X as a whole can be predicted by studying the distribution of its elements x among N of the most frequent words listed in a given frequency dictionary V_N. (The N most frequent elements of X we call X_N, i.e. $X_N = X \cap V_N$.)

Here we need only a few ideas and definitions of that theory (see ARAPOV 1974; ARAPOV/CHERC 1974).

Among the underlying concepts of this theory is the concept of the age $\tau(x)$ of a word x. The age of a word is the time interval from the word's origin till the moment of the compilation of the dictionary (the latter may be chosen as the beginning of a co-ordinate system, $t_{comp} = 0$).

Let X be some class of words x (e.g. all the words belonging to a certain grammatical category, containing a particular suffix or having a semantic property, etc.) such that the inequality $\tau(x) > \varepsilon$, $\varepsilon \neq 0$ holds for all the elements x of this class.

By definition, such a class X has not been replenished by new words since the moment ε, therefore it is natural to call X non-productive. If ε does not exist, the class X is produc-

tive.

According to the theory of productivity, we have the following three assertions:

(i) If X_i is a non-productive class, then the probability $p_i(x)$ of occupying the rank x by a word from X_i decreases, i.e. for a high-ranking x the probability of belonging to the non-productive class X_i is low.

(ii) Each non-productive class is finite, and each productive one is infinite.

(iii) If X_j is a productive class, then the probability $p_j(x)$ of occupying the rank x with a word from X_j increases, tending to some λ_j, $0 < \lambda \leq 1$. To put it differently, for all the words - exept perhaps the most frequent ones - the probability of belonging to a productive class X_j is equal to λ_j, therefore λ_j can be treated as a measure of productivity for the class X_j[3].

The problem of productivity measurement has been thoroughly studied in our previous work (ARAPOV 1974). Here we only notice that it is impossible, of course, to find the exact limit of an empirical series $p_j(x)$, if we have only a finite number of its terms. Yet knowledge of the exact value is not necessary in practice. It is enough to get an estimate of λ_j. This can be achieved in several ways.

If a rough estimate is sufficient, then it is enough to find the power of the set $X_j \cap V_N$ (we call it $|X_j \cap V_N|$), i.e. the number of elements of X_j which are among N the most frequent words, then

$$\lambda \approx \frac{1}{N}|X_j \cap V_N| .$$

Sometimes we do not know the power of $X_j \cap V_N$ but only the sum of frequencies F_N^j of words belonging to the set $X_j \cap V_N$ Then

$$\lambda_j \approx \frac{F_N^j}{F_N} ,$$

where F_N is the sum of frequencies of all the words from V_N. The two methods of evaluation of λ_j tend to be more accurate for a greater N.

Even a more accurate estimate is obtainable for the ratio of productivities:

$$\frac{\lambda_x}{\lambda_y} \approx \frac{|X_N|}{|Y_N|} , \text{ or } \frac{\lambda_x}{\lambda_y} \approx \frac{F_N^x}{F_N^y} .$$

The above mentioned properties of productive classes justify the restriction (3) on the concept of pattern. Indeed, if both the classes X and Y are non-productive, it is enough to apply criterion (1) to establish, whether a given pattern is regular; hence the problem of forecasting does not arise.

If only one of the classes -X or -Y is non-productive, then the problem of regularity of the pattern M_{xy} is a trivial one. M_{xy} is irregular because a one-to-one mapping from a finite set onto an infinite one or an inverse mapping does not exist.

It remains the same, when both X and Y are productive classes, with the productivities λ_x and λ_y respectively.

5. STATISTICS OF REGULARITY

Let us now consider the following statistics for M_{xy},

$$(2) \qquad m_{xy} = \ln\frac{x}{y} .$$

It is only natural to ask why just the logarithm of the ratio x and y (where x and y are the ranks of appropriate words) is chosen to be characteristic of the pair $\langle x, y\rangle$[4].

Our answer includes two parts of which the second is more important.

Firstly, there are some formal considerations necessary. The statistics ln x/y can be computed even if the ranks themselves are unknown. To this end we need to know only the (absolute) frequencies f_x and f_y of the words x and y, respective-

ly. Indeed, according to the Zipf's law we have

$$f_x = \frac{C}{[r(x)]^\gamma}$$

then

$$\ln \frac{x}{y} = \ln \frac{(f_y)^\alpha}{(f_x)^\alpha} = \alpha \ln \frac{f_y}{f_x}, \quad \alpha = \frac{1}{\gamma}.$$

Taking into account that α is near to 1, we can simply set m_{xy} equal to $\ln \frac{f_y}{f_x}$ instead of $\ln \frac{x}{y}$. For the co pattern there is $m_{yx} = -m_{xy}$.

Secondly, E.L. THORNDIKE (1943) and N.A. ŠECHTMAN (1978) have demonstrated that at least for some patterns the values of (x/y) tend to group around particular points which are specific for every pattern. E.L. THORNDIKE has studied the English morphological and word-forming patterns, whereas N.A. ŠECHTMAN has done so for Russian synonyms, antonyms and pairs of words connected by generic relations etc.

An obvious disadvantage of the statistics x/y is the strong non-symmetry of its distribution. For example, if observed values of these statistics are grouped around 1, then about half of them would lie within the intervall [0,1], and the other half within [1,∞]. Under this condition it is only natural to apply a logarithmic transformation to the random variable x/y.

Let us postulate two most important properties of a pattern.

(3) If the pattern M_{xy} is a homogeneous one, then the random variable (statistics) m_{xy} is distributed normally with mean $\bar{m}_{xy}$ and variance δ^2_{xy},

$$p(u) = \frac{1}{\sqrt{2\pi}\delta_{xy}} \exp\left[-\frac{1}{2}\left(\frac{\bar{m}_{xy} - u}{\delta_{xy}}\right)\right]$$

and

(4) each pattern can be represented as a composition of a finite number of homogeneous patterns.

Let us define the composition of two patterns (X_1, M_{x_1y}) and (X_2, M_{x_2y}) as a pattern $(X_1 \cup X_2, M_{x_1y} \cup M_{x_2y})$.

The composition of n patterns can be defined in the same way.

By definition the representation of a patterns as a composition is not always unique. Until now we have been dealing with patterns which can be split easily into homogeneous parts: $X_1 \cap X_2 = \emptyset$, $M_{x_1y} \cap M_{x_2y} = 0$.

The validity of the postulates (3) and (4) has been sufficiently supported by the evidence of a number of frequency lists of several languages. The data obtained and the results of their analysis will be presented in Section 6. In this passage we shall see how it is possible to use postulate (3) in the case of a homogeneous pattern.

The importance of this statement lies in the possibility of constructing a kind of neighbourhood (a subset of ranks in the dictionary V_N) $w_p(x)$ for every x in which the corresponding y lies with the given probability p.

It is important that the range of the neighbourhood is finite and does not depend on the rank x. Let us define the neighbourhood $w_p(x)$ for $x \in X$ as a set of ranks y for which the following inequality holds:

$$\left|\frac{\ln \frac{x}{y} - \bar{m}_{xy}}{\delta_{xy}}\right| \leq z$$

where z is an arbitrary number.

Hence, as it follows from (3).

$$P\{y \in W_p(x)\} = \varphi(z)$$

and

$$P\{y \bar{\in} W_p(x)\} = 1 - \varphi(z),$$

where $\varphi(z)$ is the cumulative standardized normal distribution with mean zero and variance one,

$$\varphi(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{I}^{-t^2}\, dt \ .$$

When the neighbourhood $w_p(x)$ is given, we are not obliged to scan the whole body of vocabulary to find the counterpart for x. (It was just that obligation that made the application of criterion (1) irreal.) We can look for y only in a finite fragment of the vocabulary.

Let us now formulate a new criterion of regularity more precisely.

(5) If $y \in w_p(x)$ (and therefore $\langle x, y\rangle \in M_{xy}$, $x \in X_M$) has a frequency which corresponds to probability p, $p = \varphi(z)$, then the homogeneous pattern is regular, and it is irregular if not.

Since the probability of $y \in w_p(x)$ is independent of the rank x, it is possible in principle to gather the sample that enables us to make a decision as to whether (5) holds or not, and this sample may consist of elements which lie within the list V_N of N of the most frequent words (the basis of our forecast). The rehability of the forecast obtained by such means may be established by routine methods of statistics (provided the verification of (3) has been done previously).

The implementation of criterion (5) does not make any difficulties if the necessary statistics can be chosen from such elements of the vocabulary V_N whose neighbourhood lies within V_N, $w_p(x) \subset V_N$. Unfortunately, the volume of some frequency lists is often too small; it is of the same order as the volume of the neighbourhoods of x. It is clear that x may have no partner y in such lists, simply because there is no place within V_N to lodge the neighbourhoods of x.

Indeed, let the pattern M_{xy} be a regular one with $\lambda_x > \lambda_y$ Then there are about $\lambda_x N$ elements of the set X and $\lambda_y N$ elements of Y within the dictionary V_N. Since $\lambda_x N > \lambda_y N$ it is impossible to establish a one-to-one mapping of the set X_N onto Y_N, therefore some x may have no partner.

A linguist would be especially interested in the case where both, the patterns M_{xy} and the co-pattern M'_{xy}, are regular.

The problem of the co-pattern regularity can obviously be solved by the same method as that of the pattern.

Let us give a brief summary of the results concerning one particular case, i.e. when the pattern M_{xy} is irregular.

(a) If the elements of the class $X_N \setminus X_M$ are distributed within V_N in such a way that the hypothesis of the productivity of this class can be admitted, then M_{xy} is a nearly regular pattern.

(b) If the elements of the class $X_N \cap X_M$ are distributed within V_N in such a way that the hypothesis of non-productivity of X_M is admissable, then M_{xy} is an absolutely irregular pattern.

(c) If the elements of both the class $X_N \setminus X_M$ and $X_N \cap X_M$ are distributed within V_N in such a way that the hypothesis of productivity of the two classes can be admitted, then it is a μ-regular pattern.

If the frequency list V_N is small, criterion (5) should be altered. Instead of the system of neighbourhoods $x_p(x)$ which do not coincide with each other for different x, we may consider just one neighbourhood for all $x \in X_N$, and this neighbourhood coincides with V_N. But different $x \in X_N$ have different probabilities p_x, so that the appropriate y belongs to this neighbourhood.

The following formula (3) yields this probability:

$$P_x\{y \le N\} = F(Q) = \int_{-\infty}^{Q_N} \exp\left[-\frac{1}{2}\left(\frac{\ln\frac{x}{y} - \bar{m}_{xy}}{\delta_{xy}}\right)\right] dy \ ,$$

where

$$Q_N = \frac{\ln\frac{x}{N} - \bar{m}_{xy}}{\delta_{xy}} \ .$$

Then the expectancy of the number of $x \in X_N$ which have no matching partners y within V_N, though the pattern M_{xy} is regular within the entire vocabulary, is given by the formula

(6) $$E(|X_N \setminus X_M|) = \sum_{x \in X_N} F(Q_N) \ .$$

Now the criterion of regularity can be reformulated in the following way: a pattern M_{xy} is regular, if the difference between the power of the set $X_N \setminus X_M$ and the value (6) is statistically insignificant.

Let us assume the productivity of X to be λ_1 and that of X_M to be λ_2, and take as a measure of regularity for the pattern M_{xy} their ratio $\mu = \frac{\lambda_1}{\lambda_2}$. (An interpretation for μ is given below, see p. 25).

6. EXAMPLES

The hypothesis (3) and (4) have been substantiated by the study of several frequency lists compiled on the basis of Russian, Polish, and English texts. We have singled out four quite typical examples from the bulk of gathered evidence. The first two of them concern patterns which can be regarded as homogeneous and regular. Example 3 is one of the regular but not homogeneous patterns, and example 4 is one of the homogeneous but irregular patterns.

The scheme of pattern analysis is as follows:

a) Two productive classes, X and Y, with the proper type of relationship between them have to be selected.
b) All the pairs of lexical units $M = \langle x, y \rangle$, $x \in X$, $y \in Y$ (the graph of the pattern) whose relationship is of a given type have to be picked out from the frequency list V_N; for a further study we take all of them or those for which some additional conditions are fulfilled. These conditions may concern the frequency of usage of the words picked out.
c) The fulfillment of the restriction (5) on the concept of pattern has to be checked up.
d) At the same time we single out each $x \in X$, whose partner y is absent from the list V_N, i.e. we form the class $X_N \setminus X_M$.
e) For each pair $\langle x, y \rangle$ picked out of the list the value of $\ln x/y$ has to be computed. In the examples cited below, we compute the equivalent statistics $m_{xy} = \ln \frac{f_y}{f_x}$ instead of $\ln \frac{x}{y}$.

f) Afterwards we check the independance of the parameters $\bar{m}_{xy}$, s^2_{xy} of the distribution m_{xy} from the rank x; in order to verify this hypothesis, we divide the sample M into two or three parts M_1, M_2, ... in such a way that $fx_1 > fx_2$, $x_1 \in X_{M_1}$, $x_2 \in X_{M_2}$, etc., then compare empirical moments $\bar{m}_1$, $\bar{m}_2$, ..., s_1^2, s_2^2, ... Since the zero hypothesis: $\bar{m}_1 = \bar{m}_2 = \ldots$ $s_1^2 = s_2^2 = \ldots$ has never been rejected (at the level of significance $2\alpha = 0.1$, at least), we will not dwell on that particular stage of the empirical routine.
g) Finally, we have to test the hypothesis of normality for the distribution of m_{xy} with the aid of the chi-square test.

Further treatment of M depends on the outcome of the normality test and the character of distribution (within the dictionary V_N) of such elements of the class X that have no partner from Y. This stage of analysis will be illustrated by the examples cited below.

The source of data for the first two examples is the frequency dictionary of LEWICKI et al. (1975) which was compiled on the basis of Polish newspapers.

From the list of LEWICKI's dictionary (which is arranged in the order of frequency) all the adjectives with a frequency of usage ≥ 12 have been picked out. We disregard only the ordinal numbers and such "quantifiers" as każdy 'each, every', dany 'a given one', poszczególny 'special, seperate' niektóry 'some' etc., which the authors of the dictionary have treated as adjectives.

Example 1

The Polish Frequency Dictionary gives not only the frequency of each word, but also that of all grammatical forms occuring in the sample. We divide all the word-forms into two classes:

1) all the forms of the singular (independently of their case, gender, degree of comparison),
2) all the forms of the plural (henceforth Sg and Pl respective-

ly). The frequencies f_x and f_y are defined as the sums of all the frequencies of the Sg forms (x) and the Pl forms (y) of the same adjective.

The total number of the pairs (x, y) that have been found in the list is 230. Only one of the adjectives, dziwny 'strange', has no Pl forms. For the other 229 pairs the statistics $\ln f_y/f_x$ has been computed, its distribution is given in Table 1 (see also Fig. 1).

Having been applied to the distribution of m_{xy}, the chi-square test of fitness indicates a fairly good agreement of the empirical data with the hypothesis of normality of the distribution (see Table 1). We can see that the Sg forms are used 1.83 times more frequently than the Pl forms of the same adjective; in other terms, the mean rank of Sg forms is about 0.55 that of Pl forms.

Making use of the normality of the distribution, we can easily construct neighbourhoods $w_p(x)$ for every x, i.e. find the limits of intervals that contain y for every x with the probability p. E.g. with the probability 0.5 the ratio of the frequency of the Pl forms (f_y) to the frequency of the Sg forms (f_x) is between 0.11 : 1 to 2.78, and with the probability 0.99 between 0.04 : 1 to 6.97 : 1.

To put the latter differently, there is no more than one Polish adjective out of a hundred whose Sg forms are used 23 times more or 7 times less frequently than its Pl forms.

We are dealing with the dictionary which is based on the sample where the Sg forms of dziwny have been used 14 times, while the Pl forms of this word have not even once been used (that is, $f_y < 1$). Having got the parameters of the distribution of m_{xy}, we can estimate the probability of the following event: the Sg forms are used 14 or more times more often than the Pl forms of the same x-word. This probability is equal to 0.02, that is not very low. Hence there is a chance that the absence of the Pl forms is due to random factors, and that the considered pattern is regular. Of course the Pl forms of dziwny are presented in other Polish texts.

Table 1: The Sample Distribution of the Statistic m_{xy} for the Examples 1 - 4.

Scores z*)	Examples				
	1	2	3 inanimated	3 animated	4
3.0	1	-	1	1	-
2.5	1	-	1	2	1
2.0	3	1	8		4
1.5	6	9	14	7	23
1.0	23	23	26		31
0.5	32	41	51	29	39
0.0	49	38	84		43
-0.5	49	43	71	27	24
-1.0	31	27	45		22
-1.5	18	17	30	13	10
-2.0	8	7	10		5
-2.5	7	5	4	1	-
-3.0	-	4	4	-	-
-3.5	1	1	2	-	-
	229	216	351	80	205
m_{xy}	-0.606	0.815	0.159	-2.144	-0.867
$e^{\bar{m}_{xy}}$	0.546	2.259	1.172	0.117	0.420
s_{xy}	0.995	1.017	0.939	0.675	2.166
χ^2	5.81	13.99	8.96	3.01	6.30
degrees of freedom (df)	7	7	7	3	7

90-procentile of χ^2 for 3 df is 6.25, and for 7 df 12.02.
95-procentile of χ^2 for 7 df is 14.07.

*) $z = \dfrac{\ln \frac{f_y}{f_x} - \bar{m}_{xy}}{s_{xy}}$

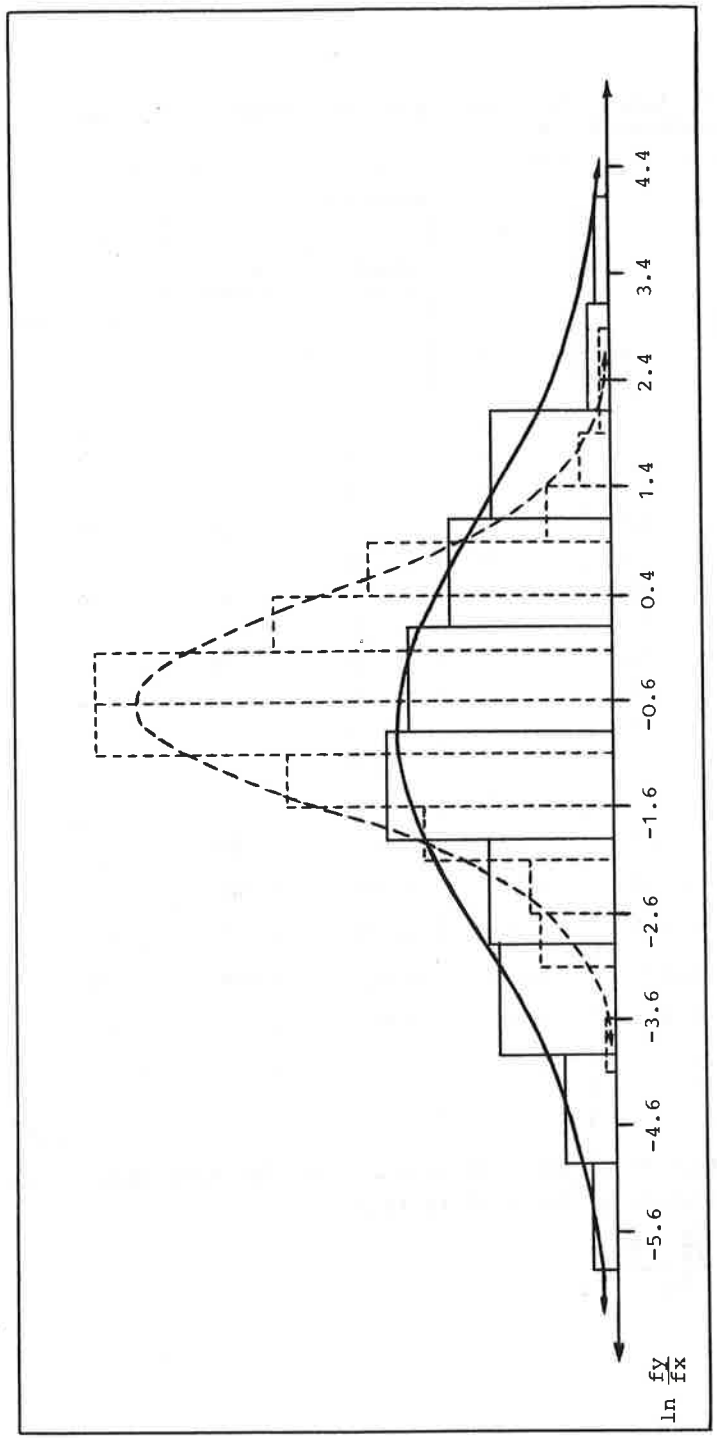


Fig. 1: Distribution of the statistic m_{xy} for the examples 1 and 4

---- : Example 1

—— : Example 4

Example 2

Let us consider a set of pairs ⟨x, y⟩ where x is an adjective in the nominative case (gender, number and degree of comparison are not important), and y is the same adjective in the genitive (in an arbitrary gender etc.); we will call these forms Nom and Gen respectively, whereas f_x and f_y will be the sum of the frequency of Nom and Gen forms.

With the aid of the same dictionary as in Example 1, we have constructed 216 pairs of adjective forms; the first components of such pairs are Nom forms, the second Gen forms. Three adjectives: konieczny (24) 'necessary', nastapujący (6) 'following' and olbrzymny (4) 'enormous' (the frequency of Nom forms are given in the brackets) have no Gen forms at all.

As in Example 1, we compute the statistics $m_{xy} = \ln \frac{f_y}{f_x}$ for all the constructed pairs (the distribution of the statistics is listed in Table 1).

The chi-square test demonstrates that there is no reason to reject the hypothesis of normality for this distribution. As we may conclude from the data presented in Table 1, the Gen forms of Polish adjectives are used on an average of 2.26 times more frequently than the Nom forms. The variance of this mean is about the same as in the case of Sg and Pl forms (with a probability of 0.5 the ratio of the frequency of the Gen forms to that of the Nom forms is between 0.88 : 1 to 4.48 : 1).

If instead of all the Nom and Gen forms we consider the forms of the nominative singular and the genitive singular seperately, and do the same for the nominative plural and genitive plural, we would obtain roughly the same means ($f_y/f_x \approx 2$), but the variance would be slightly lower.

It is interesting to compare the parameters of analogous adjectives and noun patterns. The Gen forms (of both numbers) of a Polish noun are used about 3.4 times more frequently as the Nom forms of the same noun, but the variance is substantially larger than that of the adjective pattern. As to the Russian Gen and Nom forms, their frequencies are about the same. Here our source is ŠTEJNFEL'DT's dictionary (1963) which is

compiled mainly on the basis of samples taken from children's books and magazines.

Example 3

The source of data for this example is the above mentioned dictionary of ŠTEJNFEL'DT (1963). We consider the usage of the Russian nouns in the forms of the nominative (x) and accusative (y) cases (of both numbers) (henceforth Nom and Acc respectively). The frequency of the nouns picked out from the dictionary has to be ≥ 20.

As it turns out, the statistics m_{xy} is not normally distributed in this case, i.e. the pattern cannot be regarded as homogeneous. A further analysis has shown that the Nom forms of animate nouns are used on an average 3.5 times more frequently than the Acc forms of these nouns, whereas for inanimate nouns both forms are used with a roughly equal frequency (Acc forms being used slightly more often than Nom).

Next we consider two patterns, i.e. the pattern M_1, which consists of 351 pairs of inanimate nouns, and M_2, 80 pairs of animate nouns. As it turns out, all the nouns having Nom forms have Acc forms too.

The distribution of the statistics m_{xy} for the two patterns is given in Table 1. The hypothesis of normality holds in both cases. The following curious fact is well worth noticing. In order to increase the chances of acceptance of the normality hypothesis for the two distributions, звонок '(school)bell' should be put into the category of animate nouns. It is probably due to some personification of the bell in school-orientated texts on which the dictionary sample is based.

Example 4

Let us now consider an irregular but apparently homogeneous pattern: the Russian adjectives with a suffix -н(ый)/-н'(ий) (the class X) and the matching adverbs with a suffix -н(о)/-н'(е) (the class Y), e.g. крайний 'extreme', крайне 'extremly, сильный 'strong', сильно 'strongly', огромный 'enormous', огромно 'enormously' etc.

Using ZASORINA's dictionary (1977) as a source, we can measure the productivity of the two classes. It is about 0.08 for the adjectives and 0.03 for the matching adverbs.

Only a few lexicographers regard these adverbs as regular derivates that do not deserve to be included in a dictionary (see above the example of the Russian adverb научно 'scientifically'). They may be wrong, because the real existence and use of some of these adverbs is questionable, though they can easily be derived from appropriate adjectives not only by a native speaker, but also by a student of the Russian language. Thus, adverbs like средне 'middling', громадно 'enormously', учено 'learnedly', различно 'differently', национально 'nationally', революционно 'revolutionally', буржуазно 'bourgeoisically' etc. are not probably in use in that registers of Russian on which ZASORINA's dictionary is based[5].

We have picked out all the adjectives and adverbs of the kind we are interested in from the first 30 hundreds of the most frequent words listed in ZASORINA's dictionary. The adjectives and the adverbs to match the adjectives found were looked for through the entire dictionary, without any limitation in their frequencies. Since the sample used to compile this dictionary is large, the matching words may differ in frequency from their partners by a factor as great as 1000.

205 pairs x, y have been constructed where x is an adjective and y the corresponding adverb, the statistics m_{xy} was computed for each for them. The distribution of m_{xy}, that can satisfactorily be approximated by a normal curve, is listed in Table 1 (see Fig. 1, too).

From the data listed in Table 1 it is clear that the adverb is used on the average 2.4 times less frequently than the matching adjective, but the variance of the statistics is rather large.

E.g. if we look for an interval in which y - corresponding to a given x - is lying with the probability 0.997 (the range of this intervall is $m_{xy} \pm 3\delta_{xy}$), the interval would comprise the frequencies of y that differ from those of x by a factor

which varies from 0.0006 to 280.

There are 248 adjectives with the ending -ный/-ний among the 350 hundreds of the most frequent words; 82 of them have no matching adverbs in the dictionary.

All in all, the dictionary contains 40.000 entries (N = = 40.000), therefore the expected number of the adverbs with the ending -но/-не would be about 1.200 ($\lambda_y \approx 0.03$, $\lambda_y N \approx 1.200$). The fact that the 82 relatively widely used adjectives have no matching adverbs among the 1.200 most frequent ones may suggest irregularity of the pattern in question. We shall try to varify this hypothesis.

In Table 2, besides the number n_r of the adjectives with the suffix -н(ый)/-н'(ий), which are among the successive hundreds of words, the number $\bar{n}_r$ of the adjectives without matching adverb is listed (obviously, the relative frequency of absent adverbs must be less than 10^{-6}). The fact that $\bar{n}_r$ does not virtually depend on the rank of the adjectives increases our doubts as to the regularity of the pattern. More precise information is given in the last column of Table 2.

Here we put the expected number of the adjectives without matching adverbs among the words whose ranks are within a given rank interval. According to formula (6), only about 12 words within the first 30 hundreds might be left without partners by pure chance, due to the finiteness of the dictionary.

The gap between the expected (12) and the real (82) number is too wide to be explained by statistical fluctuation. Therefore the hypothesis of irregularity seems to be the most plausible one.

The adjectives without partners are distributed among the 30 hundreds of words in such a way that the productivity hypothesis for this class can be admitted. Its productivity is of the order 1/3 of that for all the adjectives with the suffix -н(ый). Therefore the measure μ of regularity may be set equal to 2/3. This measure can be interpreted as follows.

Let us pick out at random an adjective with the proper suffix; the probability that there will exist a matching adverb with the suffix -но is equal to 2/3.

Table 2: The Distribution of the Russian Adjectives with the Suffix -н(ый)/-н'(ий) in ZASORINA's Russian Frequency Dictionary.

Rank	Frequency (not less than)	Number of adjectives n_r	Number of adjectives without matching adverbs $\bar{n}_r$	Expected number of adjectives without matching adverbs
1- 100	1093	-	-	0.0
101- 200	553	1	1	0.0
201- 300	386	7	2	0.1
301- 400	310	7	2	0.1
401- 500	254	5	3	0.0
501- 600	216	9	3	0.2
601- 700	186	5	1	0.1
701- 800	164	8	2	0.2
801- 900	146	10	5	0.3
901-1000	134	13	2	0.4
1001-1100	121	10	3	0.4
1101-1200	113	5	1	0.2
1201-1300	105	5	2	0.2
1301-1400	97	7	2	0.3
1401-1500	90	12	3	0.6
1501-1600	84	7	5	0.4
1601-1700	80	7	4	0.4
1701-1800	75	10	4	0.6
1801-1900	70	4	1	0.2
1901-2000	67	6	1	0.4
2001-2100	63	10	3	0.7
2101-2200	60	9	2	0.6
2201-2300	57	7	1	0.5
2301-2400	54	14	3	1.0
2401-2500	51	12	3	1.0
2501-2600	49	13	6	0.5
2601-2700	47	15	5	0.4
2701-2800	45	12	5	1.0
2801-2900	43	10	6	0.9
2901-3000	42	8	1	0.5
		248	82	12.3

7. SOME FINAL REMARKS

Our principal result is as follows. If there exists a word y which is semantically related to another word x (this relation may be morphologically expressed by it seems to be not necessary), then it is unlikely for the frequencies x and y to differ in an arbitrary manner. The very fact of the existence of such a dependency between the frequencies was discovered some time ago (THORNDIKE, 1943; ŠECHTMAN, 1978). The new idea is that it is possible under certain conditions to obtain a simple mathematical expression for this dependency (see postulate (3)).

This condition is not trivial. As it turns out, the frequencies (ranks) of the words would have systematical relationships, if the semantical links between them were uniformly related pairs of words that create a homogeneous pattern.

The concept of homogeneity is not a mere auxiliary notion to define the regularity; its role in classifying semantic relations may be more important.

There are many counts of the relative or absolute frequencies of grammatical categories (e.g. cases, tenses, number, etc.). These counts themselves are very useful, yet interpretations of them may be misleading. Being based on a similar count, the conclusion that one category occurs more frequently than another by a factor n, in general makes no sense unless the pattern in question is homogeneous.

For example, let the number of nouns in the singular in some sample be greater than that of nouns in the plural by a factor n. Nevertheless it is possible that there is no single reasonably coherent group of nouns for which this ratio of singular and plural holds.

The homogeneous patterns (regular ones and, strangely enough, irregular ones, too) have an additional curious property: the average value of f_y / f_x, of Pl and Sg forms of the Polish adjectives is 0.55 (see Example 1), whereas the productivity ratio is about 0.6.

Analogously, for Example 2 we have a frequency ratio of

Gen to Nom 2.26, whereas their productivity ratio is about 2.3. The same holds for Example 3 (here we have 1.17 and 1.2 for inanimate nouns and 0.12 and 0.1 for animate ones, and Example 4 (0.42 and 0.4 respectively). Hence we have a simple method for obtaining a rough estimate of the mean $\bar{m}_{xy}$. Unfortunately, such a simple method for the evaluation of the bariance of $\bar{m}_{xy}$ is unknown.

NOTES

1 It is possible to state the problem in a more general way and speak not only about pairs, but also about n-tiples of words. In this article, however, the simplicity of the formal apparatus is more valuable than its generality.

2 The frequency of usage does not appear to be an inherent property of a word as a unit of Saussure's langue: the data of frequency lists compiled on the basis of different corpora of texts (of the same language) are not in good accordance with each other. Then the results obtained below would be spread on those fragments of the language on which exclusively can be extrapolated the data of pertinent frequency lists.

3 This fact may be stated in a more precise manner. Let us examine the distribution of the number $K_{j\lambda}$ of words from a productive class X_j which should appear in the interval of ranks (x, x'), $\lambda = /x - x'/$. For every interval (x, x') except the intervals which contain the most frequent words this number has a Poisson distribution,

$$P\{k|\lambda = K\} = \frac{|\lambda j}{K!} \bar{1}^{|\lambda j} \quad , \quad = 0, 1, 2, \ldots$$

4 The choice of a basis of logarithm is irrelevant not only from the theoretical, but also the practical point of view. In an age when necessary computations are usually carried out with the aid of electronic calculators, the Briggs-logarithm has no special advantage over the natural one.

5 Yet some of them can be used as colloquialisms. Cf. Как поживаешь? - Да так, средне. 'How are you? - Just middling.'

REFERENCES

ARAPOV, M.V., Produktivnost' v estestvennom jazyke i ee izmerenie. Voprosy informacionnoj teorii i praktiki 23, 1974, 117-138.

ARAPOV, M.V., CHERC, M.M., Matematičeskie metody v istoričeskoj lingvistike. Moskva: Nauka, 1974. Deutsche Übersetzung i.Vb., Bochum: Brockmeyer.

LEWICKI, A., MASŁOWSKI, W., SAMBOR, J., WORONCZAK, J., Słownictwo współczesnego języka polskiego, t. III: Publicystyka. Warszawa, 1975.

MEL'ČUK, I.A., Opyt teorii lingvističeskix modelej "Smysl ⟷ Tekst". Moskva: Nauka, 1974.

ŠECHTMAN, N.A., Častotnost' slova i ego pozicija v tezauruse. Naučno-techničeskaja informacija, Ser. 2, N 5, 1978, 20-21.

ŠTEJNFEL'DT, E.A., Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka. Tallin, 1963.

THORNDIKE, E.L., Derivation ratios. Language 19, 1943, 27-37.

ZALIZNJAK, A.A., Grammatičeskij slovar' russkogo jazyka. Moskva: Russkij jazyk, 1977.

ZASORINA, L.N. (Red.), Častotnyj slovar' russkogo jazyka. Moskva: Russkij jazyk, 1977.

Die Notwendige Asymmetrie binärer Stammbäume

Joachim Boy, Essen

0. Wie in anderen Bereichen der Wissenschaft haben auch in der Linguistik graphische Darstellungen mit Hilfe sogenannter "binärer Stammbäume" ihren festen Platz. Sie dienen zur Darstellung der Merkmalsausprägungen verschiedener sprachlicher Elemente, die miteinander verglichen und in Beziehung gesetzt werden sollen. Besonders geläufig sind derartige Bäume etwa in der Phonologie oder in der Klassifikation von Sprachen. Allgemein lassen sich die hierarchischen Strukturen eines Systems mit Hilfe eines solchen Stammbaumes erfassen, sofern seine Eigenschaften sich in binärer Form abbilden lassen. Andererseits erlaubt die Beschreibung bzw. Analyse eines binären Stammbaums mit Hilfe geeigneter Methoden Strukturaussagen über das abgebildete System. In dieser Hinsicht kommt der Entwicklung adäquater Meß- und Beschreibungsverfahren für binäre Stammbäume eine gewisse Bedeutung für sprachwissenschaftliche Untersuchungen zu.

Der vorliegende Artikel stützt sich auf das von Sander / Altmann (1973) entwickelte Verfahren zur Bestimmung der Asymmetrie binärer Stammbäume, das sich in ähnlicher Form, allerdings in größerem Zusammenhang, nochmals bei Altmann/Lehfeldt (1980) findet. Die theoretischen Grundlagen des Verfahrens sollen hier nicht nochmals dargestellt werden. Allerdings soll als notwendige Einführung in das hier zu behandelnde Problem eine kurze Darlegung der Grundzüge des dort entwickelten Verfahrens zur Messung der Asymmetrie binärer Stammbäume vorangestellt werden.

Die Entscheidungsbäume werden dabei aufgefaßt als gerichtete Graphen, deren Knoten, ausgehend vom Anfangsknoten, untereinander bis hin zu den Endknoten durch Pfeile ("Kanten") verbunden sind.

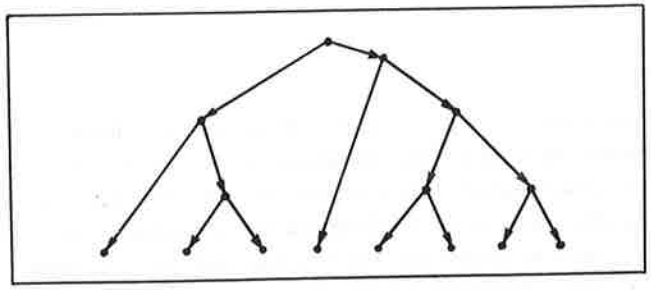


Abb. 1: Beispiel eines binären Stammbaums

Aus der Anzahl der von jedem Knoten ausgehenden Pfeile ergeben sich zwei Klassen von Knoten: von "Entscheidungsknoten" gehen jeweils zwei Kanten aus, von "Endknoten" keine. Außer diesen beiden Arten von Knoten gibt es in einem rein binären Stammbaum keine weiteren Klassen; dabei repräsentieren die Endknoten die erfaßten Untersuchungsobjekte. Die Gesamtheit aller nachfolgenden Kanten und Knoten eines beliebigen Entscheidungsknotens wird als "Unterbaum" bezeichnet; die Gesamt - Asymmetrie eines Baumes ergibt sich nach dem hier zugrundeliegenden Verfahren aus der Summe aller Einzel - Asymmetrien der Unterbäume, gewichtet durch die von der Anzahl der Endknoten abhängige "notwendige Asymmetrie" des Gesamtbaums.

Zunächst wird zu jedem Entscheidungsknoten i untersucht, wie viele Endknoten von ihm aus links bzw. rechts erreichbar sind; ihre Anzahl wird mit l_i bzw. r_i bezeichnet. Mit dem Asymmetriemaß soll nur eine rein numerische Größe für die Asymmetrie ermittelt werden, so daß nicht nach verschiedenen Arten der Asymmetrie, wie "Links-" oder "Rechtsasymmetrie", unterschieden wird. Aus diesem Grund geht in das Maß nur der

Unterschiedsbetrag der Anzahl der erreichbaren Endknoten ein, also $|l_i - r_i|$. Für einen Baum mit n Endknoten (und damit n-1 Entscheidungsknoten) ergibt sich damit zunächst als Maß der Asymmetrie:

$$A' = \frac{\sum_{i=1}^{n-1} |l_i - r_i|}{A_{max}} \qquad (1)$$

A_{max} , die maximal mögliche Asymmetrie eines Baumes mit n Endknoten, läßt sich errechnen als

$$A_{max} = \frac{(n-2)\ (n-1)}{2} ,$$

so daß sich als weiterhin vorläufige Formel zur Berechnung der Asymmetrie der folgende Ausdruck ergibt:

$$A' = \frac{\sum_{i=1}^{n-1} |l_i - r_i|}{\frac{(n-2)\ (n-1)}{2}} \qquad (2)$$

Im Sinne dieses Asymmetriemaßes können nur Bäume mit $n = 2^k$ (k eine beliebige natürliche Zahl) völlig symmetrisch sein, d.h. die Asymmetrie Null besitzen. (Die optisch geläufige Vorstellung, daß ein beliebiger Baum mit einer geraden Anzahl von Endknoten völlig symmetrisch sein kann, ist hier außer acht zu lassen. Die spiegelbildlichen Hälften eines geradzahligen Baums mit $n \neq 2^k$ Endknoten heben die jeweils entsprechenden Links- und Rechtsasymmetrien nicht gegeneinander auf, sondern summieren aus den o.g. Gründen die Beträge $|l_i - r_i|$.) Alle binären Stammbäume mit $n \neq 2^k$ Endknoten besitzen daher eine gewisse "Notwendige Asymmetrie" (weiterhin kurz als "NA" bezeichnet), die als NA(n) allein von der Anzahl der Endknoten abhängig ist. Die notwendige Asymmetrie ist bei der Berechnung der Asymmetrie mit zu berücksichtigen,

da sonst die errechneten Asymmetriegrößen verschiedener Bäume nicht miteinander vergleichbar wären. Als endgültiges Maß für die Asymmetrie eines binären Stammbaumes erhält man daher:

$$A = \frac{2\left(\sum_{i=1}^{n-1} |l_i - r_i| - NA(n)\right)}{(n-1)\ (n-2)} \qquad (3)$$

Die Berechnung von NA(n) birgt einige Probleme. Sander / Altmann und Altmann / Lehfeldt führen für NA(n) die folgende Formel ein:

$$NA(n) = NA\left(\left[\frac{n}{2}\right]\right) + NA\left(n - \left[\frac{n}{2}\right]\right) + n - 2\left[\frac{n}{2}\right] \qquad (4)$$

Die Ermittlung von NA(n) insbesondere für größere Werte von n ist wegen der notwendigen Rekursion auf n = 1, 2 oder 2^k etwas langwierig. Eine Rekursion auf diese Werte ist notwendig, da deren NA(n) als trivialer Fall gleich Null, jedenfalls aber bekannt ist. Eine Rekursion auf andere Werte von n wäre theoretisch ebenfalls möglich, doch setzt dies weitere bekannte, also vorher berechnete Größen von NA(n) voraus. Sander/Altmann (1973) geben als Umformung zur Formel (3) den folgenden Ausdruck an, dessen Handhabung sich allerdings als noch schwieriger erweist:

$$NA(n) = \frac{1}{2} \sum_{i=0}^{2^{i+1} \leq n} \left(2^i - \left|\sum_{j=0}^{2^i-1} (-1)^{\left[\frac{n+j}{2^i}\right]}\right|\right) \qquad (5)$$

Als wesentliche Hilfe bei Asymmetrieuntersuchungen an binären Stammbäumen erweisen sich daher die in den vorgenannten Arbeiten enthaltenen Tabellen der NA(n) für Werte von n zwischen 1 und 100. Für die meisten Untersuchungen gerade aus dem

sprachwissenschaftlichen Bereich dürfte dieser Umfang ausreichen; andererseits sind auch auf diesem Gebiet Beispiele denkbar (wie etwa in der Semantik oder bei der Klassifikation einer sehr großen Anzahl von Sprachen), in denen die Zahl der zu untersuchenden Objekte wesentlich größer ist. In dem vorliegenden Artikel soll daher versucht werden, vereinfachte Verfahren zur Berechnung der notwendigen Asymmetrie binärer Stammbäume zu entwickeln, die ohne größeren rechnerischen Aufwand auch die Bearbeitung umfangreicheren Materials gestatten und darüberhinaus natürlich auch die Beschreibung kleinerer Bäume ohne Rückgriff auf die Tabellendaten ermöglichen. Eine Vereinfachung der bei Sander/Altmann bzw. Altmann/Lehfeldt erarbeiteten Formeln zu diesem Zweck war nicht möglich, so daß hier ein anderer Weg gesucht wurde. Dabei wurden die in den Tabellen angegebenen Werte der NA(n) zur Grundlage genommen; anschließend wurde versucht, Rechenvorschriften zu erarbeiten, die zu denselben Ergebnissen führen, ohne sich dabei auf die vorgegebenen Verfahren zu stützen. Es werden drei verschiedene Verfahren vorgestellt. Im ersten Verfahren wird die Funktion NA(n) in die Summe einer gleichmäßigen Folge von Teil- bzw. Elementarfunktionen zerlegt und aus dieser Zerlegung eine allgemeine Formel abgeleitet. Im zweiten Verfahren werden implizite Eigenschaften der im ersten Verfahren erarbeiteten Summenfunktion zur Aufstellung einer allgemeinen Formel für NA(n) verwendet. Das dritte Verfahren stellt als Ableitung aus dem zweiten Verfahren eine stark vereinfachte Methode speziell für die Berechnung der NA(n) ohne Verwendung technischer Hilfsmittel dar. Zu allen Verfahren werden mehrere Beispiele durchgeführt.

1. Die Werte der NA(n) für n zwischen 1 und 100 sind in Tabelle 1 angegeben. Wegen der größeren Anschaulichkeit wurden die dort erhaltenen Werte für NA(n) in Abbildung 2 graphisch in Abhängigkeit von n dargestellt. Obwohl NA(n) als Funktion der diskreten Menge der n ebenfalls nur diskrete

n	NA(n)	n	NA(n)	n	NA(n)	n	NA(n)
1	0	26	10	51	21	76	32
2	0	27	10	52	20	77	34
3	1	28	8	53	21	78	34
4	0	29	8	54	20	79	34
5	2	30	6	55	19	80	32
6	2	31	4	56	16	81	36
7	2	32	0	57	17	82	38
8	0	33	5	58	16	83	40
9	3	34	8	59	15	84	40
10	4	35	11	60	12	85	42
11	5	36	12	61	11	86	42
12	4	37	15	62	8	87	42
13	5	38	16	63	5	88	40
14	4	39	17	64	0	89	42
15	3	40	16	65	6	90	42
16	0	41	19	66	10	91	42
17	4	42	20	67	14	92	40
18	6	43	21	68	16	93	40
19	8	44	20	69	20	94	38
20	8	45	21	70	22	95	36
21	10	46	20	71	24	96	32
22	10	47	19	72	24	97	36
23	10	48	16	73	28	98	38
24	8	49	19	74	30	99	40
25	10	50	20	75	32	100	40

Tabelle 1: Notwendige Asymmetrie NA(n) von Stammbäumen mit n Endknoten (aus Altmann/Lehfeldt 1980)

Werte annehmen kann, werden die Datenpunkte hier, ebenfalls aus Gründen der Anschaulichkeit, miteinander verbunden. In der Abbildung wird die Gesamtstruktur der Funktion NA(n) augenfällig: sie zerfällt durch ihre Nullstellen bei allen n mit $n = 2^k$ (k natürliche Zahl) in regelmäßige Intervalle,

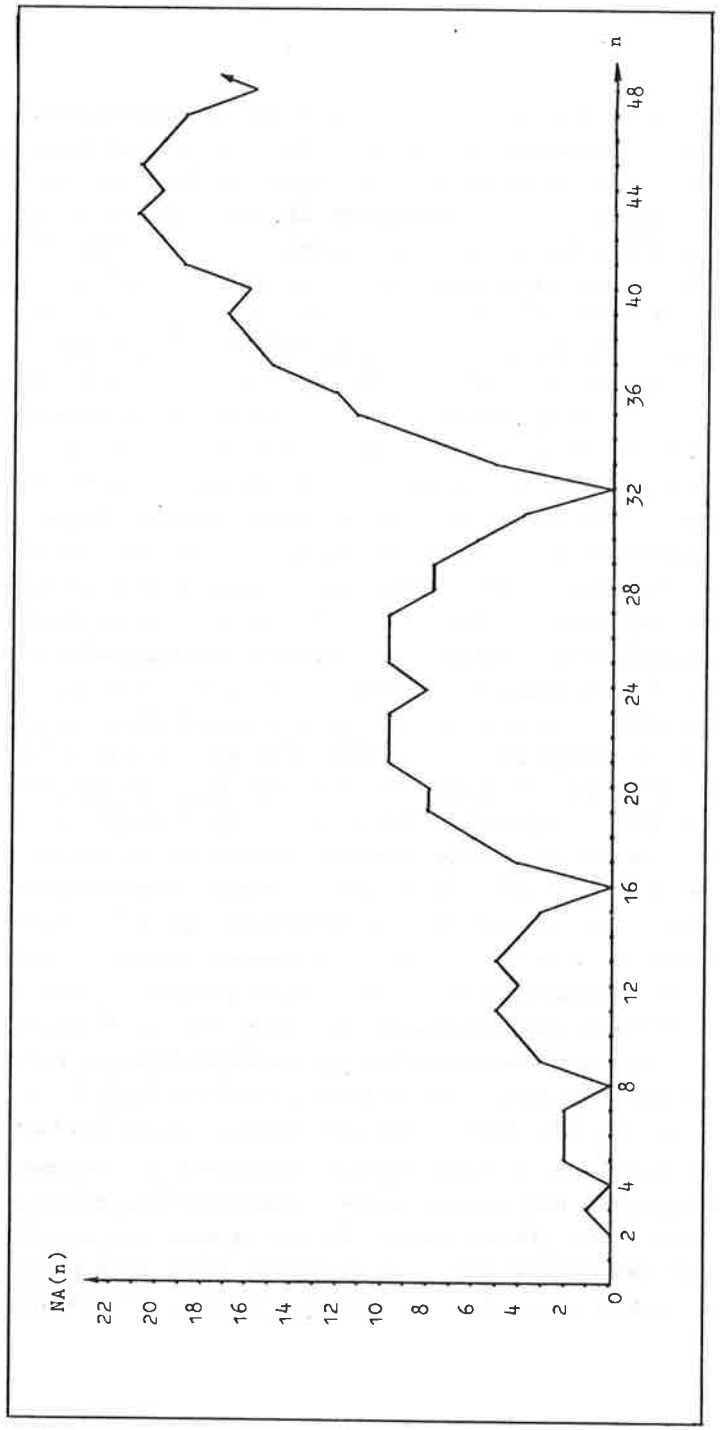


Abb. 2: Werte der Notwendigen Asymmetrie NA(n) in Abhängigkeit von der Anzahl n der Endknoten (vgl. Tabelle 1).

deren jedes doppelt so lang ist wie das vorangegangene. (Auf weitere Einzelheiten soll hier zunächst nicht eingegangen werden.) Geht man davon aus, daß sich die Struktur der binären Stammbäume, hier verstanden als Gesamtheit bzw. Aufeinanderfolge gleichartiger, regelmäßig wiederholter binärer Entscheidungen, in irgendeiner Weise ebenso durchgängig in der Feinstruktur der Funktion NA(n) abbildet, so ist für unsere Zwecke besonders die Untersuchung des ersten und damit auch kleinsten Intervalls ⟨2,4⟩ von Interesse. Dieses Intervall enthält neben den NA(n) - Werten für zwei völlig symmetrische Bäume mit $n = 2$ und $n = 4$ Endknoten auch NA(3), d.h. den Wert für die notwendige Asymmetrie eines "minimalen" asymmetrischen Baumes. Existiert eine elementare Funktion als Äquivalent zu diesem Minimalbaum, so ist für ihre Ermittlung die Untersuchung der Beziehung dieses ersten Intervalls zu den folgenden von Interesse. (Die Annahme einer durchgängigen Bedeutung des NA(n) - Wertes eines minimalen Baumes erscheint auch deswegen sinnvoll, weil sich jeder vollständig symmetrische Baum mit $n = 2^k$ Endknoten und damit NA(n) = 0 restlos in ebenfalls symmetrische Teilbäume zerlegen läßt, deren NA(n) jeweils auch stets Null beträgt. Analog hierzu wird erwartet, daß auch die von Null verschiedenen Werte von NA(n) zu denen in vorangegangenen Intervallen im Sinne einer Zerlegung in Teilbäume in Beziehung stehen. Daneben weist auch die Tatsache, daß für die Berechnung der NA(n) bisher eine R e k u r s i o n s formel verwendet wurde, in diese Richtung. Unterstützt wird die Vermutung ferner von einem weiteren Indiz: Vernachlässigt man die durch die diskrete Struktur des Wertebereichs bedingte eckige Form der Kurve in Abbildung 2, so zeigt sich eine frappierende Ähnlichkeit ihrer Gestalt mit den Bildern von Überlagerungen periodischer Schwingungen, wie sie aus der physikalischen Schwingungslehre geläufig sind. Die Annahme eines Zusammenwirkens mehrerer im weitesten Sinne gleichförmiger Teilfunktionen bei der Erzeugung der Gesamtfunktion NA(n) erscheint daher auch unter dem Aspekt dieser etwas ferner liegenden Analogie als plausibel.)

Aus Abb. 2 wird deutlich, daß die Werte für NA(n) in jedem Intervall symmetrisch zu dessen Mitte liegen; die Gesetzmäßigkeit des Anstiegs der Funktionswerte spiegelt sich in der zweiten Hälfte wieder. Reiht man das Intervall ⟨2,4⟩ zweimal so aneinander, daß der Endpunkt des ersten und der Anfangspunkt des zweiten Intervalls zusammenfallen, und subtrahiert die NA(n) - Werte des so erhaltenen Doppelintervalls (= 0,1,0, 1,0) von den entsprechenden Werten des ebensolangen Intervalls ⟨4,8⟩ (= 0,2,2,2,0), so erhält man die Wertefolge 0,1,2,1,0. Verdoppelt man in gleicher Weise das Intervall ⟨4,8⟩ und subtrahiert entsprechend die NA(n) - Werte dieses Doppelintervalls von denen des Intervalls ⟨8,16⟩, so erhält man die Differenzenfolge 0,1,2,3,4,3,2,1,0. Dieses Vorgehen läßt sich beliebig fortsetzen; die erhaltene Differenzenfolge zeigt in allen Fällen ein völlig gleichmäßiges Bild. Beginnend mit Null für $n = 2^k$ steigen ihre Werte mit um Eins wachsendem n ebenfalls jeweils um Eins an, bis sie in der Mitte des Intervalls, also für $n = \frac{3}{2}\,2^k$, ihren Maximalwert von $\frac{1}{2}\,2^k$ (oder 2^{k-1}) erreichen, um dann in der zweiten Hälfte des Intervalls ebenso gleichmäßig bis auf Null am Endpunkt des Intervalls bei 2^{k+1} abzufallen. Bezeichnet man die Differenzenfolgen als Funktionen von n in der Reihenfolge ihres Auftretens mit Indizes, so beginnt $f_1(n)$ bei $n = 2$ mit der Periode 2 und der "Amplitude" 1 ("Amplitude" hier verstanden als Differenz zwischen Null und dem Maximalwert der Funktion in der Intervallmitte und nicht, wie sonst üblich, als halbe Schwingungsweite). $f_2(n)$ beginnt bei $n = 4$ mit der Periode 4 und der Amplitude 2; $f_3(n)$ beginnt bei $n = 8$ mit der Periode 8 und der Amplitude 4 usw. Eine Darstellung der Funktionen $f_i(n)$ für i von 1 bis 5 zeigt Abbildung 3. Stellt man eine Wertetabelle dieser Funktionen auf, so erkennt man, daß sich die Werte von NA(n) vollständig durch Ordinatenaddition der $f_i(n)$ erzeugen lassen. Tabelle 2 zeigt diese Daten. Diese Tatsache entspricht exakt der optisch auffälligen Analogie aus der Schwingungslehre bezüglich des Verlaufs der Funktion NA(n) und seiner Ähnlichkeit mit Überlagerungsschwingungen.

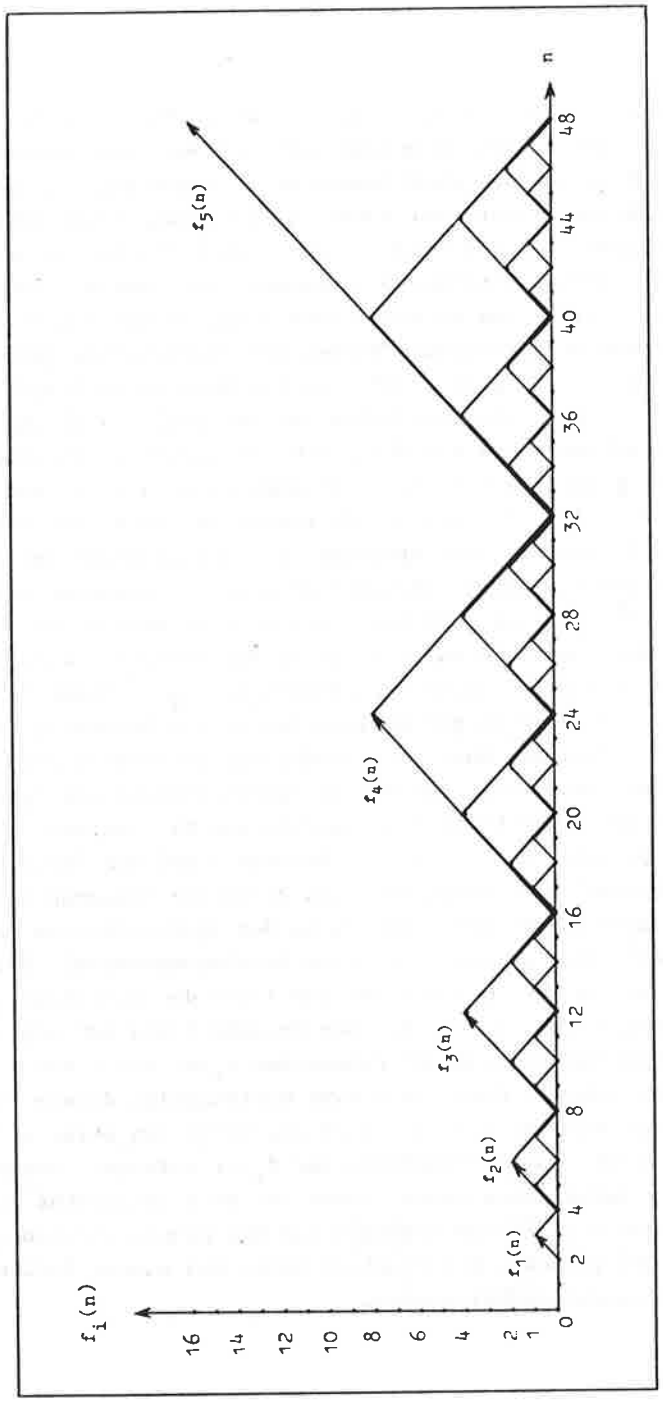


Abb. 3: Funktionswerte $f_i(n)$ für i = 1 bis 5.
(vgl. Tabelle 2).

n	$f_1(n)$	$f_2(n)$	$f_3(n)$	$f_4(n)$	$f_5(n)$	$\sum_{i=1}^{5} f_i(n) = NA(n)$
2	0					0
3	1					1
4	0	0				0
5	1	1				2
6	0	2				2
7	1	1				2
8	0	0	0			0
9	1	1	1			3
10	0	2	2			4
11	1	1	3			5
12	0	0	4			4
13	1	1	3			5
14	0	2	2			4
15	1	1	1			3
16	0	0	0	0		0
17	1	1	1	1		4
18	0	2	2	2		6
19	1	1	3	3		8
20	0	0	4	4		8
21	1	1	3	5		10
22	0	2	2	6		10
23	1	1	1	7		10
24	0	0	0	8		8
25	1	1	1	7		10
26	0	2	2	6		10
27	1	1	3	5		10
28	0	0	4	4		8
29	1	1	3	3		8
30	0	2	2	2		6
31	1	1	1	1		4
32	0	0	0	0	0	0
33	1	1	1	1	1	5
34	0	2	2	2	2	8
35	1	1	3	3	3	11
36	0	0	4	4	4	12
37	1	1	3	5	5	15
38	0	2	2	6	6	16
39	1	1	1	7	7	17
40	0	0	0	8	8	16
41	1	1	1	7	9	19
42	0	2	2	6	10	20
43	1	1	3	5	11	21
44	0	0	4	4	12	20
45	1	1	3	3	13	21
46	0	2	2	2	14	20
47	1	1	1	1	15	19
48	0	0	0	0	16	16
...	...	...	...	...	...	...

Tabelle 2: Teilfunktionen $f_1(n)$ bis $f_5(n)$ und Summe der Funktionswerte (= NA(n))

Die Anzahl der zur Berechnung von NA(n) erforderlichen Einzelfunktionen $f_i(n)$ läßt sich mit Hilfe der folgenden Überlegungen leicht festlegen. Jedes n am Beginn eines Intervalls läßt sich als Potenz einer natürlichen Zahl zur Basis Zwei darstellen. Darüberhinaus gibt es im Bereich der natürlichen Zahlen größer als Zwei keine weiteren Zweierpotenzen mit natürlichen Zahlen als Exponenten, die nicht gleichzeitig Beginn eines Intervalls mit $NA(n) = 0$ sind. Jede Teilfunktion $f_i(n)$ kann nur am Anfangspunkt eines Intervalls beginnen. Daraus ergibt sich, daß, unter Einbeziehung der Anfangspunkte der Funktionen $f_i(n)$, wie sie oben näher bezeichnet wurden, die Anzahl der zur Errechnung von NA(n) in jedem beliebigen Intervall erforderlichen Funktionen $f_i(n)$ genau dem Exponenten der Zweierpotenz am Anfangspunkt des Intervalls entspricht. Will man also ein NA(n) berechnen, so ist zunächst festzustellen, welches die nächstniedrigere Zweierpotenz (und damit der Intervallbeginn) ist; deren Exponent zur Basis Zwei ergibt die Anzahl der zu addierenden Funktionswerte. Um beispielsweise NA(43) zu berechnen, findet man als nächstniedrigere Zweierpotenz unterhalb von 43 die Zahl $32 = 2^5$; es sind also fünf Teilfunktionen $f_i(n)$ zu addieren, $f_1(43)$ bis $f_5(43)$. Die Zahl 43 läßt sich, wie jede andere natürliche Zahl, als Potenz von Zwei mit rationaler Hochzahl ausdrücken (also $43 = 2^{5,\cdots}$). Der ganzzahlige Teil des Exponenten von n selbst ergibt damit ebenso die gewünschte Zahl der Additionsschritte. Diese läßt sich daher kürzer ausdrücken als ganzzahliger Wert ("Integer", ausgedrückt durch die Klammerung $[\ldots]$) des Logarithmus von n zur Basis Zwei, also als "dualer Logarithmus" (oder ld n) von n, kurz daher: [ld n]. Allgemein erhalten wir daher als Formel für NA(n) vorläufig folgenden Ausdruck:

$$NA(n) = \sum_{i=1}^{[ldn]} f_i(n) \qquad (6)$$

Somit bleibt nur noch ein genereller Ausdruck zur Berechnung der Funktionswerte der einzelnen $f_i(n)$ zu ermitteln. Für alle $f_i(n)$ gilt, unabhängig von der Größe des i, daß sie jeweils in der Mitte zwischen zwei Nullstellen, also in der Hälfte der Periode, ihr Maximum besitzen. Für alle Perioden der Funktion $f_i(n)$ beträgt dieser Maximalwert konstant $\frac{1}{2}\,2^i$ oder 2^{i-1}. Beispielsweise beträgt der Maximalwert von $f_1(n)$ $\frac{1}{2}\,2^1 = \frac{1}{2}\,2 = 1$; der Maximalwert von $f_4(n)$ beträgt $\frac{1}{2}2^4 = 8$. Liegt n genau in der Mitte einer Periode von $f_i(n)$, so besitzt $f_i(n)$ den Wert 2^{i-1}; für alle anderen n innerhalb derselben Periode ist der Funktionswert geringer, und zwar genau um so viel, wie der Abstand von n zum Mittelpunkt der Periode beträgt, gleichgültig, ob n größer oder kleiner ist. Wie oben besprochen, beginnt die Funktion $f_i(n)$ das erste Mal bei $n = 2^i$; die Länge ihrer Periode beträgt ebenfalls 2^i. Die Anzahl der Vorgängerperioden zu derjenigen von n ergibt sich damit als

$$K = \left[\frac{n}{2^i}\right] ,$$

und da jede Periode die Länge 2^i besitzt, liegt der Beginn der zu n gehörigen Periode von $f_i(n)$ bei

$$B = K \cdot 2^i = \left[\frac{n}{2^i}\right] \cdot 2^i .$$

Zum Beispiel: Zur Berechnung von NA(43) müssen $f_1(43)$ bis $f_5(43)$ addiert werden. Der Beginn der Periode etwa von $f_4(43)$ liegt dann bei

$$\left[\frac{43}{2^4}\right] \cdot 2^4 = \left[\frac{43}{16}\right] \cdot 16 = [2{,}6875] \cdot 16 = 2 \cdot 16 = 32$$

Die Mitte dieser Periode liegt bei

$$\left[\frac{n}{2^i}\right] \cdot 2^i + 2^{i-1}, \text{ also bei } \left[\frac{43}{2^4}\right] \cdot 2^4 + 2^3 = 32 + 8 = 40$$

Wie oben ausgeführt, entspricht der Funktionswert selbst genau dem Maximalwert der Periode vermindert um den Abstand (also den Betrag der Differenz) von n zur Mitte der Periode. Als allgemeine Formel für $f_i(n)$ erhalten wir somit:

$$f_i(n) = 2^{i-1} - \left| \left[\frac{n}{2^i} \right] \cdot 2^i + 2^{i-1} - n \right| \qquad (7)$$

Für das Beispiel $f_4(43)$ ergibt sich dann:

$$f_4(43) = 2^{4-1} - \left| \left[\frac{43}{2^4} \right] \cdot 2^4 + 2^{4-1} - 43 \right|$$

$$= 8 - |32 + 8 - 43|$$
$$= 8 - |-3|$$
$$= 8 - 3$$
$$= 5$$

Mit der allgemeinen Formel für $f_i(n)$ läßt sich nunmehr in Verbindung mit Formel (6) NA(n) bestimmen als:

$$NA(n) = \sum_{i=1}^{[ldn]} \left(2^{i-1} - \left| \left[\frac{n}{2^i} \right] \cdot 2^i + 2^{i-1} - n \right| \right) \qquad (8)$$

Dieser Ausdruck läßt sich durch das Zerlegen der Summe etwas vereinfachen. Da gilt:

$$\sum_{i=1}^{r} 2^{i-1} = 2^r - 1 \quad ,$$

erhält man für NA(n) endgültig folgende Formel:

$$NA(n) = 2^{[ldn]} - 1 - \sum_{i=1}^{[ldn]} \left| \left[\frac{n}{2^i} \right] \cdot 2^i + 2^{i-1} - n \right| \qquad (9)$$

An den Beispielen n = 43 und n = 432 soll die Berechnung von NA(n) nach Formel (9) ausführlich dargestellt werden.

$$NA(43) = 2^5 - 1 - \sum_{i=1}^{5} \left| \left[\frac{43}{2^i} \right] \cdot 2^i + 2^{i-1} - 43 \right|$$

$$= 31 - (|42 + 1 - 43| + |40 + 2 - 43| + |40 + 4 - 43| + |32 + 8 - 43| + |32 + 16 - 43|)$$
$$= 31 - (|0| + |-1| + |1| + |-3| + |5|)$$
$$= 31 - 10$$
$$= 21$$

$$NA(432) = 2^8 - 1 - \sum_{i=1}^{8} \left| \left[\frac{432}{2^i} \right] \cdot 2^i + 2^{i-1} - 432 \right|$$

$$= 255 - (|432 + 1 - 432| + |432 + 2 - 432| + |432 + 4 - 432| + |432 + 8 - 432| + |416 + 16 - 432| + |384 + 32 - 432| + |384 + 64 - 432| + |256 + 128 - 432|)$$
$$= 255 - (1 + 2 + 4 + 8 + 0 + 16 + 16 + 48)$$
$$= 255 - 95$$
$$= 160$$

2. Das im folgenden Abschnitt beschriebene Verfahren zur Berechnung der NA(n) geht vom Vergleich der NA(n) - Werte für positionell übereinstimmende n in verschiedenen Intervallen aus. Die dem Verfahren zugrundeliegende Eigenschaft der NA(n)-Funktion geht deutlich aus Tabelle 1 und Abbildung 2 hervor. Betrachtet man die jeweils ersten von Null verschiedenen Werte von NA(n) in jedem Intervall, so sieht man, daß sich in aufsteigender Folge der Intervalle diese Werte um jeweils Eins unterscheiden: NA(3) = 1 ; NA(5) = 2 ; NA(9) = 3 ; NA(17) = 4.

Sie bilden, beginnend mit dem ersten Intervall, eine aufsteigende arithmetische Folge mit dem Anfangsglied NA(3) = 1 und der Differenz 1. Die jeweils zweiten von Null verschiedenen Werte eines jeden Intervalls unterscheiden sich in gleicher Weise um Zwei; sie bilden, beginnend mit NA(6) = 2, eine arithmetische Folge mit der Differenz 2. Für alle weiteren von Null verschiedenen Werte der Funktion NA(n) lassen sich analoge Folgen bilden. Diese Tatsache soll hier zur Ableitung eines weiteren Verfahrens zur Berechnung beliebiger NA(n) herangezogen werden.

Die oben beschriebene regelmäßige Aufeinanderfolge der NA(n) - Werte positionell übereinstimmender n aus verschiedenen Intervallen läßt sich sehr einfach durch die im ersten Verfahren erläuterte Zusammensetzung der NA(n) aus einzelnen gleichförmigen Teilfunktionen $f_i(n)$ erklären, deren spezielle Eigenschaften hier jedoch nicht nochmals dargestellt werden sollen. Aus Abbildung 3 wird deutlich, daß bei einem paarweisen Vergleich positionell übereinstimmender n aus unmittelbar aufeinanderfolgenden Intervallen sich für die jeweiligen $f_i(n)$ dieselben Funktionswerte ergeben; ausgenommen allerdings der $f_i(n)$ - Wert für das größte i des größeren n, da diese Funktion erst im Intervall des größeren n zu laufen beginnt. Bildet man die Differenz zwischen zweien solcher NA(n) - Werte, so heben sich dementsprechend alle "rangniedrigeren" $f_i(n)$ - Werte gegeneinander auf. Da die "ranghöchste" $f_i(n)$ - Funktion im Intervall des größeren n mit Null am Intervallbeginn einsetzt und linear mit der Steigung Eins wächst, ergibt sich die Differenz dieser beiden NA(n) - Werte als Abstand des größeren n vom Beginn seines Intervalls. Zwei beliebige positionell entsprechende n aus aufeinanderfolgenden Intervallen unterscheiden sich in ihrer Größe um den Betrag der Länge des niedrigeren Intervalls:

$$n_{i+1} - n_i = 2^{[\mathrm{ld}\, n_i]} \qquad (10)$$

Damit ergibt sich allgemein für die Differnz der entsprechenden NA(n) - Werte:

$$NA(n_i) - NA(n_i - 2^{[\mathrm{ld}\, n_i] - 1}) = n_i - 2^{[\mathrm{ld}\, n_i]} \qquad (11)$$

Dies bedeutet, daß sich der Wert jedes NA(n) bestimmen läßt als Funktion des NA(n) - Wertes des positionell entsprechenden Vorgängers im voranstehenden Intervall. Mithin läßt sich mit Hilfe eines geeigneten iterativen Verfahrens jeder NA(n)-Wert errechnen aus den elementaren Werten NA(2) = 0 und NA(3) = 1. Die hierzu erforderliche Rekursion erfolgt in Sprüngen abwärts zum jeweiligen Vorgängerintervall bis zum Intervall $\langle 2,4 \rangle$, wobei sich NA(n) ergibt als Summe der Differenzen zwischen dem positionell entsprechenden n und dem zugehörigen Beginn der durchlaufenen Intervalle.

Da jedes Intervall doppelt so lang ist wie das vorangegangene, lassen sich mit Hilfe der oben angeführten arithmetischen Folgen genau genommen nur die NA(n) - Werte aus der ersten Hälfte eines jeden Intervalls errechnen, da der Sprung von einem n aus der zweiten Hälfte eines Intervalls auf keinen positionell entsprechenden Vorgänger träfe. Aus den Symmetrieeigenschaften der Teilfunktionen $f_i(n)$ ergibt sich jedoch, daß auch (anschaulich erkennbar in Abbildung 2) die Funktion NA(n) selbst innerhalb jedes Intervalls achsensymmetrisch zur Intervallmitte verläuft. Durch "Klappung" um die Intervallmitte läßt sich daher innerhalb jedes Intervalls ein Übergang zu Werten von n mit positionell entsprechenden Vorgängern herstellen. Da für jedes Intervall, in dem sich ein beliebiges n befindet, die Mitte bei $\frac{3}{2} 2^{[\mathrm{ld}\, n]}$ liegt, läßt sich der Übergang beschreiben als Transformation

$$n \longrightarrow \tfrac{3}{2} \cdot 2^{[\mathrm{ld}\, n]} - \left| \tfrac{3}{2} \cdot 2^{[\mathrm{ld}\, n]} - n \right| \qquad (12)$$

Zum Beispiel: Für $n = 53$ liegt die zugehörige Intervallmitte bei

$$\frac{3}{2} \cdot 2^{[\mathrm{ld}\ 53]} = \frac{3}{2} \cdot 2^{[5,7279]} = \frac{3}{2} \cdot 2^5 = 48 .$$

$n = 53$ wird dann nach Formel (12) transformiert nach

$$\begin{aligned} 53 \longrightarrow\ & 48 - |48 - 53| \\ = \ & 48 - |-5| \\ = \ & 48 - 5 \\ = \ & 43 \end{aligned}$$

(Wie sich anhand von Tabelle 1 leicht überprüfen läßt, gilt tatsächlich $NA(53) = NA(43) = 21$.)

Zu bedenken ist, daß zur Errechnung der NA(n) lediglich die Summe der Differenzen zwischen den Werten von n und dem jeweiligen Intervallbeginn gebildet werden muß, so daß eine Transformation oder zumindest eine Überprüfung jedes beim Durchlaufen eines Intervalls errechneten Zwischenwerts einen zusätzlichen Verfahrensschritt bedeutet. Umgehen läßt sich dieser Schritt dadurch, daß nicht grundsätzlich die Differenz zwischen n und dem Intervall b e g i n n gebildet wird (unabhängig von der Position des n im Intervall). Aus den Symmetrieeigenscheften der Funktion NA(n) und der Transformation nach Formel (12) ergibt sich, daß der Abstand von n zur oberen Intervallgrenze für ein n aus der zweiten Intervallhälfte ebensogroß ist wie der Abstand des transformierten Wertes von der unteren Intervallgrenze. Durch die Wahl der zur Differenzbildung herangezogenen Intervallgrenze entsprechend der Position von n im Intervall läßt sich die Transformation umgehen, wie die folgende kurze Überlegung zeigt.

Aus Abbildung 2 wird deutlich, daß die Länge eines beliebigen Intervalls der Funktion NA(n) ebensogroß ist, wie der Abstand vom Nullpunkt bis zum Beginn dieses Intervalls. Folglich ist der Abstand von der Intervallmitte zum Nullpunkt

dreimal so lang wie derjenige von der Intervallmitte zur oberen Grenze des Intervalls. Multipliziert man n mit dem Faktor $\frac{4}{3}$, so liegt der erhaltene Wert daher genau dann noch im ursprünglichen Intervall, wenn n in der ersten Hälfte des Intervalls lag; für n aus der zweiten Hälfte des Intervalls liegt er im nächsthöheren Intervall. Die zur Differenzbildung heranzuziehende Intervallgrenze G erhält man dadurch sinnvollerweise als

$$G = 2^{[\mathrm{ld}\ (\frac{4}{3}\ n)]} \qquad (13)$$

Für $n = 43$ erhält man daher beispielsweise als zu berücksichtigende Grenze

$$\begin{aligned} G = \ & 2^{[\mathrm{ld}\ (\frac{4}{3} \cdot 43)]} \\ = \ & 2^{[\mathrm{ld}\ 57,33]} \\ = \ & 2^{[5,8329]} \\ = \ & 2^5 \\ = \ & 32 \end{aligned}$$

Für $n = 53$ ergibt sich dagegen:

$$\begin{aligned} G = \ & 2^{[\mathrm{ld}\ (\frac{4}{3} \cdot 53)]} \\ = \ & 2^{[\mathrm{ld}\ 70,67]} \\ = \ & 2^{[6,1430]} \\ = \ & 2^6 \\ = \ & 64 \end{aligned}$$

Da n ober- oder unterhalb der errechneten Intervallgrenze liegen kann, ist der absolute Wert der Differenz zu bilden. Bezeichnet man die beim Durchlaufen der Intervalle angenommenen Zwischenwerte der n mit Indizes i (wobei sich entsprechend der Durchlaufrichtung der Intervalle eine fallende Indexvergabe als zweckmäßig erwiesen hat), so erhält man für die einzelnen Differenzen k_i den Ausdruck

$$k_i = \left| 2^{[ld\ (\frac{4}{3} \cdot n_i)]} - n_i \right| \qquad (14)$$

Der auf einen beliebigen Zwischenwert n_i folgende Zwischenwert im nächstniedrigeren Intervall, also n_{i-1}, muß mit n_i positionell übereinstimmen, d.h. er muß den gleichen Abstand von der Untergrenze des niedrigeren Intervalls besitzen wie n_i von der Untergrenze des höheren (bzw. von der Obergrenze im Fall von n_i aus der zweiten Hälfte des Intervalls). Damit erhält man für n_{i-1} allgemein:

$$n_{i-1} = 2^{i-1} + \left| 2^{[ld\ (\frac{4}{3} \cdot n_i)]} - n_i \right| \qquad (15)$$

Der Wert von NA(n) ergibt sich, wie oben ausgeführt, als Summe der Differenzbeträge zwischen den durchlaufenen Zwischenwerten n_i und den jeweils zu berücksichtigenden Intervallgrenzen. Die Anzahl der Summenglieder entspricht der Anzahl der durchlaufenen Intervalle, so daß n als Ausgangswert mit dem höchsten Index i eingesetzt wird als

$$n = n_{[ld\ n]} \qquad (16)$$

Aus den Formeln (14), (15) und (16) ergibt sich somit allgemein für NA(n):

$$NA(n) = \sum_{i=[ldn]}^{1} \left| 2^{[ld\ (\frac{4}{3} \cdot n_i)]} - n_i \right| \qquad (17)$$

mit $n_{[ld\ n]} = n$

und $n_{i-1} = 2^{i-1} + \left| 2^{[ld\ (\frac{4}{3} \cdot n_i)]} - n_i \right|$

Verwendet man anstelle der ausgeschriebenen Differenzbeträge die schon in Formel (14) eingeführte Bezeichnung k_i, so erhält man vereinfacht aus Formel (17):

$$NA(n) = \sum_{i=[ldn]}^{1} k_i \qquad (18)$$

mit $n_{[ld\ n]} = n$

und $n_{i-1} = 2^{i-1} + k_i$

Für das Beispiel NA(43) erhalten wir damit:

$[ld\ 43] = 5$

$$NA(43) = \sum_{i=5}^{1} \left| 2^{[ld\ (\frac{4}{3} \cdot n_i)]} - n_i \right|$$

$n_5 = n = 43$; $k_5 = \left| 2^{[ld\ (\frac{4}{3} \cdot 43)]} - 43 \right|$

$= \left| 2^{[ld\ 57{,}33]} - 43 \right|$

$= \left| 2^5 - 43 \right|$

$= \left| 32 - 43 \right|$

$= 11$

$n_4 = 2^4 + 11 = 27$; $k_4 = \left| 2^{[ld\ (\frac{4}{3} \cdot 27)]} - 27 \right|$

$= \left| 2^{[ld\ 36]} - 27 \right|$

$= \left| 2^5 - 27 \right|$

$= 5$

$n_3 = 2^3 + 5 = 13$; $k_3 = \left| 2^{[ld\ (\frac{4}{3} \cdot 13)]} - 13 \right|$

$= \left| 2^{[ld\ 17{,}33]} - 13 \right|$

$= \left| 2^4 - 13 \right|$

$= 3$

$$n_2 = 2^2 + 3 = 7 \quad ; \quad k_2 = \left| 2^{[1d\ (\frac{4}{3} \cdot 7)]} - 7 \right|$$
$$= \left| 2^{[1d\ 9{,}33]} - 7 \right|$$
$$= \left| 2^3 - 7 \right|$$
$$= 1$$

$$n_1 = 2^1 + 1 = 3 \quad ; \quad k_1 = \left| 2^{[1d\ (\frac{4}{3} \cdot 3)]} - 3 \right|$$
$$= \left| 2^{[1d\ 4]} - 3 \right|$$
$$= \left| 2^2 - 3 \right|$$
$$= 1$$

Damit ergibt sich NA(43) als

$$NA(43) = \sum_{i=5}^{1} k_i$$
$$= 11 + 5 + 3 + 1 + 1$$
$$= 21$$

Die Durchführung dieses Beispiels zeigt, daß das zweite Verfahren insbesondere durch die exakte Ermittlung der Logarithmen in der Anwendung etwas langwierig ist. In der Praxis ergibt sich, sofern man ohne technische Hilfsmittel nach diesem Verfahren arbeitet, dadurch eine Vereinfachung, daß man den zur Errechnung der zu berücksichtigenden Intervallgrenze erforderlichen Ausdruck

$$2^{[1d\ (\frac{4}{3} \cdot n_i)]}$$

ersetzt durch die Abschätzung, ob n_i näher an der unteren oder an der oberen Intervallgrenze liegt. Entsprechend bestimmt man dann k_i als

$$k_i = \left| 2^i - n_i \right|$$

bzw.

$$k_i = \left| 2^{i+1} - n_i \right| .$$

Im folgenden Beispiel wird bei der Berechnung des Wertes von NA(432) entsprechend verfahren.

$[1d\ n] = [1d\ 432] = 8$

$n_8 = n = 432 \quad ; \quad k_8 = \left| 2^9 - 432 \right| = \left| 512 - 432 \right| = 80$

$n_7 = 2^7 + 80 = 208 \quad ; \quad k_7 = \left| 2^8 - 208 \right| = \left| 256 - 208 \right| = 48$

$n_6 = 2^6 + 48 = 112 \quad ; \quad k_6 = \left| 2^7 - 112 \right| = \left| 128 - 112 \right| = 16$

$n_5 = 2^5 + 16 = 48 \quad ; \quad k_5 = \left| 2^6 - 48 \right| = \left| 64 - 48 \right| = 16$

$n_4 = 2^4 + 16 = 32 \quad ; \quad k_4 = \left| 2^5 - 32 \right| = \left| 32 - 32 \right| = 0$

$n_3 = 2^3 + 0 = 8 \quad ; \quad k_3 = \left| 2^3 - 8 \right| = \left| 8 - 8 \right| = 0$

$n_2 = 2^2 + 0 = 4 \quad ; \quad k_2 = \left| 2^2 - 4 \right| = \left| 4 - 4 \right| = 0$

$n_1 = 2^1 + 0 = 2 \quad ; \quad k_1 = \left| 2^1 - 2 \right| = \left| 2 - 2 \right| = 0$

$$NA(432) = \sum_{i=8}^{1} k_i$$
$$= 80 + 48 + 16 + 16 + 0 + 0 + 0 + 0$$
$$= 160$$

Aus diesem Beispiel wird ersichtlich, daß die Iteration abgebrochen werden kann, sobald sich erstmals für ein k_i der Wert Null ergibt. Alle folgenden Werte k_{i-r} sind dann ebenfalls gleich Null, so daß für die Berechnung von NA(n) die Summierung der bis dahin errechneten Differenzen ausreicht.

3. Wie die ausführlich dargestellten Beispiele zur Berechnung von NA(n) mit Hilfe des ersten und zweiten Verfahrens zeigen, lassen sich auch die Werte für größere n mit relativ geringem Aufwand bestimmen, wobei sowohl die Zahl der erforderlichen Verfahrensschritte als auch der mathematische Aufwand vergleichsweise gering sind. Beide Verfahren sind jedoch nicht in erster Linie für die Anwendung ohne technische Hilfsmittel konzipiert. Eine wesentliche Vereinfachung der Berechnung von NA(n) "in Handarbeit" bietet erst das im folgenden Abschnitt beschriebene dritte Verfahren. Die Verfahren 1 und 2

lassen sich sehr einfach programmieren; ihre Einführung war erforderlich, da die in ihnen zum Ausdruck kommenden mathematischen Gesetzmäßigkeiten die Voraussetzung für die Erstellung des dritten Verfahrens bilden.

Das hier darzustellende Verfahren basiert ebenfalls auf der Addition der Differenzbeträge k_i, berechnet diese jedoch nicht in der durch Formel (14) dargestellten Weise. Es sei ein beliebiges k_i gegeben. Nach der Transformation aus Formel (12), die direkt zur Definition von k_i in Formel (14) führte, kann kein k_i größer sein als die halbe Länge des Intervalls, in dem sich das zugehörige n_i befindet; diese halbe Intervallänge entspricht der Länge des Vorgängerintervalls. War im zweiten Verfahren ein n_i "zu groß", wurde es, wenn auch nur implizit, durch die Transformation in Formel (12) an der Intervallmitte gespiegelt, um von diesem transformierten Zwischenwert aus zu einem positionell entsprechenden Vorgänger zu gelangen. Die Art der Transformation war für alle n_i gleich; daraus folgt, daß jedes n_i innerhalb seines Intervalls nur zwei Positionen innehaben kann. Die Position des n_i innerhalb seines Intervalls wird durch die Größe des zuvor errechneten Zwischenwerts bestimmt, der seinerseits zwei Positionen innerhalb seines Intervalls einnehmen kann usw. Der einzig feststehende Wert innerhalb dieser Kette ist die Anzahl der Endknoten des zugrundeliegenden binären Stammbaums, also der Eingangswert n. Mit Hilfe von n und dem oben erwähnten Prüfkriterium, ob nämlich der folgende k_i - Wert größer ist als die halbe Intervallänge, läßt sich sehr leicht die Folge der k_i konstruieren. Die halbe Intervallänge als Prüfkriterium dient dabei gegebenenfalls auch zur Korrektur eines Wertes.

Zunächst ist festzustellen, wie weit n vom Intervallbeginn entfernt ist. Es wird also die Differenz k' gebildet als

$$k' = 2^{[ld\ n]} - n \qquad (19)$$

Nun ist zu prüfen, ob k' entsprechend den oben genannten Bedingungen kleiner oder gleich der halben Intervallänge ist, ob

also gilt:

$$k' \leq 2^{[ld\ n]} - 1 \qquad (20)$$

Ist dies der Fall, wird k' als endgültig angesehen und als k bzw. k_i mit dem fallenden Index i entsprechend dem zweiten Verfahren versehen weiterverwendet. Ist k' größer als die halbe Intervallänge, wird es in Anlehnung an Formel (12) transformiert, indem man k' von der halben Intervallänge subtrahiert. Das Resultat k ist in jedem Fall positiv, da k' trivialerweise kleiner ist als die Länge des Intervalls, in dem sich n befindet. Anschaulich entspricht dieser Subtraktion die Spiegelung an der Intervallmitte, mathematisch der im zweiten Verfahren angewandte Übergang von der unteren zur oberen Intervallgrenze. Als vorläufige Differenz k'_{i-1} im nächstniedrigeren Intervall wird nun einfach zunächst k_i angenommen. Auch dieses Vorgehen entspricht dem Rechengang des zweiten Verfahrens; aus der Forderung nach positioneller Übereinstimmung zwischen den unmittelbar aufeinander folgenden n_i ergab sich die Festlegung in Formel (15), später übernommen in den Formeln (17) und (18). Nunmehr wird dasselbe Prüfkriterium wie oben herangezogen, mit dessen Hilfe festgestellt wird, ob k'_{i-1} kleiner oder gleich 2^{i-1} ist. Gegebenenfalls wird k'_{i-1} von 2^{i-1} subtrahiert usw. Dieser Vorgang wird so lange fortgesetzt, bis als letzte Differenz k_1 errechnet ist. Das dritte Verfahren läßt sich also formal folgendermaßen zusammenfassen:

$$NA(n) = \sum_{i=[ldn]}^{i} k_i$$

wobei

$$k_i = \begin{cases} k'_i & \text{für } k'_i \leq 2^{i-1} \\ 2^i - k'_i & \text{sonst} \end{cases}$$

mit $k'_{[ldn]} = n - 2^{[ldn]}$

und $k'_{i-1} = k_i$ (21)

NA(n) als Summe der k_i wird also ebenso definiert wie im zweiten Verfahren. Der Unterschied zwischen beiden Verfahren liegt darin, daß nunmehr die k_i und damit NA(n) ohne Berechnung der Zwischenwerte n_i bestimmt werden. Hierdurch wird eine wesentliche Vereinfachung und Abkürzung des Rechenganges erreicht. Für die Ermittlung von NA(n) mit Hilfe dieses Verfahrens ohne Verwendung technischer Hilfsmittel hat sich das folgende Schema als zweckmäßig erwiesen:

Man unterteilt zunächst das Rechenfeld kreuzförmig. In die linke obere Ecke des Feldes wird der Wert für n eingetragen. Auf gleicher Höhe wie n, jedoch rechts von der senkrechten Linie, wird die nächstniedrigere Zweierpotenz mit ganzzahligem Exponenten notiert, also $2^{[ldn]}$. Darunter, nunmehr unterhalb der waagerechten Linie, werden alle weiteren Zweierpotenzen mit ganzzahligen Exponenten in fallender Folge bis hin zu $2^0 = 1$ eingetragen. Für das Beispiel $n = 43$ ergibt sich bis hierhin das folgende Bild:

43	32
	16
	8
	4
	2
	1

Als nächstes wird die Differenz $n - 2^{[ldn]}$ gebildet (in unserem Beispiel 43 - 32) und das Ergebnis rechts neben $2^{[ldn]}$, hier 32, notiert, also noch oberhalb der waagerechten Linie. Diese Differenz, in unserem Beispiel 11 als k'_5 (da [ld 43] = 5 ist), ist die einzige aller vorläufigen Differenzen k'_i, die während des gesamten Rechenvorgangs überhaupt festgehalten wird. Nun wird weiter ganz schematisch nach der Prüfbedingung aus Formel (21) verfahren: Ist k'_i kleiner oder gleich der nächstniedrgeren Zweierpotenz, die jeweils links in der folgenden Zeile notiert ist, so wird es unverändert rechts in der nächsten Zeile

eingetragen; ist k'_i größer als 2^{i-1}, so wird die Differenz $2^i - k'_i$ gebildet und das Resultat rechts in der nächsten Zeile notiert. So wird weiter verfahren, bis neben allen vorher eingetragenen Zweierpotenzen kleinere oder gleich große k_i - Werte stehen. Die Addition aller dieser k_i - Werte, d.h. ohne den vorläufigen Wert $k'_{[ldn]}$ rechts in der Kopfzeile, ergibt NA(n).

Das Beispiel zur Berechnung von NA(43) soll ausführlich in allen Schritten dargestellt werden; als Hilfe für die Erläuterungen werden zusätzlich die einzelnen Zeilen numeriert.

Zeile				
1	43	32	11	$= k'_5$
2		16	11	$= k_5$
3		8	5	$= k_4$
4		4	3	$= k_3$
5		2	1	$= k_2$
6		1	1	$= k_1$
7			21 = NA(43)	

Wie oben beschrieben, werden zunächst in das vorbereitete Schema die Werte für n (= 43), $2^{[ldn]}$ (= 32) und die kleineren Zweierpotenzen eingetragen. Dann wird die Differenz 43 - 32 gebildet und das Resultat, 11, als k'_5 rechts von 32 notiert. 11 ist kleiner als 16 (die Zweierpotenz in Zeile 2) und wird daher (als k_5) rechts neben die 16 geschrieben. Nun wird 11 mit der nächstkleineren Zweierpotenz verglichen (also mit 8 in Zeile 3). 11 ist größer als 8 und wird daher von 16 subtrahiert. Das Resultat, 5, wird (als k_4) neben die 8 geschrieben. 5 wird nun mit der nächstkleineren Zweierpotenz, also 4, verglichen. Da 5 größer als 4 ist, wird die Differenz 8 - 5 gebildet und das Resultat, 3, als k_3 neben die 4 geschrieben. Da weiterhin 3 größer ist als 2, wird die Differenz 4 - 3 gebildet und das Resultat, 1, als k_2 neben die 2 in Zeile 5 geschrieben. 1 ist

gleich groß wie die nächste Zweierpotenz, also 1 in Zeile 6, und wird daher als k_1 dorthin übernommen. Zum Abschluß werden nun die Werte von k_5 bis k_1 addiert, also die rechts notierten Zahlen unterhalb der waagerechten Linie. Das Ergebnis (11 + + 5 + 3 + 1 + 1 = 21) bildet, in Zeile 7, den gesuchten Wert für NA(43).

Einen Fall, in dem $k_{[ldn]}$ bereits geändert werden muß, bietet, wie alle n aus der zweiten Hälfte eines Intervalls, das Beispiel n = 53. Die Differenz 53 - 32 = 21 in der Kopfzeile muß, da 21 größer ist als 16 (die Zweierpotenz in der nächsten Zeile), von 32 subtrahiert werden, so daß sich k_5 in der ersten Zeile unterhalb der waagerechten Linie ergibt als 32 - 21 = 11. Da auch alle weiteren k_1 - Werte mit denen des vorigen Beispiels übereinstimmen, ergibt sich als Resultat für NA(53) ebenfalls der Wert 21:

53	32	21
	16	11
	8	5
	4	3
	2	1
	1	1
		21 = NA(53)

Mit Hilfe des hier präsentierten dritten Verfahrens lassen sich, wie auch die beiden weiteren Beispiele für n = 432 und n = 4321 zeigen, rasch und mit minimalem rechnerischem Aufwand die NA(n) - Werte auch für sehr große binäre Stammbäume ermitteln. Mit den Zweierpotenzen bilden auch die k_1 eine monoton fallende Folge; man kann daher hinreichend kleine k_1 - Werte so lange ständig wiederholt untereinander schreiben, bis links in der Folgezeile eine kleinere Zweierpotenz erscheint (vgl. den k_1 - Wert 225 im Beispiel NA(4321)). Dies bedeutet auch, daß die Errechnung weiterer Folgewerte unterbleiben kann, sobald als Differenz k_i die Werte 0 oder 1 erscheinen. Diese können

dann, wie in den beiden folgenden Beispielen ersichtlich, bis zur letzten Zeile übertragen werden.

432	256	176	
			$176 = k'_8 > 128 \Rightarrow k_8 = 256 - 176 = 80$
	128	80	$80 = k_8 > 64 \Rightarrow k_7 = 128 - 80 = 48$
	64	48	$48 = k_7 > 32 \Rightarrow k_6 = 64 - 48 = 16$
	32	16	$16 = k_6 \leq 16 \Rightarrow k_5 = k_6 = 16$
	16	16	$16 = k_5 > 8 \Rightarrow k_4 = 16 - 16 = 0$
	8	0	$0 = k_4 \leq 4,2,1 \Rightarrow k_3, k_2, k_1 = k_4 = 0$
	4	0	
	2	0	
	1	0	
			160 = NA(432)

4321	4096	225
	2048	225
	1024	225
	512	225
	256	225
	128	31
	64	31
	32	31
	16	1
	8	1
	4	1
	2	1
	1	1
		998 = NA(4321)

LITERATUR

Altmann, G., Lehfeldt, W.
1980 Einführung in die Quantitative Phonologie. Bochum. (Quantitative Linguistics. 7.)

Sander, H-D., Altmann, G.
1973 Asymmetrie binärer Stammbäume. In: Phonetica 28, 171-181.

Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen

U. Rothe, Bochum

1. In der vorliegenden Arbeit soll - speziell für die romanischen Sprachen - die Gültigkeit des Menzerathschen Gesetzes unter dem Aspekt der Bezugsetzung der Bedeutungsmenge eines Wortes zu der Wortlänge untersucht werden.

In der von Menzerath (1954:101) generell formulierten Hypothese 'Je größer das Ganze desto kleiner die Teile' wird die Länge des sprachlichen Konstrukts (z. B. des Wortes) in Bezug gesetzt zu den es konstituierenden Teilen (z. B. den Silben). Diese Beziehung gilt für verschiedene sprachliche Ebenen und Bereiche: Die Länge eines Elements ist eine Funktion der Länge seiner hierarchisch nächsthöheren Einheit; aber auch die Menge der semantischen Repräsentanten kann als eine Funktion der Länge des Ausdrucks angegeben werden (vgl. Altmann, Beöthy, Best 1982).

Letztere der Beziehungen anhand von drei romanischen Sprachen (Französisch, Portugiesisch, Spanisch) zu überprüfen, ist Ziel der folgenden Arbeit. Beobachtungen, die sich aufgrund der erzielten Ergebnisse machen lassen, können eventuell zu Rückschlüssen auf die morphologisch-semantische Lexemstruktur der untersuchten Sprachen im einzelnen sowie generell führen.

2. Der Lösungsweg geht über einen von Altmann (1980) vorgeschlagenen Ansatz zum Nachweis der Annahme, daß die Größe der Konstituente (y) eine monoton fallende Funktion der Konstruktgröße (x) ist.

In dem ursprünglichen Ansatz wurden von Altmann zwei Konstanten angesetzt. Eine Reihe von Untersuchungen aus verschiedenen Bereichen zeigt aber, daß die Differentialgleichung

$$\frac{dy}{y} = \frac{bdx}{x}$$

die untersuchten Trends hinreichend erfaßt. Die empirischen Kurven verlaufen so, daß ein ausgleichender Störungsparameter überflüssig erscheint (vgl. Altmann, Beöthy, Best 1982, Köhler 1982, Gerlach 1982, Heups in diesem Band).

Um eine Funktionsgleichung zu erhalten, wird die Differentialgleichung aufgelöst in:

$$y = ax^b$$

Wie die Erfahrung gezeigt hat, gibt der Parameter a dieser Kurve die theoretische mittlere Zahl der Bedeutungen (y_t) für die Wörter mit der Länge 1 an, wobei ein Wort mit der Länge 1 entweder ein einbuchstabiges oder ein einsilbiges Wort ist.

Zur Schätzung der Koeffizienten a und b wurde mithilfe einer logarithmischen Transformation eine Linearisierung vorgenommen:

$$y = ax^b \Longrightarrow \ln y = \ln a + b \cdot \ln x$$

Die einzelnen Koeffizienten wurden mit der Methode der kleinsten Quadrate geschätzt, indem mit einer partiellen Ableitung das Minimum von

$$W = \sum_i n_i \, (\ln y_i - \ln a - b \cdot \ln x_i)^2$$

gesucht wurde. Bei Verwendung der üblichen Methoden der (gewichteten) linearen Regression ergeben sich die unter den jeweiligen Tabellen (§ 3.) angegebenen Werte.

Für die Berechnungen wurde pro Sprache mittels einer systematischen Lexikonzählung eine Stichprobe von 1000 Wörtern erstellt. Es wurden die Wortlängen - gemessem einmal in der Zahl der Buchstaben, einmal in der Zahl der Silben - ihren jeweiligen Bedeutungsmengen gegenübergestellt. Gezählt wurde das letzte Wort jeder oder jeder zweiten Seite, je nach Umfang des benutzten Wörterbuchs. Die verwendeten Lexika sind in der Literaturliste aufgeführt. Wörter, die an sich keine Bedeutung haben, wie Eigennamen und Wortbildungselemente, wurden nicht in die Stichprobe mit einbezogen. Traf man jedoch auf ein solches Lexem oder enthielt das Stichwort lediglich einen Verweis auf andere Stellen, so wurde das davor befindliche Wort genommen. Ergab dies wiederum Probleme, so wurde das erste Wort der folgenden Seite gewählt. Zahlenmäßig unterbelegte Wortlängenklassen wurden von der Berechnung ausgeschlossen, da sie keine repräsentativen Werte liefern.

3. Das Ergebnis der Überprüfung ist in Form von Tabellen und anhand von Graphiken dargestellt.

Die verwendeten Notationen bedeuten:

x_i Zahl der Buchstaben bzw. Silben im Wort
y_i mittlere Zahl der Bedeutungen
n_i Zahl der Wörter mit der Länge x_i
y_t berechnete Werte (hier die theoretische Bedeutungszahl)
F empirischer Wert des F-Kriteriums mit den angegebenen Freiheitsgraden
P die Wahrscheinlichkeit, mit der man einen so hohen oder noch extremeren F-Wert theoretisch erwarten kann.

Das Signifikanzniveau wurde auf $\alpha = 0{,}05$ festgesetzt.

3. 1. Die für das Französische sich ergebenden Daten und Werte sind in Tabelle 1 bzw. 2 sowie in Abbildung 1 bzw. 2 dargestellt.

Tabelle 1. Mittlere Bedeutungszahl nach Buchstabenlänge.

x_i	n_i	y_i	y_t
2	5	4.4000	10.3820
3	14	5-8571	6.7863
4	56	4.8036	5.0190
5	87	4.3333	3.9719
6	128	3.4453	3.2807
7	146	3.1986	2.7911
8	142	2.1408	2.4264
9	120	2.2250	2.1445
10	98	1.6735	1.9209
11	65	1.6923	1.7375
12	50	1.7200	1.5860
13	33	1.3939	1.4583
14	20	1.5500	1.3493
15	11	1.0909	1.2551
16	7	1.1429	1.1730

Die Kurve lautet:

$$y = 21.4755x^{-1.048608};\ F_{1,13} = 135.48;\ P = 3 \cdot 10^{-8}\ .$$

Wie ersichtlich ist, liefert die Kurve für die mittlere Bedeutungszahl nach Buchstabenlänge im Französischen ein hoch signifikantes Resultat.

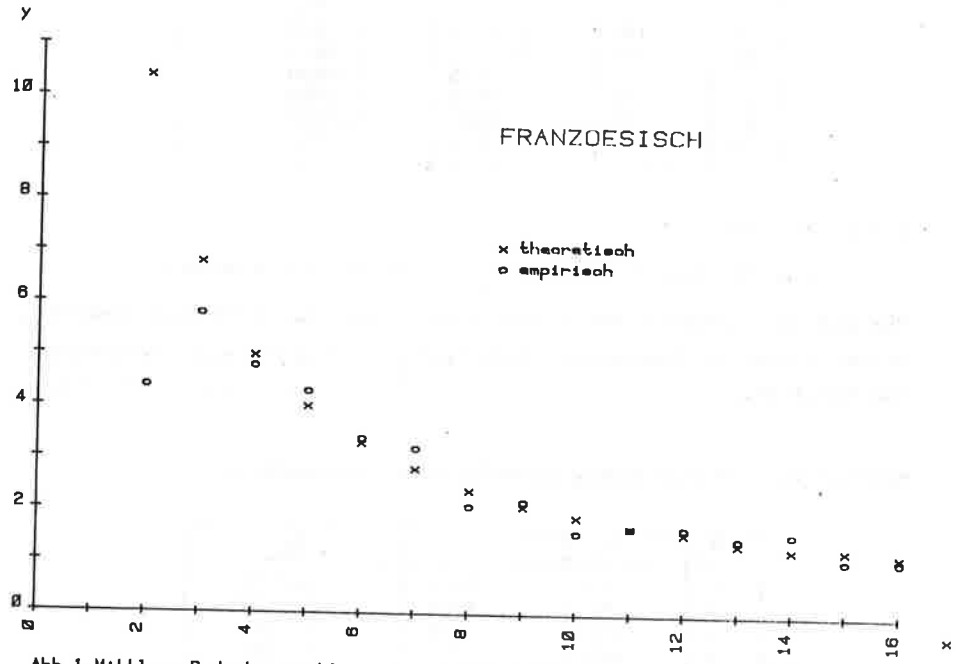


Abb. 1 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Buchstaben

Tabelle 2. Mittlere Bedeutungszahl nach Silbenlänge.

x_i	n_i	y_i	y_t
1	141	4.8369	5.1956
2	344	3.1715	2.9208
3	309	1.9709	2.0854
4	135	1.5852	1.6420
5	50	1.5200	1.3641
6	9	1.0000	1.1723

Die Kurve lautet:

$$y = 5.1956x^{-0.830925};\ F_{1,4} = 112.56;\ P = 0.0004$$

Auch für die mittlere Bedeutungszahl nach Silbenlänge ergibt sich für das Französische ein hoch signifikantes Resultat.

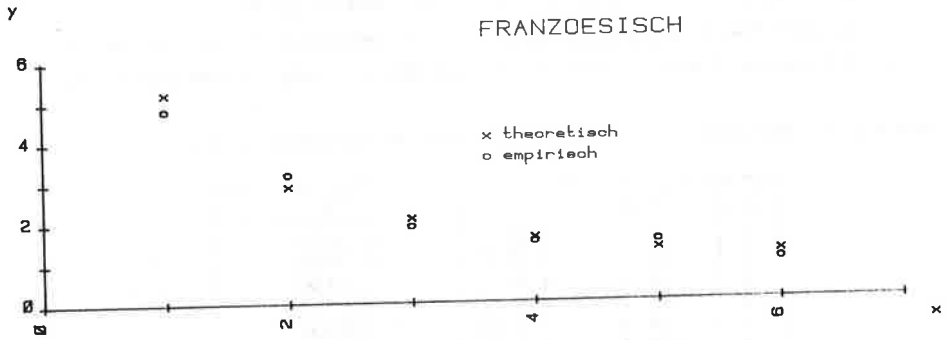


Abb. 2 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Silben

3. 2. Die für das Portugiesische ermittelten Werte sind in Tabelle 3 bzw. 4 und in Abbildung 3 bzw. 4 wiedergegeben.

Tabelle 3. Mittlere Bedeutungszahl nach Buchstabenlänge.

x_i	n_i	y_i	y_t
3	13	6.7792	8.3967
4	50	4.2600	6.0842
5	85	3.9412	4.7390
6	123	3.7805	3.8639
7	142	3.4269	3.2513
8	167	3.1677	2.7997
9	142	2.1972	2.4538
10	110	2.1545	2.1807
11	63	2.0635	1.9600
12	55	1.5818	1.7780
13	20	1.4000	1.6256
14	14	1.5714	1.5961

Die Kurve lautet:

$$y = 28.733x^{-1.119780};\ F_{1,10} = 77.14;\ P = 0.000005$$

Für die Abhängigkeit der Bedeutungszahl von der Wortlänge gemessen in Buchstaben im Portugiesischen liefert die Kurve ein signifikantes Resultat.

Tabelle 4. Mittlere Bedeutungszahl nach Silbenlänge.

x_i	n_i	y_i	y_t
1	15	5.5333	11.2672
2	191	5.4503	5.0194
3	339	3.1976	3.1278
4	308	2.1981	2.2361
5	126	1.6349	1.7236
6	25	1.3600	1.3934

Bei den in Silbenzahl gemessenen Wortlängen des Portugiesischen ergab sich die obigen Werten zugrundeliegende Kurve

$$y = 11.2672x^{-1.166537};\ F_{1,4} = 65.34;\ P = 0.0013$$

Die Signifikanz bei dieser Kurve ist zwar nicht so hoch wie bei den Kurven der anderen Errechnungen; dennoch ist das Ergebnis zufriedenstellend.

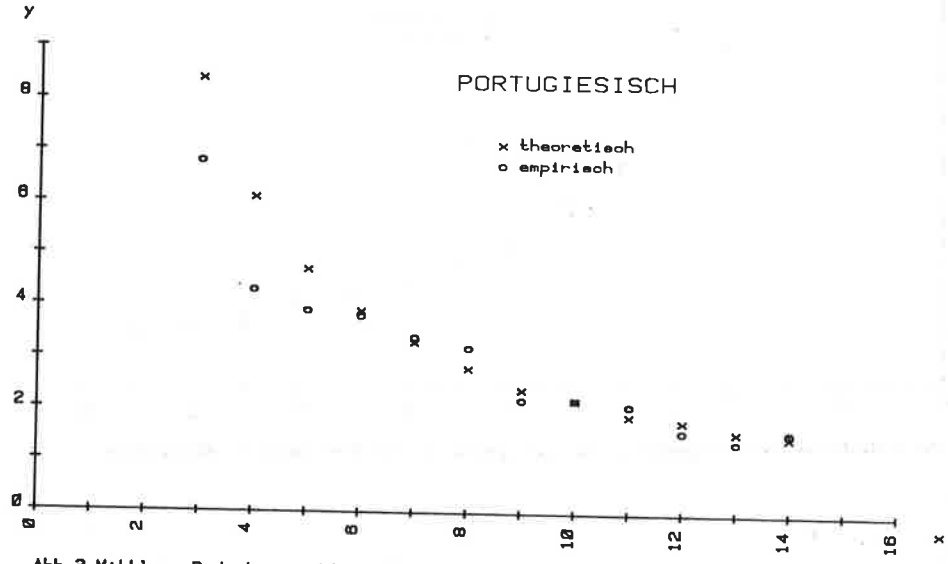


Abb. 3 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Buchstaben

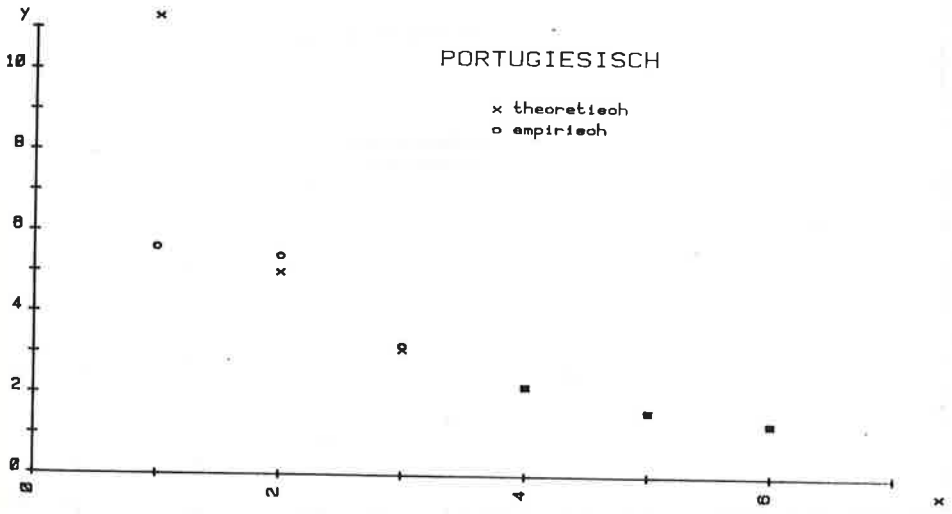


Abb. 4 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Silben

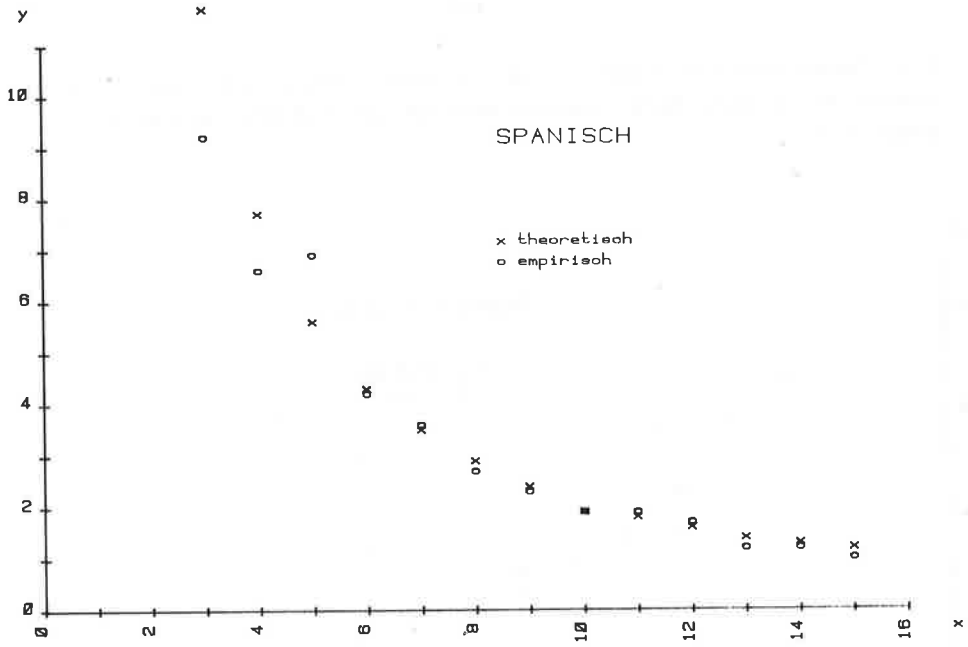


Abb. 5 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Buchstaben

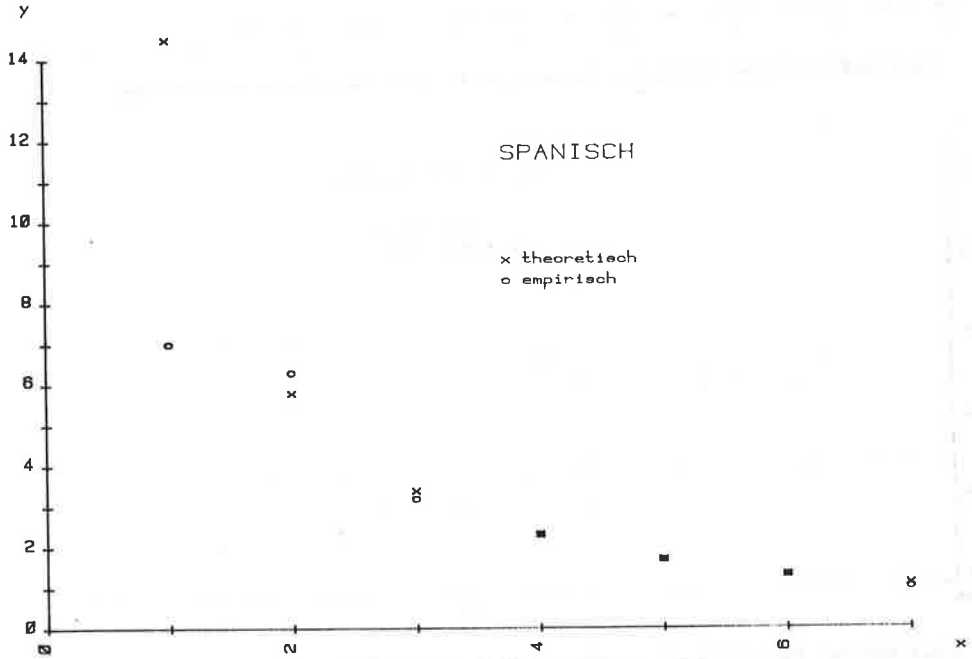


Abb. 6 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Silben

3. 3. Für das Spanische schließlich ergaben sich die aus Tabelle 5 bzw. 6 und aus Abbildung 5 bzw. 6 ersichtlichen Ergebnisse:

Tabelle 5. Mittlere Bedeutungszahl nach Buchstabenlänge.

x_i	n_i	y_i	y_t
3	13	9.1538	11.7244
4	54	6.6111	7.7366
5	101	6.9614	5.6041
6	110	4.1636	4.3061
7	172	3.5174	3.4462
8	149	2.7047	2.8415
9	141	2.3617	2.3968
10	95	1.9368	2.0583
11	87	1.9655	1.7935
12	37	1.7027	1.5816
13	18	1.2222	1.4088
14	10	1.2000	1.2657
15	7	1.0000	1.1456

Es ergibt sich die Kurve:

$$y = 57.3530x^{-1.445052};\ F_{1,11} = 270{,}03;\ P = 4 \cdot 10^{-9}$$

Diese Kurve für die Abhängigkeit der Bedeutungszahl von der in Buchstaben gemessenen Wortlänge des Spanischen weist sich durch eine sehr hohe Signifikanz aus.

Tabelle 6. Mittlere Bedeutungszahl nach Silbenlänge.

x_i	n_i	y_i	y_t
1	10	7.0000	14.4673
2	218	6.2615	5.7407
3	326	3.1840	3.3431
4	283	2.3463	2.2780
5	121	1.6694	1.6917
6	25	1.2400	1.3266
7	11	1.0000	1.0801

Die Kurve lautet:

$$y = 14.4673x^{-1.333490};\ F_{1,5} = 128.73;\ P = 0.000093$$

Wie bei allen anderen fünf Überprüfungen ergibt sich auch hier wieder ein eindeutig signifikantes Resultat.

3. 4. Zum Vergleich werden auf der folgenden Seite (Abb. 7 und 8) die theoretischen Kurven noch einmal in einer Graphik gegenübergestellt, diesmal gemeinsam mit den theoretischen Werten des Deutschen, Slovakischen und Ungarischen aus der Untersuchung von Altmann, Beöthy, Best (1982).

Wie man sieht, besteht zwischen den romanischen und den anderen Sprachen ein Unterschied sowohl im Anstieg der Kurven (b) als auch in der Ordinate am Anfang der Kurve (a). Letzteres steht im Zusammenhang mit der durchschnittlichen Bedeutungszahl der kürzesten Wörter. Eine mögliche Erklärung für den unterschiedlichen Anstieg der Kurven könnte in unterschiedlich umfangreichen Wortschätzen bzw. unterschiedlich umfangreichen Wörterbüchern begründet sein. Auch vorhandene sprachenspezifische Abweichungen in den Sinnrelationen, wie Homonymie, Synonymie, Hyponymie, Markiertheit, etc., könnten für den unterschiedlichen Kurvenanstieg verantwortlich sein.

Die Tatsache aber, daß es sich hier um einen recht strengen Zusammenhang zwischen der Inhalts- und der Ausdrucksebene handelt, deutet darauf hin, daß eine Interpretation bzw. theoretische Ableitung des Parameters b mit ziemlich großen Schwierigkeiten verbunden ist. Wahrscheinlich spielen hier nicht nur lexikalische Eigenschaften eine wichtige Rolle.

Zusammenfassend läßt sich sagen: die erbrachten Nachweise für die Vermutung, daß das Menzerathsche Gesetz auch für die Beziehung zwischen Wortlänge und Bedeutungsmenge gilt, dürften ausreichen. Das Gesetz bleibt somit nicht auf die Ausdrucksseite der Sprache beschränkt.

Die nachgewiesene Beziehung ist durch eine monoton fallende Funktion charakterisiert, die - je nach Sprache - mehr oder weniger steil abfällt. Welche diversen Zusa-menhänge im einzelnen für den jeweiligen Kurvenverlauf verantwortlich sind, kann nur nach umfangreichen Forschungsarbeiten entschieden werden.

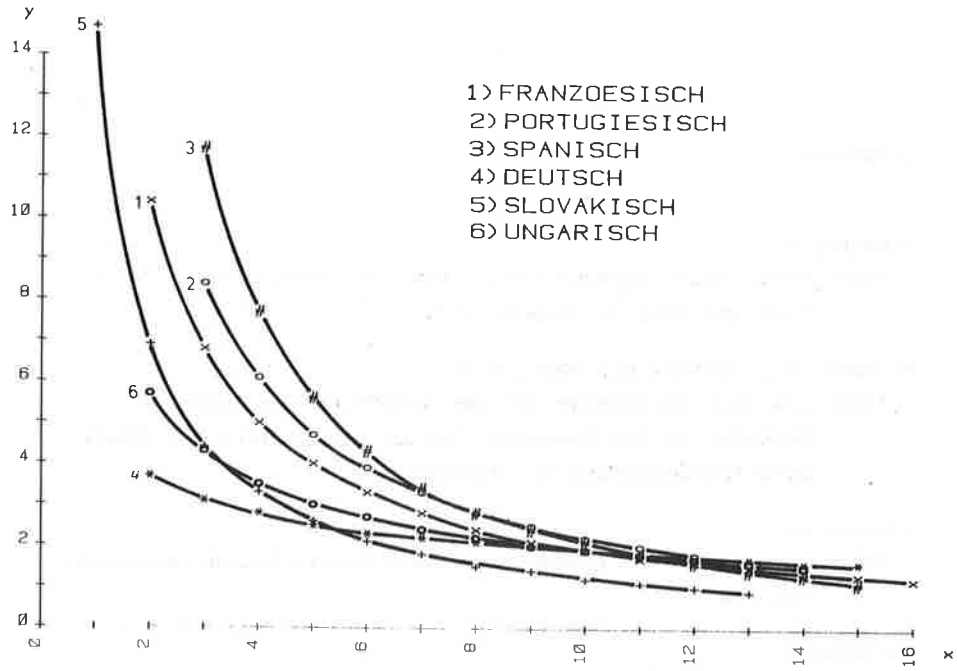


Abb. 7 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Buchstaben

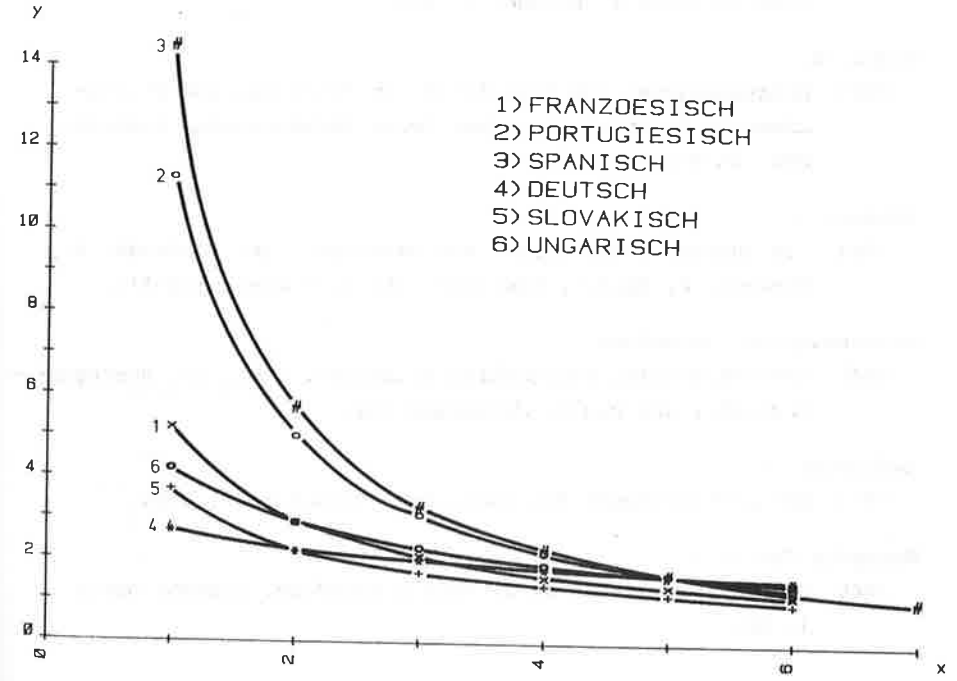


Abb. 8 Mittlere Bedeutungszahl in Abhaengigkeit von der Wortlaenge in Silben

LITERATUR

Altmann, G.
1980 Prolegomena to Menzerath's Law. In: Grotjahn, R. (Ed.), Glottometrika 2. Bochum, 1-10

Altmann, G., Beöthy, E., Best, K.-H.
1982 Die Bedeutungsmenge und das Menzerathsche Gesetz. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543.

Dubois, J.
1977 Larousse de la langue française. Lexis. Paris: Larousse, 16. Ed.

Gerlach, R.
1982 Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In: Lehfeldt, W., Strauss, U. (Eds.), Glottometrika 4. Bochum, 95-102.

Heups, G.
1983 Untersuchungen zum Verhältnis von Satzlänge und Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In diesem Band.

Köhler, R.
1982 Das Menzerathsche Gesetz auf Satzebene. In: Lehfeldt, W., Strauss, U. (Eds.), Glottometrika 4. Bochum, 103-113.

Melhoramentos, Brockhaus
1961 Novo Michaelis, Diccionário Ilustrado. (vol. II, Portuguese-English), São Paulo: Melhoramentos.

Menzerath, P.
1954 Die Architektonik des deutschen Wortschatzes. Bonn.

Menendez Pidal, R.
1965 Diccionario Durvan de la Lengua Española. Bilbao: Durvan, 1. Ed.

Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen

Gabriela Heups, Göttingen

1. Untersuchungsinteresse der vorliegenden Arbeit ist in Anlehnung an MENZERATHs Arbeiten aus den Jahren 1928 und 1954 und den daraus von ALTMANN abgeleiteten möglichen Hypothesen die Frage nach dem Verhältnis von Satzlänge zu Clauselänge in ausgewählten deutschen Texten verschiedener Textklassen. MENZERATH hat seine Ergebnisse zum Verhältnis von Laut, Silbe und Lautdauer in einem "Allgemeinen Quantitätsgesetz" zusammengefaßt; seine Beobachtungen sind wortstatistischer Natur. Die Resultate seiner Arbeiten zeigen Einzelaspekte über den Zusammenhang zweier Variablen - Silbe und Laut - innerhalb einer Sprache; sie demonstrieren das Verhältnis zwischen der Länge eines Sprachkonstrukts zu seinen Konstituenten. ALTMANN (1980 : 1) formuliert dies so: "The longer a language construct the shorter its components (constituents)".

Um sowohl MENZERATHs empirische Entdeckungen als auch ALTMANNs generelle Hypothesen theoretisch für gültig erklären zu können, erscheint es unumgänglich, möglichst viele ableitbare Hypothesen empirisch zu überprüfen, um der Behauptung: "je größer das Ganze, umso kleiner die Teile" (MENZERATH 1954 : 101) gesetzmäßigen Status zu verleihen.
Einige der von ALTMANN aufgestellten Hypothesen, z.B. die zum Verhältnis zwischen Phonemen und Morphemen (GERLACH 1981), sind bereits überprüft worden. Diese Arbeit hat ihre Anregung in ALTMANNs Hypothese XI (1980 : 9) gefunden, die lautet: "In long sentences (measured in number of clauses) the clauses are shorter and vice versa otherwise the sentence looses its clearness".

Zwei wesentliche Aspekte erscheinen hier interessant nachzuprüfen. Einmal, o b ein Verhältnis zwischen Satz- und Clauselänge tatsächlich besteht; zum anderen läßt die Hypothese die spekulative Überlegung zu, daß in längeren Sätzen die Clauses kürzer werden m ü s s e n , damit der Inhalt (Informationsgehalt) einer sprachlichen Äußerung für den Rezipienten überhaupt noch verständlich ist.

Wir werden hier also explizit auf das Verhältnis zwischen Satzlänge und Clauselänge als Untersuchungsgegenstand diese Arbeit eingehen und formulieren deshalb die folgende Hypothese:

"Je länger ein Satz, gemessen in der Anzahl der Clauses, desto kürzer die Clauses, gemessen in der Wortzahl".

2. Das Material beläuft sich auf 10.668 Sätze als Untersuchungseinheiten, die sich aus fünf Textklassen mit insgesamt 13 Einzeltexten rekrutierten (Tab. 1). Die ausgewählten Textklassen sind in Anlehnung an PIEPERs "Textgruppen" zum Teil übernommen worden (vgl. PIEPER 1979 : 45).
Die 13 bearbeiteten Einzeltexte sind den 5 Textklassen ohne weitere Differenzierung direkt zugeordnet worden. So finden wir den Romantext des 19. Jahrhunderts und den aus der Trivialliteratur der neuesten Zeit unmittelbar den Romantexten subsumiert, obgleich sich in den jeweiligen Einzeldaten erhebliche Unterschiede erkennen lassen (s. Anhang).

3. Die für unsere Untersuchung relevanten Beobachtungskriterien sind Satzlänge und Clauselänge. Die Bezugseinheiten, mit denen wir operieren, sind die Variablen Wort, finites Verb, Clause und Satz. Ein SATZ ist definiert als eine endliche Sequenz von Graphemen und Spatien, die zwischen zwei satzabschließenden Zeichen steht, und durch Großschreibung des ersten Graphems am Satzanfang gekennzeichnet ist (Definitionen in Anlehnung an WINTER 1974 und PIEPER 1979). Satzabschließende Zeichen sind: der Punkt (.) und das Fragezeichen (?) und das Ausrufungszeichen (!).

Tab. 1: Verteilung der untersuchten Sätze n = 10668 auf 13 Einzeltexte in 5 Textklassen (Einzelresultate s. Anhang).

Textklasse	Einzeltexte	Anzahl der Sätze	Sätze gesamt pro Textklasse
GESETZESTEXTE (juristische Publikationen)	GG	691	
	BVerfGG	354	1045
WISSENSCHAFTL. TEXTE (in wiss. Reihen publ.)	Philologie	1002	
	Chemie	518	
	BWL	516	2036
ZEITUNGSTEXTE (in öffent. Medien publ.)	Politik	1112	
	Feuilleton	1072	2184
BRIEFE (Gruß- u. Abschiedsformeln)	Leserbriefe	490	
	Romanbriefe	532	1022
ROMANTEXTE (als 'Roman' ersch.)	Roman 19. Jh.	1035	
	Roman 20. Jh.	1241	
	Roman 20. Jh.	1037	
	Roman Triviallit. 20. Jh.	1068	4381
Summe	13 Einzeltexte	10668	10668

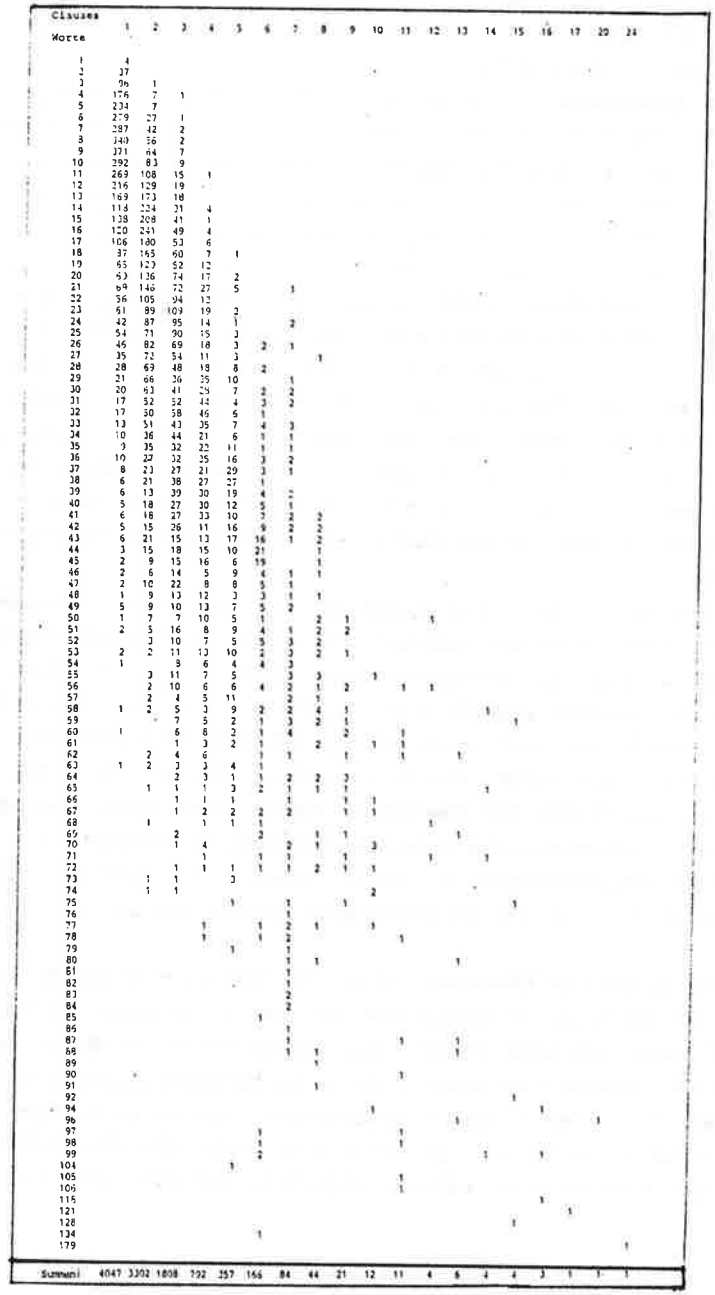
Steht ein Doppelpunkt (:), so gilt der Satz dann als abgeschlossen, wenn das erste Graphem des nächsten Wortes groß geschrieben ist.

Die SATZLÄNGE (SL) kann zum einen durch die den Satz konstituierende Anzahl der Wörter bestimmt werden, wobei das Wort als eine "in Morpheme gegliederte, relativ selbständige Graphemkette, die durch Lücken begrenzt - und in bestimmten Fällen geteilt - (verstanden) wird, als Träger einer einheitlichen grammatischen Bedeutung auftritt und als Lexem repräsentiert werden kann" (HOFFMANN 1976 : 261).

Man errechnet zunächst die Summe der in den zu untersuchenden Sätzen enthaltenen Wörter und dividiert dann diese Summe durch die Anzahl der ausgezählten Sätze. Dieser Quotient ist ein Durchschnitts- oder Mittelwert.
Zum anderen, und für unsere Analyse von besonderem Interesse, läßt sich die Satzlänge durch die Anzahl der Clauses festlegen. Hier werden allerdings keine Durchschnittswerte, sondern absolute Daten ermittelt. Mit Hilfe der Bestimmung der Clauses pro Satz ist es uns möglich, SATZTYPEN zu bilden und diese nach "Einclausern", "Zweiclausern" etc. zu klassifizieren.
CLAUSES (auch "Teilsätze") sind festgelegt durch die Anzahl der im Satz vorkommenden finiten Verben, wobei FINITE VERBEN durch ihre Konjugationsmorpheme gekennzeichnet sind (PIEPER 1979 : 24). Die CLAUSELÄNGE (CL) läßt sich aus der Anzahl der Wörter pro Satz, die sich um ein finites Verb gruppiert, berechnen, wobei die im Satz enthaltenen finiten Verben obligatorische Konstituenten der Clauses, also Indizes für die Clauselänge, sind. Der Quotient aus Wörtern pro Satz und finiten Verben pro Satz ist wiederum ein Durchschnittswert.

4. Nach Segmentierung und Auszählung des Datenmaterials erhielten wir ein Untersuchungskorpus von 10.668 Sätzen. Auf diese 10.668 untersuchten Einheiten verteilen sich insgesamt 23.745 Clauses und 233.931 Wörter. Die durchschnittliche SL besteht demnach aus ca. 2 Clauses, genauer: SL = 2.2258 oder 21 bis 22 Wörtern. Der Durchschnittswert für die CL liegt bei fast 10 Wörtern oder: CL = 9.8518 (vgl. Tab. 2).
Am häufigsten in unserer Untersuchung tritt der 1-clausige Satztyp auf; er umfaßt mit seiner absoluten Häufigkeit von 4047 Sätzen über ein Drittel (37.93 %) der Gesamtzahl. Ihm folgt in geringer Distanz der Zweiclauser; er erscheint 3302mal, nimmt also gut ein Drittel (30.95 %) aller Sätze in Anspruch. Der Dreiclauser kommt bereits weitaus weniger häufig vor; er umfaßt mit 1808 Sätzen nur noch etwa ein Sechstel der Gesamtzahl. An vierter Stelle steht der 4-clausige Satztyp mit ungefähr einem Dreizehntel bzw. 792-maligem Vorkommen (7.42 %). Der Fünfclauser erreicht nicht einmal ein Deißigstel (3.35 %) der gesamten Sätze, und minimal ist

Tab. 2: Verteilung der Satztypen (in Clauses) und Satzlängen (in Wörtern).



5. In Abschnitt 4 wurden die quantitativen Ergebnisse unserer Untersuchung tabellarisch und graphisch vorgestellt. Im folgenden werden diese Daten zur Überprüfung der in Kap. 1 formulierten Hypothese herangezogen.
Unsere Hypothese: "Je länger ein Satz, gemessen in der Anzahl der Clauses, desto kürzer die Clauses, gemessen in der Wortzahl", enthält die Überlegung, daß ein Zusammenhang zwischen Satz- und Clauselänge (innerhalb des von uns bearbeiteten Stichprobenkorpus) besteht. Ein erster Schritt zur Überprüfung möglicher funktionaler Zusammenhänge erfolgt über die Bildung von Durchschnittswerten für die Clauselänge in den verschiedenen Satztypen. Die zu bildenden Quotienten dienen dabei als eine erste Annäherung an die Erfassung von Beziehungen zwischen den untersuchten Charakteristika. Regel ist hier, die Häufigkeit einer Variablen durch die für eine andere Variable gemessene Häufigkeit zu dividieren. Wir ermitteln im folgenden, aufgrund der nicht relevanten Häufigkeit der höherclausigen Sätze, ebenfalls nur die Mittelwerte für den 1- bis 11-clausigen Satztyp, deren Häufigkeit mindestens 10 Belege übersteigt.
Zunächst werden die unter einem Satztyp erfaßten Satzlängen (gemessen in Wörtern) mit dem Wert ihres absoluten Vorkommens multipliziert. Sind alle Multiplikationsergebnisse aufgelistet, wird für jeden Satztyp gesondert die Endsumme gebildet, die also alle auf einen Satztyp fallenden Wörter enthält. Dividiert man diese Summe durch die Anzahl der ausgezählten Clauses für jeden Satztyp, so erhält man die jeweilige durchschnittliche Clauselänge. Es ist zu beachten, daß beim Zweiclauser, Dreiclauser etc. die absolute Satzhäufigkeit im Nenner jeweils mit 2, 3 usw. zu multiplizieren ist, um die entsprechende Anzahl der Clauses zu erhalten.
Aus der folgenden Übersicht (Tab. 3) läßt sich ersehen, daß der Wert der berechneten Quotienten kleiner wird, wenn die Anzahl der Clauses pro Satz steigt. Die Abnahmequote ist dabei nicht konstant, sondern verringert sich, je höherclausig der Satz wird. Ein möglicher funktionaler Zusammenhang der betrachteten Variablen wird also nicht linear verlaufen (vgl. Abb. 2). Nach der Ermittlung der durchschnittlichen Clauselänge hat sich also gezeigt,

Tab. 3: Durchschnittliche Clauselänge

Satztyp in Clauses	Durchschnittliche Clauselänge
1	12.4122
2	10.2700
3	9.5500
4	9.0319
5	8.5076
6	8.0040
7	7.9201
8	7.1733
9	6.8413
10	7.0833
11	7.4380

daß ein funktionaler Zusammenhang zwischen Satzlänge und Clauselänge besteht. Die Tatsache, daß die Messwerte bei x = 10 wieder ansteigen, kann (a) entweder durch die kleine Anzahl der Messungen, oder (b) durch eine unbekannte, jedoch relevante Strömung, oder (c) durch die Vereinigung der Stichproben und Überlagerung mehrerer Trends oder schließlich (d) durch einen systematischen Faktor hervorgerufen werden. In diesem Stadium der Untersuchung läßt sich das Problem noch nicht klären.
Die generelle Hypothese über die Abhängigkeit der Komponentenlänge von der Konstruktlänge wird nun mit der von ALTMANN (1980) abgeleiteten Kurve

$$f(x) = Ax^b \; e^{-cx}$$

konkretisiert. In der ursprünglichen Differentialgleichung tritt außer der Proportionalität der Abnahmequote zur unabhängigen Variablen (-c) noch eine Proportionalität der Abnahmequote zur Konstruktlänge (b/x) auf.
Nach der Berechnung der oben aufgeführten Koeffizienten, lassen sich die empirisch erhaltenen Daten nun mit den theoretisch gewonnenen Funktionswerten vergleichen (s. Tab. 4).

der Anteil des 6-clausigen Satztyps: knapp ein Fünfundsechzigstel (1.55 %).
Ein-, Zwei- und Dreiclauser zusammen ergeben 85.8 % aller Satztypen; Ein- bis Sechsclauser decken nahezu das gesamte Untersuchungsmaterial ab, ihre Häufigkeit liegt bei 98.15 %. Die noch verbleibenden Satztypen splitten sich wie folgt: Sieben-, Acht- und Neunclauser erscheinen zusammen 149 mal (1.4 %). Ihre Häufigkeit liegt also jeweils unter der 1%-Grenze (s. Abb.1). Das jeweilige Einzelvorkommen der 10- bis 24-clausigen Sätze liegt sogar unter 0.2 %. Selbst in ihrer Summe erreichen diese Satztypen lediglich 0.45 % aller untersuchten Einheiten. Der 1- bis 11-clausige Satztyp wurde für unsere Berechnungen herangezogen (die Grenze für den Elfclauser lag bei seinem absoluten Auftreten von über 10 mal); die übrigen Satztypen wurden für weitere Berechnungen eliminiert, da sie als Daten innerhalb unserer Analyse als nicht repräsentativ galten.

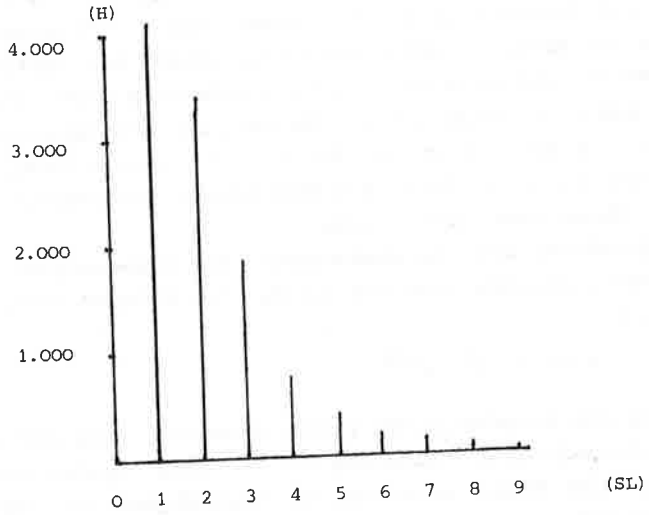


Abb. 1: Histogramm zur 1- bis 9-clausigen Satztypenhäufigkeit (H = absolute Häufigkeit, SL = Satzlänge in Clauses).

Tab. 4: Beobachtete und berechnete Werte für die Gesamtdaten

SL in Clauses	Anzahl der Sätze	beobachtete CL ($f_e(x)$)	berechnete CL ($f_t(x)$)
1	4047	12.4122	12.3638
2	3302	10.2700	10.4111
3	1808	9.5500	9.4547
4	792	9.0319	8.8560
5	357	8.5076	8.4370
6	166	8.0040	8.1244
7	84	7.9201	7.8816
8	44	7.1733	7.6875
9	21	6.8413	7.5292
10	12	7.0833	7.3983
11	11	7.4380	7.2889

6. Zur Überprüfung, ob die berechnete Kurve von der beobachteten signifikant abweicht, benutzen wir den F-Test (s. SACHS 1972) nach der Formel

$$F_{2,n-3} = \frac{SSR \;/\; 2}{SSE \;/\; (n-3)}$$

unter der Voraussetzung, daß die logarithmischen Werte alle Bedingungen einer multiplen linearen Regression erfüllen. Wir verwendeten die gewichteten Daten, wobei das Gewicht jedes Messwertes durch die Zahl der Sätze dargestellt wird. In unserem Fall haben wir 2 und 11-3=8 Freiheitsgrade und erhalten

$$F_{2,8} = 467.47,$$

was mit einer Wahrscheinlichkeit von P = 0.000000005 zu erwarten wäre.
Die empirischen wie die theoretisch ermittelten Daten zeigen also gute Übereinstimmung (s. Abb. 3).

7. Unsere Ergebnisse lassen sich u.a. für Betrachtungen im Bereich der quantitativen Stilistik oder auch innerhalb der Text-

klassifikationsversuche verwenden, unter dem Aspekt, daß es auf unterschiedlichen sprachlichen Ebenen Interrelationen, die sich als Gesetze formulieren lassen, zwischen den jeweils betrachteten Sprachkonstrukten gibt.
Mittels der Kriterien Satzlänge, Clauselänge und Satztypenhäufigkeit lassen sich Abgrenzungs- bzw.Differenzierungsversuche zwischen Einzeltexten anstellen. Das Phänomen Satzlänge ist bereits in einigen Arbeiten hinlänglich untersucht worden. FUCKS (1955) und FUCKS/LAUTER (1965) gebrauchten dieses Kriterium zu autorspezifischen Stilabgrenzungen; DOLEŽEL (1965) charakterisierte es als den Stilindex von Texten. PIEPER (1979) und HOFFMANN (1976) befaßten sich ebenfalls mit der Satzlänge als einer von vielen möglichen Variablen zur Textgruppen- bzw. Textklassenbildung.
Auch die in der vorliegenden Arbeit durchgeführten Berechnungen von Satz- und Clauselänge können dem Vergleich von Einzeltexten als textklassen- und stilspezifische Charakteristika dienen. Die Kategorienbildung nach Häufigkeitsmustern, die Gemeinsamkeiten oder Abstufungen zwischen Einzeltexten ermöglichen, erfolgt durch numerische Charakterisierung. Funktionalzusammenhänge zwischen zwei oder mehreren sprachlichen Konstituenten können mit Hilfe statistischer Mittel (z.B. Durchschnittswerte) errechnet und dargelegt, sprachlich Gestaltetes somit deskriptiv erfaßt werden.
Die Variablen Satzlänge, Clauselänge und Satztypenhäufigkeit dieser Arbeit können für die Kategorisierung von Texten und Stilen als erste Anhaltspunkte nützlich sein, stellen aber lediglich die Behandlung eines Einzelproblems dar. Alle ausschließlich mit diesen Variablen durchgeführten Differenzierungsversuche sind daher nur als "erste Annäherung in Form von Kennwerten für einen Text" (PIEPER 1979:115) zu verstehen.

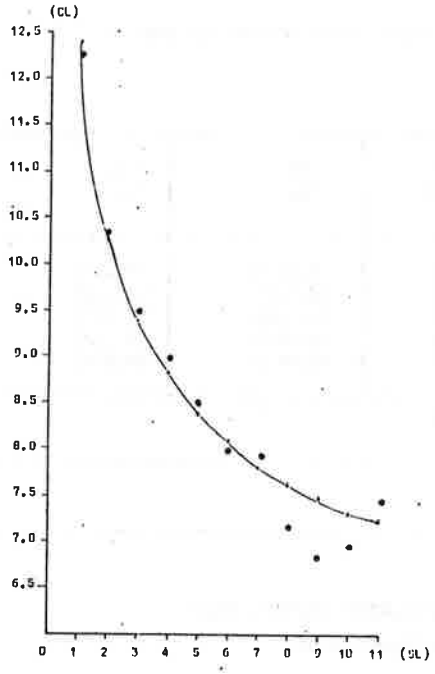


Abb. 2: Empirisch ermittelte Punkte und theoretisch berechnete Kurve: Clauselänge als Funktion der Satzlänge

ANHANG

Um künftige Vergleiche zu erleichtern, bringen wir in jeder Tabelle des Anhangs neben den empirischen Daten die Werte der berechneten Kurve $f(x) = Ax^b e^{-cx}$, den F-Test und die Wahrscheinlichkeit des F-Wertes.
Alle Kurven und Tests werden nur für diejenigen Clauselängen durchgeführt, bei denen die Anzahl der Sätze größer oder gleich 10 ist.

Tab. I: Messwerte in den GESETZESTEXTEN. Hier: GG für die BRD und BVerfGG

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	575	575	8060	14.0174	14.0012
2	284	568	6900	12.1479	12.3808
3	121	363	4333	11.9366	11.5973
4	43	172	1762	10.2442	11.1226
5	18	90	982	10.9111	10.8062
$f(x) = 13.7813x^{-0.211269}\ e^{0.015824x}$					
$\hat{F}_{2,2} = 17.90$; $P = 0.0529$					

Tab.II: Messwerte in den WISSENSCHAFTLICHEN TEXTEN. Hier: "Demokratie und bürgerliche Gesellschaft"

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	174	174	4854	27.8966	27.3857
2	381	762	13081	17.1666	17.7345
3	271	813	12233	15.0467	14.3843
4	121	484	6149	12.7045	12.7934
5	38	190	2192	11.5368	11.9680
6	12	72	780	10.8333	11.5593
$f(x) = 24.5839x^{-0.782577}\ e^{0.107931x}$					
$\hat{F}_{2,3} = 85.79$; $P = 0.0023$					

Tab.III: Messwerte in den WISSENSCHAFTLICHEN TEXTEN. Hier: "Glasfaserverstärkte Kunststoffe"

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	248	248	3731	15.0444	15.0354
2	170	340	3501	10.2971	10.3373
3	63	189	1779	9.4127	9.2826
4	28	112	1028	9.1786	9.2987
$f(x) = 11.4919x^{-0.92825}\ e^{0.268764x}$					
$\hat{F}_{2,1} = 568.25$; $P = 0.0296$					

Tab.IV: Messwerte in den WISSENSCHAFTLICHEN TEXTEN. Hier: "Produktivität und Rationalisierung"

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	271	271	4327	15.9668	15.9711
2	160	320	3800	11.8750	11.8510
3	61	183	1799	9.8306	9.9000
4	19	76	666	8.7632	8.6814
$f(x) = 16.1778x^{-0.411911}\ e^{-0.012856x}$					
$\hat{F}_{2,1} = 1856.1$; $P = 0.0164$					

Tab. V: Messwerte in den ZEITUNGSTEXTEN. Hier: Politik

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	449	449	6622	14.7483	14.7603
2	347	694	7893	11.3732	11.3262
3	189	567	5641	9.9489	10.0245
4	79	316	2973	9.4082	9.4064
5	32	160	1474	9.2125	9.1137
$f(x) = 13.6377x^{-0.496183}\ e^{0.079102x}$					
$\hat{F}_{2,2} = 1610.25$; $P = 0.0006$					

Tab.VI: Messwerte in den ZEITUNGSTEXTEN. Hier: Feuilleton

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	414	414	5717	13.8092	13.8114
2	344	688	7616	11.0698	11.0465
3	179	537	5374	10.0075	10.1082
4	82	328	3257	9.9299	9.7737
5	31	155	1520	9.8065	9.7400
6	17	102	987	9.6765	9.8936
$f(x) = 12.4851x^{-0.467924}\ e^{0.100959}$					
$\hat{F}_{2,3} = 617.21$; $P = 0.0001$					

Tab.VII: Messwerte in den BRIEFEN. Hier: sämtliche Briefe aus T. Mann und J.W. Goethe, sowie die Leserbriefe

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	349	349	4060	11.6332	11.4516
2	328	656	5378	8.1982	8.6200
3	176	528	4272	8.0909	7.7728
4	83	332	2719	8.1898	7.5468
5	39	195	1538	7.8872	7.6304
6	24	144	1055	7.3264	7.9152
7	14	98	694	7.0816	8.3571
$f(x) = 9.8457x^{-0.627786}\ e^{0.151099x}$					
$\hat{F}_{2,4} = 24.49$; $P = 0.0057$					

Tab.VIII: Messwerte in den ROMANTEXTEN. Hier: "Die erste Polka"

Cl/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	379	379	3322	8.7652	8.7417
2	328	656	5253	8.0076	8.0821
3	230	690	5394	7.8174	7.7739
4	112	448	3448	7.6964	7.5997
5	80	400	2999	7.4975	7.4957
6	47	282	2067	7.3298	7.4348
7	25	175	1312	7.4971	7.4030
8	15	120	867	7.2250	7.3923
$f(x) = 8.5950x^{-0.13762}\ e^{0.016929x}$					
$\hat{F}_{2,5} = 135.90$; $P = 0.00004$					

Tab. IX: Messwerte in den ROMANTEXTEN. Hier: "Buddenbrooks"

CL/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	283	283	3296	11.6466	11.7140
2	297	594	6054	10.1919	9.9906
3	198	594	5320	8.9562	9.2016
4	120	480	4214	8.7792	8.7459
5	68	340	2880	8.4706	8.4574
6	34	204	1686	8.2647	8.2681
7	19	133	1114	8.3759	8.1441
8	10	80	644	8.0500	8.0663
$f(x) = 11.4124x^{-0.267228}\ e^{0.026085}$					
$\hat{F}_{2,5} = 128.54$; $P = 0.00005$					

Tab. X: Messwerte in den ROMANTEXTEN. Hier: "Ein Mädchenherz in Fammen"

CL/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	434	434	3394	7.8203	7.8339
2	379	748	5135	6.7744	6.7235
3	185	555	3372	6.0757	6.1607
4	46	184	1057	5.7446	5.7983
5	20	100	579	5.7900	5.5380
$f(x) = 7.7963x^{-0.227466}\ e^{0.004812x}$					
$\hat{F}_{2,2} = 115.95$; $P = 0.0086$					

Tab. XI: Messwerte in den ROMANTEXTEN. Hier: "Wie kommt das Salz ins Meer ?"

CL/Satz	Anzahl Sätze	Anzahl Clauses	Worte gesamt	CL emp.	CL theor.
1	471	471	2849	6.0488	6.0416
2	284	568	3212	5.6549	5.6875
3	135	405	2282	5.6346	5.6045
4	59	236	1340	5.6780	5.6271
5	21	105	596	5.6762	5.7079
6	15	90	544	6.0444	5.8276
7	12	84	477	5.6786	5.9765
$f(x) = 5.7486x^{-0.158872}\ e^{0.049719}$					
$\hat{F}_{2,4} = 30.22$; $P = 0.0039$					

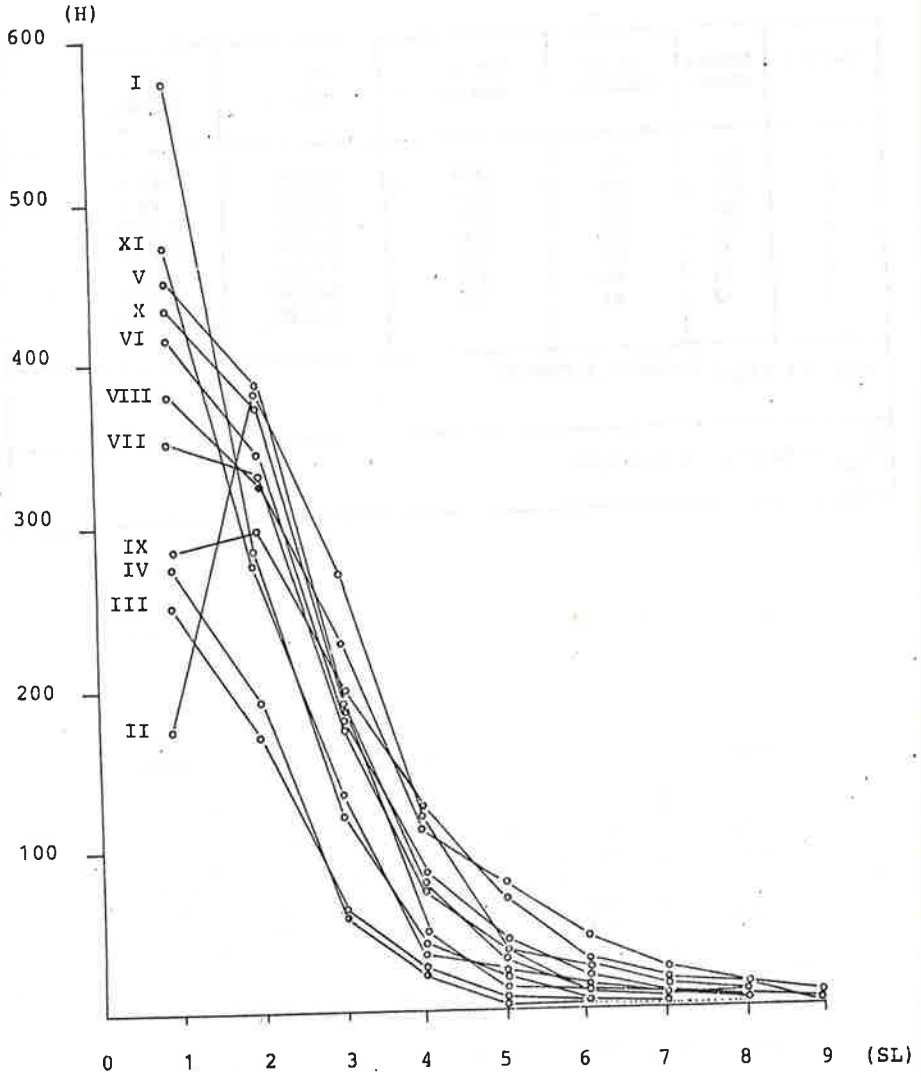


Abb. 3: Häufigkeitspolygone für die 1- bis 9- clausigen Satztypen aus den Einzeltexten.
(I bis XI analog der Tabellennumerierung in Kap. 6.2 die Werte in Abb. 2 sind die Summen dieser Häufigkeiten).
(H = absolute Häufigkeit, SL = Satzlänge in Clauses).

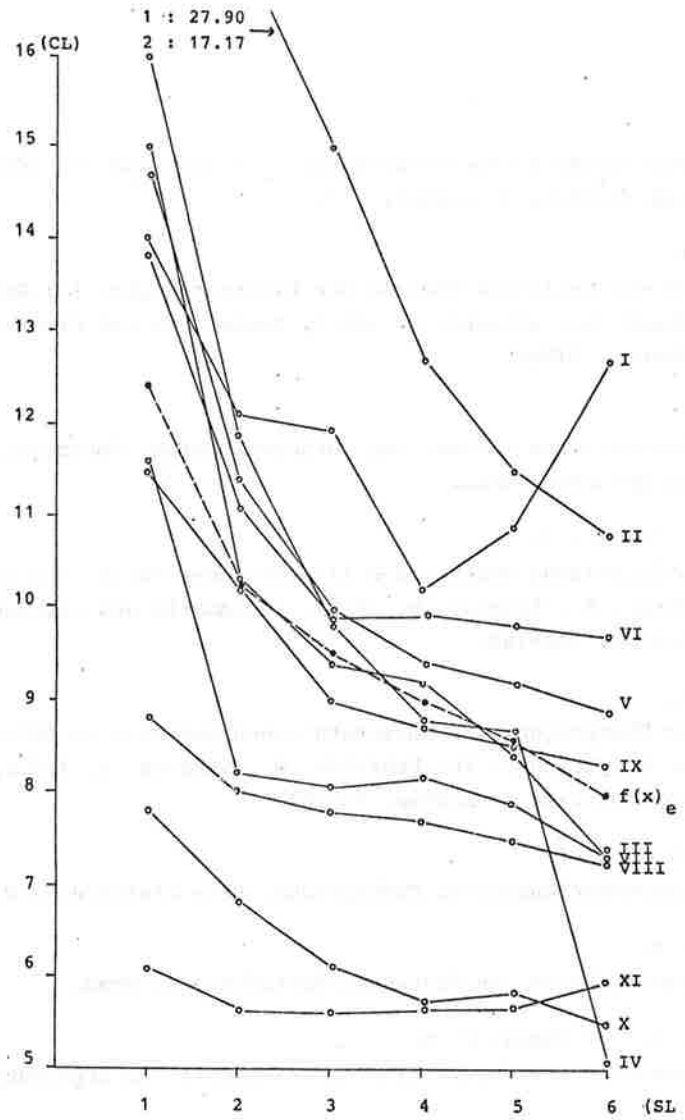


Abb. 4: Die durchschnittlichen Clauselängen (CL) in den Einzeltexten. (SL ≙ Satzlänge; I bis XI analog der Tabellennumerierung in 6, $f(x)_e$ analog Kap. 4.2).

LITERATUR

Altmann, G.
1980 Prolegomena to Menzerath's Law. In: Grotjahn, R. (Ed.), Glottometrika 2. Bochum, 1-10.

Doležel, L.
1965 Zur statistischen Theorie der Dichtersprache. In: Gunzenhäuser, R., Kreuzer, H. (Ed.), Mathematik und Dichtung. München, 275ff.

Fucks, W.
1955 Mathematische Analyse von Sprachelementen, Sprachstilen und Sprachen. Köln.

Fucks, W., Lauter, J.
1956 Mathematische Analyse des literarischen Stils. In: Gunzenhäuser, R., Kreuzer, H. (Ed.), Mathematik und Dichtung. München, 107-122.

Gerlach, R.
1982 Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie. In: Lehfeldt, W., Strauss, U. (Ed.), Glottometrika 4. Bochum, 95-102.

Hoffmann, L.
1976 Kommunikationsmittel Fachsprache. Eine Einführung. Berlin.

Menzerath, P.
1954 Architektonik des deutschen Wortschatzes. Bonn.

Menzerath, P., de Oleza, J. M.
1928 Spanische Lautdauer. Eine experimentelle Untersuchung. Berlin.

Pieper, U.
1979 Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse. Tübingen.

Sachs, L.
1974 Angewandte Statistik. Berlin.

Winter, W.
1974 Untersuchungen zur Quantitativen Stilistik. Prosa der Gegenwart in schriftlicher Form. Unveröffentlicht. Kiel.

Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse

R. Köhler, Essen

1. MARKOV-KETTEN

Während sich die Algebraische Linguistik mit der Formalisierung struktureller und kombinatorischer Beschreibungen mit Hilfe deduktiver Systeme befaßt und sich mit subjektiver Überprüfung (Introspektion, "kompetenter" Hörer/Sprecher) begnügt, verwendet die Quantitative Linguistik stochastische Modelle, die einerseits deterministische Gesetze oder Regeln als Spezialfälle einschließen, andererseits Methoden sowohl für eine kontrollierte Heuristik als auch zur objektiven empirischen Überprüfung von Aussagen zur Verfügung stellen.

Unter den mathematischen Mitteln, die die Quantitative Linguistik zu Zwecken der Beschreibung und der Modellbildung anwendet, spielt die Wahrscheinlichkeitstheorie eine wichtige Rolle u.a. bei der Erfassung und Darstellung von Struktur, Variabilität und Evolution sowie der Kombinatorik linguistischer Einheiten.
Bei linguistischen Untersuchungen ist sehr oft die Kombinierbarkeit hinsichtlich der Konkatenation von Interesse, d.h. die lineare Verkettbarkeit von Einheiten oder Klassen von Einheiten in der zeitlichen (bzw. bei der geschriebenen Sprache in einer räumlichen) Dimension. Im Rahmen der Wahrscheinlichkeitstheorie wird die Zugehörigkeit eines Elements zu einer Klasse (Kategorie) als Ausprägung einer Zufallsvariablen aufgefaßt. Das entsprechende quantitative Modell, das das dynamische Verhalten einer Zufallsvariablen in einer beliebigen

Dimension charakterisiert, nennt man allgemein einen stochastischen Prozeß. Werden sowohl die Einheiten wie die Dimension als diskrete Parameter aufgefaßt (wie es in der Linguistik meistens der Fall ist), wird von einer Markov-Kette gesprochen; die Werte, die die Zufallsvariable annehmen kann, heißen Zustände der Kette.
Für die Definition der Zustände gibt es grundsätzlich keine Beschränkungen, so daß eine Markov-Kette als Modell für einen weiten Bereich sprach- und literaturwissenschaftlicher Fragestellungen geeignet ist. Folgende zwei Beispiele können zur Verdeutlichung dienen:

- Ein wichtiges Gestaltungsmittel in poetischen Texten ist die Wiederholung von Elementen (z.B. die Alliteration). Diese Art von Wiederholung ist mit deterministischen Mitteln nicht erfaßbar ("es gibt keine feste Regel"); in ALTMANN/DILLER/SAPPOK/STRAUß (erscheint) wird dazu ein spezieller stochastischer Prozeß abgeleitet.
- BRAINERD (1976) analysiert Texte als Markov-Ketten, deren Zustände für die Zugehörigkeit von Wörtern zu Wortarten stehen.

Auch für die grammatische Beschreibung von Sprachen sind Markov-Ketten einsetzbar; die Grenze ihrer Anwendbarkeit liegt da, wo Selbsteinbettungen von Kategorien (besonders bei unendlicher Einbettungstiefe) beschrieben werden müssen, wie es für die Syntax natürlicher Sprachen typisch ist (vgl. CHOMSKY 1957, OSGOOD 1963, LEVELT 1974). In solchen Fällen müssen probabilistische Grammatiken anderen Typs herangezogen werden (vgl. SUPPES 197o, LEVELT 1974).

Für die nachfolgende Darstellung des Markov-Modells beschränken wir uns auf dichotomische Variable; dies hat den Vorteil besserer Überschaubarkeit (besonders bei Ketten höherer Ordnung). Für Ketten mit mehr als zwei Zuständen gilt das Gesagte analog.

1.1 MARKOV-KETTEN ERSTER ORDNUNG

Untersuchen wir beispielsweise eine Phonemfolge auf das Vorhandensein eines Merkmals (z.B. "konsonantisch"), dann gehen wir davon aus, daß die jeweilige Wahrscheinlichkeit einer Variablenausprägung unabhängig von der Stelle j in der Kette ist (d.h. daß die Übergangswahrscheinlichkeit von einem Phonem zum anderen nicht davon abhängt, wie lange der Sprecher schon redet). Ketten mit dieser Eigenschaft nennt man homogene Markov-Ketten[1]. Da außerdem die Anzahl der möglichen Zustände begrenzt ist, handelt es sich um eine endliche stationäre Markov-Kette, die allein durch die Matrix der Übergangswahrscheinlichkeiten definiert ist. In Abb.1 ist eine Markov-Kette graphisch dargestellt, deren Zustände "L" und "K" für lang- bzw. kurzvokalige Silben eines Textes stehen könnten. Hier sind die Kanten des Graphen mit den bedingten Wahrscheinlichkeiten der Zustände bewertet; es gilt:
$p_{ab} = P(X_{j+1} = b \mid X_j = a)$, wo $a,b \in \{L,K\}$. Also ist p_{KL} die

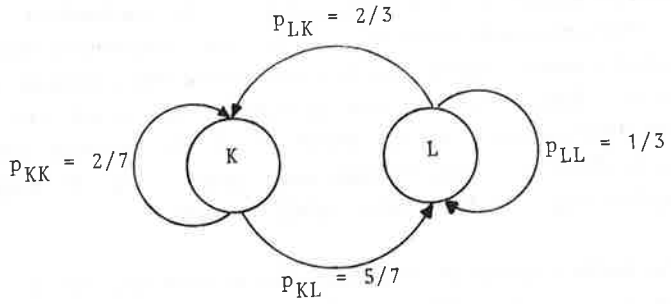


Abb.1 Graphische Darstellung einer binären stationären Markov-Kette erster Ordnung

Wahrscheinlichkeit dafür, daß hinter einer kurzvokaligen Silbe an einer beliebigen Stelle j eine langvokalige an der Stelle j+1 erscheint. Die zugehörige Matrix der Übergangswahrscheinlichkeiten wird auf folgende Weise notiert:

$$P = \begin{pmatrix} p_{LL} & p_{LK} \\ p_{KL} & p_{KK} \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \\ 5/7 & 2/7 \end{pmatrix}. \qquad (1.1)$$

Wegen der Beschränkung auf Dichotomien kann man hier die Zustände auch mit "o" (Merkmal nicht vorhanden) und "1" (Merkmal vorhanden) bezeichnen:

$$P = \begin{pmatrix} p_{oo} & p_{o1} \\ p_{1o} & p_{11} \end{pmatrix}. \qquad (1.2)$$

Diese Schreibweise soll von hier an beibehalten werden.

Die Matrix P der Übergangswahrscheinlichkeiten gibt uns also für jede Stelle j einer Kette die Wahrscheinlichkeit dafür an, daß an der darauffolgenden Stelle j+1 der Zustand o bzw. der Zustand 1 eintritt (vgl. Abb.2a). Wir können aber auch die Wahrscheinlichkeiten für den übernächsten Zustand (vgl. Abb.2b) berechnen, indem wir die Matrix quadrieren:

$$P^2 = \begin{pmatrix} p_{oo}^{(2)} & p_{o1}^{(2)} \\ p_{1o}^{(2)} & p_{11}^{(2)} \end{pmatrix}. \qquad (1.3)$$

Allgemein gibt die r-te Potenz der Matrix die Übergangswahrscheinlichkeiten für die Zustände der Kette nach r Schritten (vgl. Abb.2c) an:

$$P^r = \begin{pmatrix} p_{oo}^{(r)} & p_{o1}^{(r)} \\ p_{1o}^{(r)} & p_{11}^{(r)} \end{pmatrix}. \qquad (1.4)$$

a) ...oo1o1ooX...
j

b) ...oo1o1ooXX...
j j+2

c) ...oo1o1ooXX...XXX...
j j+r

Abb.2 Zum Übergang nach r Schritten

Beispiel 1:

Unter der Annahme, daß die Folge von betonten (Zustand 1) und unbetonten (Zustand o) Silben eines gegebenen Textes eine Realisation der Markov-Kette mit der Matrix

$$P = \begin{pmatrix} 1/3 & 2/3 \\ 5/7 & 2/7 \end{pmatrix}$$

ist, gibt die zweite Potenz von P,

$$P^2 = \begin{pmatrix} 37/63 & 26/63 \\ 65/147 & 82/147 \end{pmatrix}$$

die Wahrscheinlichkeiten für den jeweils übernächsten Zustand (an der Stelle j+2) an. $p_{11}^{(2)} = 82/147 \approx 0.56$ bezeichnet also z.B. die Wahrscheinlichkeit, daß nach einer betonten Silbe die übernächste wieder betont ist.

Da die Zeilensummen der Wahrscheinlichkeiten immer 1 sind, können wir die Matrix auch folgendermaßen schreiben:

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \qquad (1.5)$$

mit $0 \leq p,q \leq 1$; $|1-p-q| < 1$. Daraus bestimmt sich (vgl. BHAT 1972:21ff) die Matrix der Übergangswahrscheinlichkeiten als

$$P^r = \begin{pmatrix} \frac{q + p(1-p-q)^r}{p+q} & \frac{p - p(1-p-q)^r}{p+q} \\ \frac{q - q(1-p-q)^r}{p+q} & \frac{p + q(1-p-q)^2}{p+q} \end{pmatrix} \qquad (1.6)$$

1.2 MARKOV-KETTEN HÖHERER ORDNUNG

Für linguistische Zwecke ist es also nötig, Abhängigkeiten zwischen Merkmalsausprägungen erfassen zu können, die über die direkte Nachbarschaft hinausreichen. Die Wahrscheinlichkeit dafür, daß ein Phonem z.B. konsonantisch ist, hängt in den Sprachen, die Konsonanten- oder Vokalcluster bilden, nicht nur vom unmittelbaren Vorgänger ab. Dementsprechend wird die Zufallsvariable - statt wie bisher an zwei - an drei oder mehr Stellen bzw. Zeitpunkten gleichzeitig betrachtet, und daher sind die Wahrscheinlichkeiten drei- oder mehrdimensional verteilt (s. Abb.3).

Um das Modell der Markov-Ketten dieser Forderung gemäß zu erweitern, müssen komplexe Zustände derart definiert werden, daß wieder nur zwei verschiedene Zeitpunkte gleichzeitig eine Rolle spielen. Sollen z.B. die Ausprägungswahrscheinlichkeiten der dichotomischen Variablen X_j nicht nur von X_{j-1}, also

dem direkt vorhergehenden Zustand, sondern auch von X_{j-2} abhängen, so müssen die entsprechenden dreidimensionalen Ereignisse (vgl. Abb. 3) auf zweidimensionale abgebildet werden. Dies geschieht durch Verdopplung der mittleren Glieder (vgl. Abb. 4), so daß die Zustände der Markov-Kette geordnete Paare aus elementaren Ereignissen sind. Dabei ist zu beachten, daß der Zustand oo bzw. der Zustand 1o der neuen Kette

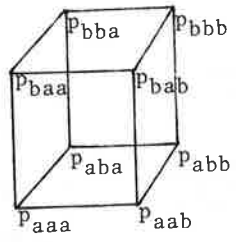


Abb.3 Dreidimensionale Verteilung von Übergangswahrscheinlichkeiten

nicht in einem Schritt in den Zustand 11 oder 1o übergehen kann, so daß die Matrix der Übergangswahrscheinlichkeiten von vornherein Nullen enthält. Die Matrix für unser Beispiel hat demnach folgendes Aussehen:

$$p = \begin{pmatrix} p_{oooo} & p_{ooo1} & o & o \\ o & o & p_{o11o} & p_{o111} \\ p_{1ooo} & p_{1oo1} & o & o \\ o & o & p_{111o} & p_{1111} \end{pmatrix} \qquad (1.7)$$

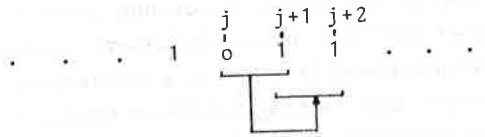


Abb.4 Markov-Kette zweiter Ordnung: Übergang vom Zustand o1 in den Zustand 11

mit $p_{abbc} = P(X_{j+2} = c \mid X_{j+1} = b, X_j = a)$. Man spricht hier von einer Markov-Kette zweiter Ordnung. Analog entstehen Ketten dritter und höherer Ordnung.

Beispiel 2:

Um armenische Texte als Phonemkette im Hinblick auf das Merkmal "vokalisch" als Realisation einer Markov-Kette zu beschreiben, benötigt man eine Kette zweiter Ordnung:

$$P_{voc_A} = \begin{pmatrix} 0.22 & 0.78 & 0 & 0 \\ 0 & 0 & 0.56 & 0.44 \\ 0.32 & 0.68 & 0 & 0 \\ 0 & 0 & 0.60 & 0.40 \end{pmatrix} .$$

Um vorherzusagen, ob ein Phonem an einer Stelle j vokalisch ist oder nicht, muß die Merkmalsausprägung der beiden vorhergehenden Phoneme bekannt sein. Nach der Matrix P_{voc_A} ist die Wahrscheinlichkeit dafür, daß nach zwei Vokalen ein Nicht-Vokal erscheint, gleich $p_{111o} = 0.6o$.

1.3 BEHARRUNGSTENDENZ

Ein weiterer Parameter der Markov-Kette, der sich aus der Matrix der Übergangswahrscheinlichkeiten gewinnen läßt, ist die Wahrscheinlichkeit, mit der die Kette über n Schritte in einem gegebenen Zustand verharrt.
Mit p haben wir die Wahrscheinlichkeit für den o1-Übergang und mit 1-p die Wahrscheinlichkeit für den oo-Übergang bezeichnet. Die Wahrscheinlichkeit für einen o1-Übergang nach einem oo-Übergang ist folglich p(1-p), die Wahrscheinlichkeit für einen o1-Übergang nach n oo-Übergängen ist $p(1-p)^n$; wir erhalten die geometrische Verteilung.

$$P(r_o = n) = p(1-p)^n \qquad n = 1,2... \qquad (1.8)$$
$$P(r_1 = n) = q(1-q)^n$$

Hier bezeichnet r_a die Anzahl der aufeinanderfolgenden Ausprägungen a ($a \in \{o,1\}$) von der Stelle j+1 an, wenn die Kette in j im Zustand a ist. Die Erwartungen und Varianzen bestimmen sich aus der geometrischen Verteilung:

$$E(r_o) = \frac{1-p}{p} \; ; \qquad E(r_1) = \frac{1-q}{q} \qquad (1.9)$$

$$V(r_o) = \frac{1-p}{p^2} \; ; \qquad V(r_1) = \frac{1-q}{q^2} \; . \qquad (1.1o)$$

Beispiel 3:

Eine empirisch gewonnene Matrix der Übergangswahrscheinlichkeiten für das phonematische Merkmal "rund" der indonesischen Phoneme ist

$$P_{rnd_I} = \begin{pmatrix} o.9352 & o.o648 \\ o.9859 & o.o141 \end{pmatrix} .$$

Befindet sich die Kette an der Stelle j im Zustand o, dann ist die Wahrscheinlichkeit, daß sie für weitere drei Schritte in diesem Zustand bleibt,

$$P(r_o = 3) = (o.o648)\cdot(o.9352)^3 = o.o53.$$

Die zu erwartende Anzahl von o-Zuständen in unmittelbarer Folge ist

$$E(r_o) = \frac{o.9352}{o.o648} = 14.4321$$

mit der Varianz

$$V(r_o) = \frac{o.9352}{(o.o648)^2} = 222.7176.$$

2. SCHÄTZUNGS- UND TESTVERFAHREN

2.1 SCHÄTZEN VON ÜBERGANGSWAHRSCHEINLICHKEITEN AUS EMPIRISCHEN DATEN

In vielen Fällen hat man zunächst keine theoretisch fundierten Annahmen über die Wahrscheinlichkeiten, so daß man sie aus einem Korpus abschätzen muß. Dazu erfaßt man die beobachteten Häufigkeiten n_{ab} von Übergängen der Kette vom Zustand a in den Zustand b für alle s möglichen Zustände in einer Häufigkeitstabelle (Abb.5). Hier sind a und b je nach Ordnung der

a \ b	o	1	2	. . .	s-1	n_a
o	n_{oo}	n_{o1}	n_{o2}	. . .	$n_{o,s-1}$	n_o
1	n_{1o}	n_{11}	n_{12}	. . .	$n_{1,s-1}$	n_1
2	n_{2o}	n_{21}	n_{22}	. . .	$n_{2,s-1}$	n_2
.	.	.	.	. . .	.	.
.	.	.	.	. . .	.	.
.	.	.	.	. . .	.	.
s-1	$n_{s-1,o}$	$n_{s-1,1}$	$n_{s-1,2}$	. . .	$n_{s-1,s-1}$	n_{s-1}
	n'_o	n'_1	n'_n	. . .	n'_{s-1}	N

Abb.5 Häufigkeitstabelle für Übergänge

Kette einfache oder zusammengesetzte Zustände. Jede Zeile der Tabelle kann als Stichprobe des Umfangs n_a aus einer multinominalen Verteilung betrachtet werden. Für die zugehörigen Wahrscheinlichkeiten p_{ab} gilt $\sum_b p_{ab} = 1$, so daß die Wahrscheinlichkeit für das Zustandekommen der beobachteten Zeile a

$$p_a = \frac{n_a!}{n_{ao}!\, n_{a1}! \cdots n_{a,s-1}!}\, p_{ao}^{n_{ao}}\, p_{a1}^{n_{a1}} \cdots p_{a,s-1}^{n_{a,s-1}} \qquad (2.1)$$

ist. Für die gesamte Tabelle ist entsprechend die Realisationswahrscheinlichkeit gleich $\prod_a p_a$ (vgl. BHAT 1972:97). Die Verteilung der Zeilensummen ist von den p_{ab} unabhängig und kann als $A(n_{ab})$ separiert werden. Die Likelihood-Funktion ist demnach

$$L = A(n_{ab}) \prod_{a=o}^{s-1} \left(\frac{n_a! \prod_{b=o}^{s-1} p_{ab}^{n_{ab}}}{\prod_{b=o}^{s-1} (n_{ab}!)} \right) . \qquad (2.2)$$

Die Maximum-Likelihood-Schätzung für die Parameter p_{ab} erhält man schließlich durch Nullsetzen der Ableitung (s. BHAT 1972):

$$\frac{\partial \ln L}{\partial p_{ab}} = o \longrightarrow \hat{p}_{ab} = \frac{n_{ab}}{n_a} . \qquad (2.3)$$

Beispiel 4:

Unter der Voraussetzung, daß eine beobachtete Folge von Ausprägungen des Merkmals "continuant" in einem georgischen Korpus eine Realisation einer Markov-Kette zweiter Ordnung ist, soll aus der zweidimensionalen Häufigkeitstabelle die Matrix der Übergangswahrscheinlichkeiten abgeschätzt werden.

	oo	o1	1o	11	
oo	1	57	o	o	58
o1	o	o	13o	465	595
1o	57	537	o	o	594
11	o	o	464	1358	1822
	58	594	594	1823	3o69

Nach (2.3) erhält man durch Division der Einzelhäufigkeiten durch die zugehörige Zeilensumme die Maximum-Likelihood-Schätzungen für die Übergangswahrscheinlichkeiten. Die Matrix ist also

$$\hat{P}_{cnt_G} = \begin{pmatrix} o.o172 & o.9828 & o & o \\ o & o & o.2185 & o.7815 \\ o.o96o & o.9o4o & o & o \\ o & o & o.2547 & o.7453 \end{pmatrix} .$$

2.2 TESTEN EINER MATRIX VON ÜBERGANGSWAHRSCHEINLICHKEITEN

Will man überprüfen, ob vorliegende Daten eine Realisation einer Markov-Kette mit gegebener Matrix sein kann, so kann man das z.B. mit Hilfe des Likelihood-Ratio-Tests $2\hat{I}.tun^2$. Dabei ist

$$2\hat{I} = 2 \sum_a \sum_b n_{ab} \ln \frac{n_{ab}}{n_a p_{ab}} \qquad (2.4)$$

mit s(s-1) Freiheitsgraden asymptotisch χ^2-verteilt; s ist die Anzahl der Zustände. Kommen in der Matrix Nullwahrscheinlichkeiten vor, so muß für jede von ihnen ein Freiheitsgrad abgezogen werden (vgl. KULLBACK/KUPPERMAN/KU 1962:596). Dies gilt jedoch nicht für diejenigen Nullen, die durch Abbildung einer Markov-Kette höherer Ordnung auf eine zweidimensionale Übergangswahrscheinlichkeit mit komplexen Zuständen entstanden sind (vgl. BARTLETT 1978:276; s. Abb.4 und (1.7)).
Zu Rechenzwecken schreiben wir (2.4) in folgender Form:

$$2\hat{I} = 2 \sum_a \sum_b n_{ab} \ln n_{ab} - 2 \sum_a n_a \ln n_a - 2 \sum_a \sum_b n_{ab} \ln p_{ab}. \qquad (2.5)$$

2.3 TESTEN DER ORDNUNG EINER MARKOV-KETTE

Das Likelihood-Ratio-Kriterium für den Test der Hypothese, daß eine Markov-Kette mit s Zuständen von der Ordnung m ist, gegen die Hypothese, daß die Kette von der Ordnung m+1 ist, lautet allgemein (s. ANDERSON/GOODMAN 1957:1o2):

$$2\hat{I} = 2 \sum_{a..f} n_{abcd..ef} \ln \frac{\hat{p}_{abcd..ef}}{\hat{p}_{bcd..ef}} . \qquad (2.6)$$

$2\hat{I}$ ist asymptotisch χ^2-verteilt, wenn die Nullhypothese gültig ist. Die Anzahl der Freiheitsgrade beträgt in diesem Fall $s^{m-1}(s-1)^2$ (vgl. HOEL 1954:432).

Beispiel 5:

Gegeben sei eine Matrix P_{th} von Übergangswahrscheinlichkeiten, die aus theoretischen Überlegungen stammen:

$$P_{th} = \begin{pmatrix} o.2 & o.8 \\ o.4 & o.6 \end{pmatrix} .$$

Es soll getestet werden, ob die beobachteten Häufigkeiten

	o	1	
o	15o	619	769
1	618	1o67	1685
	768	1886	2454

eines Merkmals aus der Markov-Kette mit der Hypothese verträglich sind, daß der Kette die Matrix P_{th} zugrundeliegt. Nach (2.5) berechnen wir:

$2\hat{I} =$ $2($ $\sum_a \sum_b n_{ab} \ln n_{ab}$ $5.o1o6 \cdot 15o$
$+ 6.4281 \cdot 619$
$+ 6.4265 \cdot 618$
$+ 6.9726 \cdot 1o67$

$- \sum_a n_a \ln n_a$ $- 6.6451 \cdot 769$
$- 7.4295 \cdot 1685$

$- \sum_a \sum_b n_{ab} \ln p_{ab}$ $+ 1.6o94 \cdot 15o$
$+ o.2231 \cdot 619$
$+ o.9163 \cdot 618$
$)$ $+ o.51o8 \cdot 1o67$

$2 \cdot$ 3.9758

$2\hat{I} =$ 7.9516

Die Matrix enthält keine Null,so daß die Anzahl der Freiheitsgrade s(s-1) = 2·1 = 2 ist. Da $2\hat{I}$ größer ist als $\chi^2_{2;0.05} = 5.99$, wird die Nullhypothese mit der Fehlerwahrscheinlichkeit 0.05 abgelehnt.

Davon muß je ein Freiheitsgrad für jede Null abgezogen werden, die in der Matrix der $n_{bcd..ef}$ vorkommt, da sie eine Beschränkung der Variabilität der Matrix der $n_{abcd..ef}$ darstellt. Für die Berechnung von $2\hat{I}$ wird o ln o als o definiert.

Eine Korrektur für $2\hat{I}$ im Falle von Nullfrequenzen wird von KU (1963) abgeleitet. Danach wird für jede beobachtete Nullfrequenz die Zahl 1 von $2\hat{I}$ subtrahiert.
Der Test der Ordnung eins gegen die zweite Ordnung bzw. den der zweiten gegen die dritte Ordnung lautet[3]:

$$2\hat{I}_{1/2} = 2 \sum_{a\,b\,c} n_{abc} \ln \frac{n_{abc}}{\frac{n_{ab}\, n_{bc}}{n_b}} \qquad (2.7)$$

mit $s(s-1)^2$ Freiheitsgraden;

$$2\hat{I}_{2/3} = 2 \sum_{a\,b\,c\,d} n_{abcd} \ln \frac{n_{abcd}}{\frac{n_{abc}\, n_{bcd}}{n_{bc}}} \qquad (2.8)$$

mit $s^2(s-1)^2$ Freiheitsgraden. Zu Berechnungszwecken schreiben wir für (2.7)

$$2\hat{I}_{1/2} = 2 \Big(\sum_{a\,b\,c} n_{abc} \ln n_{abc} + \sum_b n_b \ln n_b - \sum_{a\,b} n_{ab} \ln n_{ab} - \sum_{b\,c} n_{bc} \ln n_{bc} \Big) \qquad (2.9)$$

und für (2.8)

$$2\hat{I}_{2/3} = 2 \Big(\sum_{a\,b\,c\,d} n_{abcd} \ln n_{abcd} + \sum_{b\,c} n_{bc} \ln n_{bc} - \sum_{a\,b\,c} n_{abc} \ln n_{abc} - \sum_{b\,c\,d} n_{bcd} \ln n_{bcd} \Big). \qquad (2.10)$$

Aus der allgemeinen Form (2.6) gewinnen wir analog

$$2\hat{I}_{o/1} = 2 \sum_{a\,b} n_{ab} \ln \frac{n_{ab}}{\frac{n_a\, n_b}{N}} \qquad (2.11)$$

$$= 2 \Big(\sum_{a\,b} n_{ab} \ln n_{ab} + N \ln N - \sum_a n_a \ln n_a - \sum_b n_b \ln n_b \Big) \qquad (2.12)$$

mit $(s-1)^2$ Freiheitsgraden und

$$2\hat{I}_{3/4} = 2 \sum_{a\,b\,c\,d\,e} n_{abcde} \ln \frac{n_{abcde}}{\frac{n_{abcd}\, n_{bcde}}{n_{bcd}}} \qquad (2.13)$$

$$= 2 \Big(\sum_{a\,b\,c\,d\,e} n_{abcde} \ln n_{abcde} + \sum_{b\,c\,d} n_{bcd} \ln n_{bcd} - \sum_{a\,b\,c\,d} n_{abcd} \ln n_{abcd} - \sum_{b\,c\,d\,e} n_{bcde} \ln n_{bcde} \Big) \qquad (2.14)$$

mit $s^3(s-1)^2$ Freiheitsgraden. Der Test $2\hat{I}_{o/1}$ ist gleichbedeutend mit dem Test der statistischen Unabhängigkeit der aufeinanderfolgenden Merkmalsausprägungen. In unserem Fall ist die Anzahl der Freiheitsgrade wegen s=2 stets 2^{m-1}, wenn die Ordnung m-1 gegen die Ordnung m getestet wird.

Beispiel 6:

Wollen wir z.B. überprüfen, ob die erste Ordnung für das Merkmal "consonantal" im Indonesischen ausreicht, dann testen wir diese Nullhypothese gegen die Alternative, daß die vorliegende Kette zweiter Ordnung ist.

Aus der Häufigkeitstabelle[4]

	o	o	
oo	43	11o	153
o1	647	286	933
1o	1o9	823	932
11	285	o	285
	1o84	1219	23o3

berechnen wir $2I_{1/2}$ nach (2.9):

$$\sum_{abc}\sum\sum n_{abc} \ln n_{abc} = \begin{aligned} &3.7612 \cdot 43 \\ &+ 4.7oo5 \cdot 11o \\ &+ 6.4723 \cdot 647 \\ &+ 5.656o \cdot 286 \\ &+ 4.6913 \cdot 1o9 \\ &+ 6.713o \cdot 823 \\ &+ 5.6525 \cdot 285 \\ &+ o \cdot o \end{aligned}$$

$$= 14131.o85$$

$$\sum_{b} n_b \ln n_b = (43 + 11o + 1o9 + 823) \cdot 6.9893 + (647 + 286 + 285 + o) \cdot 7.1o5o$$

$$= 16237.2767$$

$$\sum_{ab}\sum n_{ab} \ln n_{ab} = \begin{aligned} &5.o3o4 \cdot 153 \\ &+ 6.8384 \cdot 933 \\ &+ 6.8373 \cdot 932 \\ &+ 5.6525 \cdot 285 \end{aligned}$$

$$= 15133.243$$

$$\sum_{bc}\sum n_{bc} \ln n_{bc} = \begin{aligned} &(43 + 1o9) \cdot 5.o239 \\ &+ (11o + 823) \cdot 6.8384 \\ &+ (647 + 285) \cdot 6.8373 \\ &+ (286 + o) \cdot 5.656o \end{aligned}$$

$$= 15133.87$$

$$2\mathfrak{I} = 2\,(14131.o85 + 16237.2767 - 15133.243 - 15133.87)$$
$$= 2 \cdot 1o1.248$$
$$= 2o2.496.$$

In der Tabelle ist eine Nullfrequenz enthalten, so daß wir von diesem Wert die Zahl 1 subtrahieren müssen:

$$2\mathfrak{I}_{korr} = 2o1.469.$$

Die zugehörige Tabelle erster Ordnung enthält keine Null; es wird kein Freiheitsgrad abgezogen: $FG = 2(2-1)^2 = 2$.

Da $\chi^2_{2;o.o5} = 5.99 < 2\mathfrak{I}_{korr} = 2o1.469$ lehnen wir die Nullhypothese, daß die Kette der Ordnung eins ist, ab. Die Irrtumswahrscheinlichkeit bei dieser Entscheidung liegt bei o.o5.

3. AUTOKORRELATION

Ein Verfahren, das bereits mehrfach zu Textanalysen verwendet worden ist[5], ist die Autokorrelation; sie wird im vorliegenden Beitrag vor allem deshalb behandelt, weil sie in enger Beziehung zum Markov-Modell steht.

3.1 EMPIRISCHE AUTOKORRELATIONSANALYSE

Bei diesem Verfahren wird jeweils die Ausprägung eines Merkmals an allen Stellen (bzw. Zeitpunkten) j mit der Ausprägung desselben Merkmals an allen Stellen j+r verglichen, wobei r = 1,..,R; j = 1,..,L-r; L die Länge des zu untersuchenden Korpus ist, und R die maximale Verschiebungsweite angibt.

Dies ist gleichbedeutend mit dem Vergleich der Merkmalsausprägungen bei j mit denen bei j-r, r = 1,..,R; j = 1+r,..,L. Eine Autokorrelationsfunktion ist immer symmetrisch zur Verschiebungsweite Null bei j, sodaß beide möglichen Richtungen der Variablenwirkung berücksichtigt sind, auch wenn man meist nur den Bereich der positiven Verschiebungen explizit darstellt.
Da unsere Variablen dichotomisch sind (Ausprägung "o" := Merkmal trifft nicht zu; Ausprägung "1" := Merkmal trifft zu), sind als Beobachtungen die Paare (o,o), (o,1), (1,o) und (1,1) möglich. An erster Stelle jedes Paares steht die beobachtete Ausprägung an der Stelle j, und an zweiter Stelle die an der Stelle j+r beobachtete (vgl. Abb.2). Ihre Häufigkeiten im Korpus können wir in Tabellen erfassen, wie sie die Abb.6 zeigt.

Für eine solche tetrachorische Tabelle ist der Φ-Koeffizient definiert, der sich z.B. aus dem Pearsonschen Koeffizienten ableiten läßt:

$$\rho = \frac{Cov(X_j, X_{j+r})}{\sqrt{V(X_j)\ V(X_{j+r})}} \qquad (3.1)$$

X_j \ X_{j+r}	o	1	
o	n_{oo}	n_{o1}	n_o
1	n_{1o}	n_{11}	n_1
	n'_o	n'_1	N

Abb.6 Zweidimensionale Häufigkeitsverteilung der Variablen X an den Stellen j und j+r

oder

$$\rho = \frac{E(X_j X_{j+r}) - E(X_j)E(X_{j+r})}{\sqrt{[E(X_j^2) - E^2(X_j)]\,[E(X_{j+r}^2) - E^2(X_{j+r})]}} \qquad (3.2)$$

$E(X_j)$ steht für die mathematische Erwartung von X_j; das Quadrat dieser Erwartung, $[E(X_j)]^2$, schreiben wir kurz $E^2(X_j)$.

Nun ist $E(X_j) = \frac{1}{N}(n_{1o} + n_{11}) = \frac{n_1}{N}$; $E(X_{j+r}) = \frac{n'_1}{N}$. (3.3)

Durch Quadrieren ändern sich die Variablenausprägungen nicht: $o^2 = o$; $1^2 = 1$. Daher sind die Varianzen einfach:

$$V(X_j) = \frac{n_1}{N} - \left(\frac{n_1}{N}\right)^2 ; \quad V(X_{j+r}) = \frac{n'_1}{N} - \left(\frac{n'_1}{N}\right)^2 \qquad (3.4)$$

$$V(X_j) = \frac{n_1}{N}\left(1 - \frac{n_1}{N}\right) ; \quad V(X_{j+r}) = \frac{n'_1}{N}\left(1 - \frac{n'_1}{N}\right)$$

$$V(X_j) = \frac{n_1\, n_o}{N^2} ; \quad X(X_{j+r}) = \frac{n'_1\, n'_o}{N^2}$$

Die Kovarianz bestimmt sich folgendermaßen:

$$E(X_j)E(X_{j+r}) = \frac{n_1\, n_1'}{N^2}$$

und $$E(X_j X_{j+r}) = \frac{\sum_j x_j x_{j+r}}{N} = \frac{n_{11}}{N}\,.$$

Daher ist

$$\rho = \frac{\frac{n_{11}}{N} - \frac{n_1\, n_1'}{N^2}}{\sqrt{\frac{n_1\, n_o\, n_1'\, n_o'}{N^4}}}\,. \qquad (3.5)$$

Da sich die Häufigkeiten n_{o1} und n_{1o} höchstens um den Betrag 1 unterscheiden können, gilt

$\lim_{N\to\infty}(n_{1o} - n_{o1}) = o$; daraus folgt auch $\lim_{N\to\infty}(n_1 - n_1') = o$ und $\lim_{N\to\infty}(n_o - n_o') = o$. Für großes N ist der Korrelationskoeffizient also

$$\lim_{N\to\infty}\rho = \frac{\frac{n_{11}}{N} - \frac{n_1^2}{N^2}}{\frac{n_1\, n_o}{N^2}}$$

$$= \frac{N\, n_{11} - n_1^2}{n_1\, n_o}$$

$$= \frac{(n_1 + n_o)\, n_{11} - n_1^2}{n_1\, n_o}$$

$$= \frac{n_1\,(n_{11} - n_1) + n_o\, n_{11}}{n_1\, n_o}$$

$$= \frac{n_{11}}{n_1} - \frac{n_{1o}}{n_o} \qquad (3.6)$$

In dieser Form läßt sich der Korrelationskoeffizient am leichtesten aus den empirisch gefundenen Häufigkeitsverteilungen berechnen.

Beispiel 7:

Aus der Häufigkeitstabelle

	o	1	
o	1187	846	2033
1	847	189	1036
	2034	1035	3069

ergibt sich der Korrelationskoeffizient

$$\Phi = \frac{189}{1036} - \frac{847}{2033} = 0.23419.$$

Die Signifikanz des Autokorrelationskoeffizienten kann mit Hilfe der Beziehung

$$N\Phi^2 = \chi^2 \tag{3.7}$$

getestet werden.

3.2 AUTOKORRELATION UND MARKOV-KETTE

Die Determinante der Matrix P

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} p_{oo} & p_{o1} \\ p_{1o} & p_{11} \end{pmatrix},$$

ist

$$\begin{aligned} \det(P) &= (1-p)(1-q) - pq \\ &= 1 - p - q + pq - pq \\ &= 1 - p - q \\ &= p_{oo} - p_{1o} \end{aligned} \tag{3.8}$$

und entsprechend ist (vgl. MILLER 1951)

$$\det(P^r) = p_{oo}^{(r)} - p_{1o}^{(r)}. \tag{3.9}$$

Die Größe $\rho_1 = p_{oo} - p_{1o}$ ist die Autokorrelation der Variablen mit der Verschiebungsweite von 1. Außerdem gilt:

$$\rho_r = p_{oo}^{(r)} - p_{1o}^{(r)}. \tag{3.1o}$$

ρ_r ist also eine bei $o < \rho < 1$ monoton fallende und bei $o > \rho > -1$ gedämpft oszillierende Funktion von r.
Dieser Autokorrelationskoeffizient ist allerdings für Ketten der Ordnung $m > 1$ nicht definiert, ebensowenig für Matrizen erster Ordnung mit mehr als zwei nichtnumerischen Zuständen[6];

die Determinanten solcher Matrizen höherer Ordnung haben aber ähnliche Eigenschaften wie ρ.

Beispiel 8:

Aus der Matrix

$$P = \begin{pmatrix} o.33 & o.67 \\ o.7o & o.3o \end{pmatrix}$$

bestimmen wir die Autokorrelation der Variablen mit $\rho_1 = o.33 - o.7o = -o.37$; die Autokorrelation nach vier Schritten ist $\rho_4 = (-o.37)^4 \approx o.o19$.

Setzen wir nun in (3.8) anstelle der Wahrscheinlichkeiten deren Maximum-Likelihood-Schätzwerte ein und formen um, erhalten wir

$$\begin{aligned} \hat{\rho} &= \hat{p}_{oo} - \hat{p}_{1o} \\ &= \hat{p}_{oo} + (\hat{p}_{11} - 1) \\ &= (1 - \hat{p}_{o1}) + (\hat{p}_{11} - 1) \\ &= \hat{p}_{11} - \hat{p}_{o1} \\ &= \frac{n_{11}}{n_1} - \frac{n_{o1}}{n_o} \approx \frac{n_{11}}{n_1} - \frac{n_{1o}}{n_o} = \Phi. \end{aligned} \tag{3.11}$$

Damit haben wir eine Beziehung zwischen dem in Abschnitt 3.1 abgeleiteten empirischen Korrelationskoeffizienten Φ und dem theoretischen Koeffizienten ρ gefunden.

Zu beachten ist, daß Φ im Unterschied zu ρ für jedes r aus den Häufigkeitstabellen der beobachteten Variablenpaare (X_j, X_{j+r}) gesondert berechnet werden muß.

3.3 TEST DES AUTOKORRELATIONSKOEFFIZIENTEN

Unter der Voraussetzung, daß eine beobachtete Folge von Merkmalsausprägungen die Realisation einer Markov-Kette erster Ordnung ist, sollte man erwarten, daß die empirischen Werte Φ_r von den aus der Matrix berechneten ρ_r nicht sehr verschieden sind. Um festzustellen, ob die Schwankungen von Φ um ρ zufällig sind, verwenden wir folgenden Test:

Für große N und kleine ρ ist Φ normalverteilt (vgl. HALD 1967:6o9). Transformieren wir folgendermaßen auf eine u-Variable, können wir die Abweichungen durch die N(o,1) Verteilung testen:

$$u = \frac{\Phi-\rho}{1-\rho^2} \cdot \sqrt{N-1} \qquad (3.12)$$

d.h. bei $|u| > 1.96$ lehnen wir mit einer Irrtumswahrscheinlichkeit von o.o5 die Nullhypothese ab, daß $\Phi = \rho$ ist.

Beispiel 9:

Haben wir z.B. bei dem phonematischen Merkmal "coronal" in einem deutschen Korpus einen empirischen Autokorrelationskoeffizienten Φ_3 = o.o5oo31496 für die Verschiebungsweite r = 3 bei N = 192o Beobachtungen festgestellt, und ist der theoretische[7] Koeffizient aus der Matrix der Übergangswahrscheinlichkeiten ρ_3 = -o.oo6o68813, dann testen wir die Nullhypothese

$$H_o : \Phi_3 = \rho_3 \quad \text{gegen} \quad H_1 : \Phi_3 \neq \rho_3$$

mit

$$u_3 = \frac{o.o5oo31496 - o.oo6o68813}{1 - (-o.oo6o68813)^2} \cdot \sqrt{1919}$$

$$= 2.457642655.$$

Wegen $|u_3| > 1.96$ lehnen wir die Nullhypothese ab: der empirische Autokorellationskoeffizient unterscheidet sich signifikant vom theoretischen, d.h. die Abweichung ist nicht zufällig.

4. MASCHINELLE KORPUSANALYSE

Quantitative Modelle verlangen in der empirischen Arbeit einen so großen Daten- und Rechenaufwand, daß in den meisten Fällen die Benutzung einer elektronischen Rechenanlage sinnvoll ist.

Liegt ein zu untersuchendes Korpus in maschinenoperabler Form auf Datenträger vor und kann eine Tabelle der verwendeten Symbole mit Zuordnung zu den Merkmal(bündel)sausprägungen eindeutig erstellt werden, so kann für die in diesem Beitrag dargestellten Verfahren das vom Verfasser erstellte Programmsystem AUTOMARK verwendet werden, dessen Einzelfunktionen mit Hilfe einfacher Anweisungen im Dialog steuerbar sind. Das in FORTRAN geschriebene Programmsystem steht wissenschaftlichen Anwendern auf Anfrage zur Verfügung.

Ein ausführliches Beispiel soll abschließend veranschaulichen, wie eine maschinelle empirische Korpusanalyse mit Hilfe von Autokorrelation und Markov-Modell durchgeführt werden kann. Als Gegenstand der Untersuchung dient das dynamische Verhalten phonematischer Merkmale des Deutschen.

Für die Beschreibung des Phonemsystems benutzen wir die distinktiven Merkmale von CHOMSKY und HALLE (1968:298ff). Die Kürzel bedeuten:

son	sonorant	rnd	round
voc	vocalic	lat	lateral
cns	consonantal	lng	length
cor	coronal	cnt	continuant
high	high	back	back
low	low	tns	tense
ant	anterior.		

Da wir diese Merkmale nicht als ausschließlich distinktiv, sondern darüberhinaus als konstitutive Komponenten des linguistischen Konstrukts "Phonem" [8] betrachten, sind in der Phonem-Merkmal-Matrix (anstelle der sonst üblichen Nullen) die redundanten Merkmalsausprägungen enthalten[9] (s. Abb.7). Ein Phonem kann nun

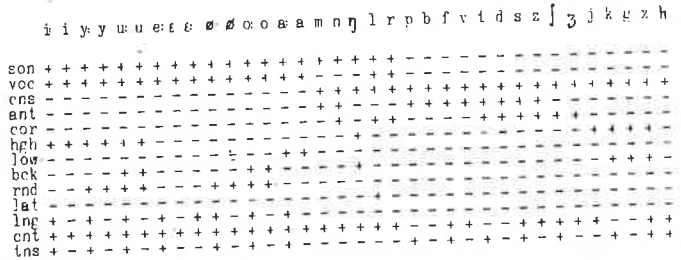


Abb.7 Phonem-Merkmal-Matrix für das Deutsche

als ein n-stelliger Vektor (eine Spalte der Matrix mit den Ausprägungen der n - in unserem Fall 13 - Merkmale) angesehen werden, eine Phonemfolge als Folge von Vektoren in der Zeit bzw. im Raum (s. Abb.8). Außer in der Phonem-Merkmal-Matrix, wo wir die gebräuchlichen Symbole "+" und "-" verwendet haben, stehen wieder "1" und "o" für das Zutreffen bzw. Nichtzutreffen eines Merkmals.

Das Programmsystem AUTOMARK setzt nun mit Hilfe der Phonem-Merkmal-Matrix eine in maschinenlesbare Symbole transkribierte Phonemfolge - für jedes Merkmal gesondert - in eine Folge von binären Merkmalsausprägungen um, die als Realisation einer Markov-Kette angesehen wird (vgl. Abb.9). Auf diese Kette werden die oben beschriebenen Verfahren angewendet. Als Ergebnis druckt das Programm Häufigkeits- und Wahrscheinlichkeitstabellen für die Übergänge[10], Häufigkeitstabellen der empirischen Autokorrelation, den Test der Ordnung der Markov-Kette, die empirischen und die theoretischen Korrelationskoeffizienten, sowie den Test der Abweichung (den Wert der u-Variablen und das Signifikanzurteil). Außerdem wird auf dem Drucker eine graphische Darstellung der

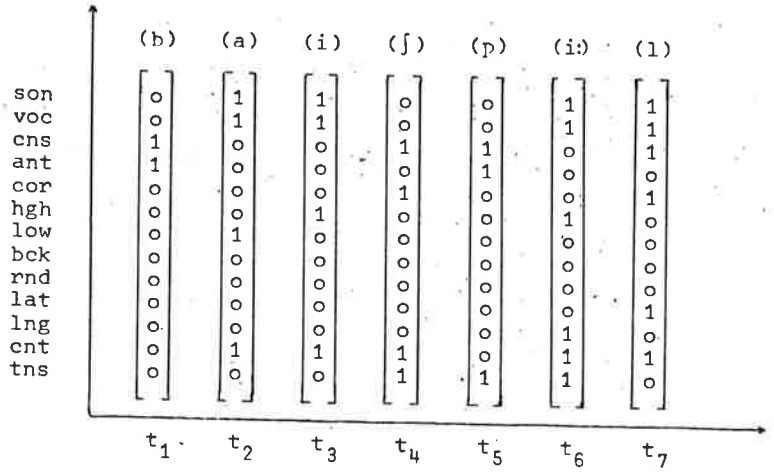


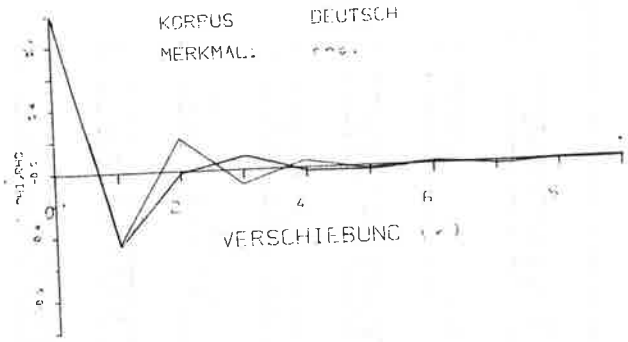
Abb.8 Die Phonemfolge /baiʃpi:l/ als Folge von Vektoren in der Zeit

empirischen Autokorrelationsfunktion erzeugt. Falls ein Plotter vorhanden ist, kann auf diesem Gerät eine graphische Ausgabe von empirischer und theoretischer Autokorrelation erfolgen. Um die Übersichtlichkeit zu erhöhen, werden die beiden Funktionen farblich unterschieden, und die zum Graphen der Funktionen gehörenden Punkte werden miteinander verbunden.

(fernerʃpraxerainmanhatetsvaizœne)
1o11o1111o1o1oo11o11o1o111oo1o1o

Abb.9 Ausschnitt aus einer Phonemfolge und der dazugehörigen binären Markov-Kette für das Merkmal cns

Im Anschluß ist exemplarisch ein Programm-Output für das Merkmal cns ("consonantal") in einem Deutschen Textkorpus wiedergegeben.



KORPUS : DEUTSCH
MERKMAL: 3 (cons.)

1111	1110	1101	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001	0000
15	16	143	5	236	26	32	0	85	42	115	2	45	1	1	0 | 0
1	84	150	52	94	232	45	7	15	239	149	25	6	32	6	0 | 1

HAEUFIGKEITEN

	0	1	
000	7	7	7
001	33	51	84
010	27	264	291
011	281	100	381
100	7	77	84
101	258	330	588
110	57	323	380
111	100	16	116
	763	1168	1931

WAHRSCHEINLICHKEITEN

	0	1
000	0.0000	1.0000
001	0.3929	0.6071
010	0.0928	0.9072
011	0.7375	0.2625
100	0.0833	0.9167
101	0.4388	0.5612
110	0.1500	0.8500
111	0.8621	0.1379

HAEUFIGKEITEN

	0	1	
00	7	84	91
01	291	381	672
10	84	588	672
11	380	116	496
	762	1169	1931

WAHRSCHEINLICHKEITEN

	0	1
00	0.0769	0.9231
01	0.4330	0.5670
10	0.1250	0.8750
11	0.7661	0.2339

HAEUFIGKEITEN

	0	1	
0	91	672	763
1	672	496	1168
	763	1168	1931

WAHRSCHEINLICHKEITEN

	0	1
0	0.1193	0.8807
1	0.5753	0.4247

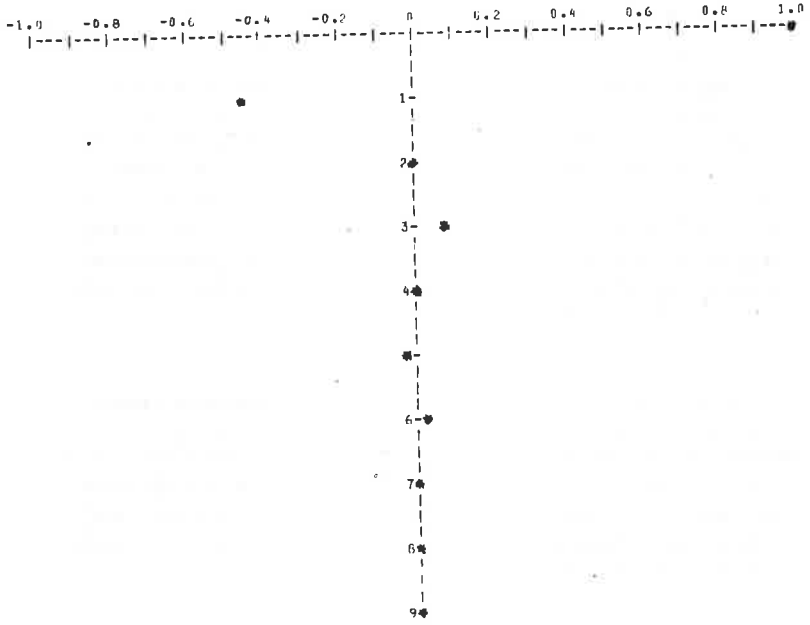
MATRITZEN ERSTEN ORDNUNG FUER VERSCHIEBUNGEN VON 0 BIS 9 (EMP.HAEUFIGKEITEN):

00	91	91	298	341	290	286	308	291	305	306
01	672	672	465	422	472	474	454	470	456	455
10	672	673	465	422	473	474	454	471	457	456
11	496	497	704	746	695	693	712	695	708	708

ORDNUNG DER KETTE M = 3. CHI-QUADRAT = 14.100

TEST DER ORDNUNG:

2<I> [0/1] = 441.0873413086 DF = 1
2<I> [1/2] = 156.0664062500 DF = 2
2<I> [2/3] = 14.2460937500 DF = 4
2<I> [3/4] = 13.0351562500 DF = 7



KORPUS : DEUTSCH MERKMAL: 3(cons.) MODUS: 1 (WORTGR. IGN.)

AUTOKORRELATIONEN:	ANZ. DER BEOBACHTUNGEN	THEOR. KOEFF.	TEST	
PHI(0) = 1.000000000000	N = 1.0	RHO = 1.000000000000	U = 0.0000	
PHI(1) = -0.465947637858	N = 1933.0	RHO = -0.465947637858	U = 0.0000	
PHI(2) = -0.097212264403	N = 1932.0	RHO = -0.207588245163	U = -9.8791	**SIGNIFIKANT**
PHI(3) = 0.085618440656	N = 1931.0	RHO = -0.094796167145	U = 7.9974	**SIGNIFIKANT**
PHI(4) = -0.024388273515	N = 1930.0	RHO = 0.042217717436	U = -2.9748	**SIGNIFIKANT**
PHI(5) = -0.020216788441	N = 1929.0	RHO = -0.019704526759	U = -0.5739	
PHI(6) = 0.014834154193	N = 1928.0	RHO = 0.005584412998	U = 0.2568	
PHI(7) = -0.021535181457	N = 1927.0	RHO = -0.011696823276	U = -0.7661	
PHI(8) = 0.006513757959	N = 1926.0	RHO = 0.011677753696	U = 0.2916	
PHI(9) = 0.010349555413	N = 1925.0	RHO = -0.023051597486	U = 0.4913	

ANMERKUNGEN

1 Wenn in anderem Zusammenhang Zweifel über eine solche Zeitunabhängigkeit bestehen, läßt sich diese Eigenschaft der Markov-Kette testen. S. dazu z.B. ANDERSON/GOODMAN 1957:97ff.
Ein Beispiel für die Anwendung von nicht-homogenen Markov-Ketten in der Metrik findet man in GROTJAHN 1979:213

2 In den Fällen der Tests für die Wahrscheinlichkeiten und für die Ordnung von Markov-Ketten können stets χ^2- und ω^2-Tests einerseits, sowie Likelihood-Ratio-Tests andererseits verwendet werden. Beide Möglichkeiten haben Vor- und Nachteile, die nicht ohne Weiteres gegeneinander abwägbar sind. Wir entscheiden uns hier mit KULLBACK/KUPPERMAN/KU 1962 für den 2I-Test, der asymptotisch χ^2-verteilt ist und gegenüber χ^2 rechnerische Vorzüge besitzt.

3 s. KULLBACK/KUPPERMAN/KU 1962:598, 6o5

4 Zur Vereinfachung lassen wir in den Tabellen die Verdopplung des mittleren Zustands weg, sodaß die Null-Häufigkeiten und -Wahrscheinlichkeiten in der Darstellung entfallen.Die Tabellenzeilen sind entsprechend eingerückt.

5 vgl. z.B. NEWMAN 1951b, 1952, GROTJAHN 1979

6 Die numerische Interpretation der Ausprägungen "O" und "1" einer nominalskalierten Zufallsvariablen führt nur im dichotomischen Fall nicht zu Widersprüchen.

7 Es sollte betont werden, daß der aus der Matrix berechnete Koeffizient nur dann der theoretische Autokorrelationskoeffizient ist, wenn die Übergangswahrscheinlichkeiten in der Matrix keine mit Stichprobenfehlern behafteten Schätzwerte sind, d.h. wenn die Matrix aus theoretischen Überlegungen stammt oder aus der Grundgesamtheit gewonnen wurde.

8 vgl. hierzu ALTMANN/LEHFELDT 198o:52ff

9 Zum Auffüllen einer nichtredundanten Matrix siehe ALTMANN/LEHFELDT 1973:74

1o Die erste Tabelle ist aus Platzgründen um $9o^{o}$ im Uhrzeigersinn gedreht.

LITERATUR

Altmann, G. / Lehfeldt, W.
1973 Allgemeine Sprachtypologie. München
198o Einführung in die Quantitative Phonologie. Bochum

Altmann, G. / Diller, / Sappok, / Strauß, U.
Wiederholungen in der Poesie. erscheint

Anderson, T. / Goodman, L.A.
1957 Statistical Inference about Marcov Chains. Annals of Mathematical Statistics 28, 89-11o

Bartlett, M.S.
1978 An Introduction to Stochastic Processes with Special Reference to Methods and Applications. Cambridge

Bhat, U.N.
1972 Elements of Applied Stochastic Processes. New York

Brainerd, B,
1976 On the Marcov Nature of Text. Linguistics 176, 5-3o

Chomsky, N.
1957 Syntactic Structures. The Hague

Chomsky, N. / Halle, M.
1968 The Sound Pattern of English. New York

Grotjahn, R.
1979 Linguistische und statistische Methoden in Metrik und Textwissenschaft. Bochum

Hald, A.
1967 Statistical Theory. New York

Hoel, P.G.
1954 A test for Markoff chains. Biometrika 41 43o-433

Ku, H.H.
1963 A Note on Contingency Tables Involving Zero Frequencies and the 2$\uparrow$ Test. Technometrics 5 398-4oo

Kullback, S. / Kupperman, M. / Ku, H.H.
1962 Tests for Contingency Tables and Marcov Chains. Technometrics 4.4, 573-6o8

Levelt, W.J.M.
1974 Formal Grammars in Linguistics and Psycholinguistics. The Hague

Miller, G.A.
1951 Finite Marcov Processes in Psychology. Psychometrica 17.2, 149-167

Newman, E.B.
1951a Computational Methods Useful in Analyzing Series of Binary Data. American Journal of Psychology 64, 252-262
1951b The Pattern of Vowels and Consonants in Various Languages. American Journal of Psychology 64, 369-379
1952 A New Method for Analyzing Printed English. Journal of Experimental Psychology 44, 114-125

Osgood, Ch.E.
1963 On Understanding and Creating Sentences. American Psychologist 18, 735-751

Suppes, P.
197o Probabilistic Grammars For Natural Languages. Synthese 22, 95-116

Markierungen paraphonetischer Information: Kurventypen, Kombinationen und Strukturen

Peter Winkler, Konstanz

Summary

Emphatic sequences have been selected from recordings of a dyadic conversation in which striking combinations of the acoustical parameters pitch, intensity, speed, and periodicity occured. In classifying these combinations and calculating the probability of certain patterns of combinations a typical process can be discovered: First, the speaker has to produce the sound pattern (or "neutral" combination) which is necessary for the phonetic realization of a phonem. Within a span of milliseconds, this sound structure will be changed into another combination and again rechanged into the next phonemic sound pattern. The whole structure of the utterance, the gestalt, will not be destroyed by this fluent changes, e.g. the intonation contour is audible as a integrated something. But the utterance sounds 'extraordinarly', 'emphatically'. The redundancy of the linguistic-phonemic sound structure is occupated to express the paraphonetic content. The changements work at the segmental or the sub-segmental level. The paraphonetic information consists in this changement, not in an acoustical substance (for example, no other features than those which can be used for constituting the phoneme, too, are added). The paraphonetic information has no 'slow development', but is rather a short-timed insert into the neutral parts of phonemic segments. The changements start from the phonemically prestructured configuration, and the parameters return to this configuration after the displacement. The results indicate some prototypical combinations and internal structures of paraphonetic vs. linguistically determined sound information.

1. *Phonetische Markierungen emphatischer Äußerungen*

Der Gegenstand der folgenden Analyse soll durch zwei Beispiele demonstriert werden: Abb. 1 zeigt die Signalparameter eines emphatisch gesprochenen Satzes. Es handelt sich um eine Frage, die im colloquial style während eines nicht sehr offiziellen Gespräches von einem Sprecher an eine Gesprächspartnerin gerichtet wurde ("Was mußtest Du?"). Inhaltlich scheint diese Frage höchst belanglos zu sein; auch die Intonationskontur, die

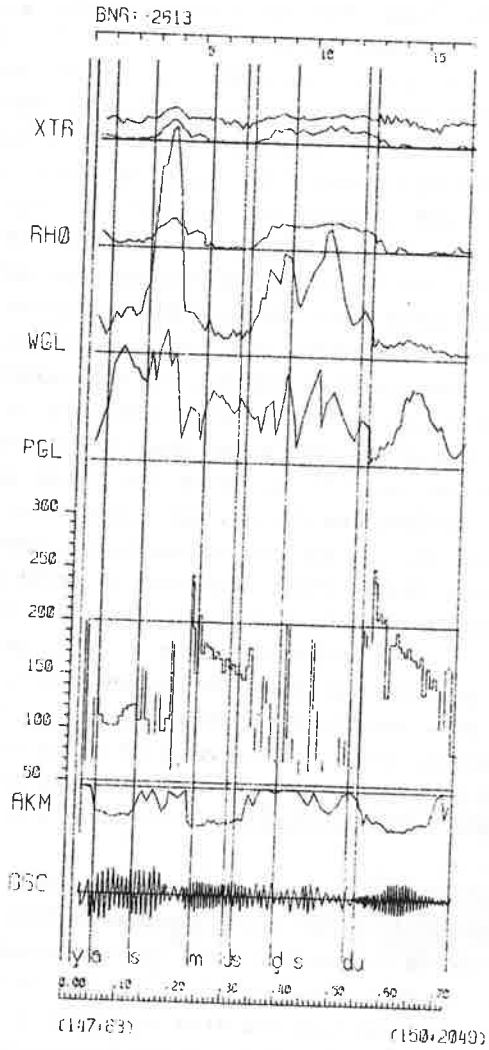


Abb. 1: Akustische Parameter einer emphatischen Äußerung (Sprecher)

häufig dem Hörer Zusatzinformationen gibt, ist in diesem Fall nicht weiter aufschlußreich. Sie ist eine Kombination aus einem steigenden und zwei fallenden Teilen: [ɣ/as\mosʁsd\u]. Diese Form kann ebensogut bei speziellen Fragen auftreten, die keinen emphatischen Charakter haben (z.B. "/Haben Sie das \Buch 'Der weiße \Hai'?"). Signal- und ohrenphonetisch kann die Besonderheit dieser Frage etwas näher beschrieben werden; die Abb. 1 enthält die Signalparameter AKM (Maximum der Autokorrelationsfunktion), den resultierenden F_o, den Pegel (PGL), die interne 'Geschwindigkeit' (WGL = Weglänge des Signals), die Nulldurchgangsdichte (RHØ) und die Extremwerte (XTR); dazu das Pseudooszillogramm (jede 1o. Schalldate) sowie eine ohrenphonetisch festgelegte Segmentierung und Transkription. Der Sprecher beginnt den aufsteigenden Teil des Grundtones (der F_o ist bei stimmhaften Anteilen in Form einer zusammenhängenden, treppenartigen Kurve gezeichnet, sonst als irreguläre Strichfolge) bei ca. 11o Hz, senkt die Stimme kurz auf 1oo Hz und hebt sie sofort bis 12o Hz. Der zweite Teil beginnt mit einem Sprung auf 18o Hz, vor dem ein kurzzeitiges, noch höher plaziertes Einsetzen der Stimme liegt (25o Hz); er endet bei einer Frequenz von 155 Hz. Der dritte Teil beginnt bei 19o Hz und hört mit 14o Hz auf (der in der Kurve sichtbare Vorlauf bei 25o Hz kann noch zum /d/ gehören, das ohrenphonetisch jedoch völlig anders 'geortet' worden ist). Innerhalb der stimmhaften Passagen gibt es einige Unterbrechungen, die koartikulatorisch nicht nur nicht notwendig, sondern besonders 'aufwendig' sind (gemeint sind die F_o - Unterbrechungen in dem Zeitabschnitten 58. - 6o. msec und 61.-69. msec [\dᵕu]).

Im Anfangsteil des Satzes (in der 2. - 13. msec) ist die Periodizität größer O, d.h. der Vokal enthält nichtperiodische Anteile. Das nachfolgende /s/ wird schnell und mit einem ausgeprägten breitbandigem Rauschen ausgesprochen, das gut in Abb. 2 dokumentiert ist (diese Abbildung ist eine intensivere Analyse des Satzes). Der Tiefpaß zeigt nach dem vierten Zeitabschnitt (1 Skalenstrich = 1ooo Daten bei f_s = 2oKHz = 1/2o sec) keinen Grundton mehr an, der Pegel sinkt und die Geschwindigkeit steigt. Das Spektrum (DFT) indiziert ein Verschwinden der Vokalformanten des /a/, das besagte breitbandige Rauschspektrum baut sich auf, das

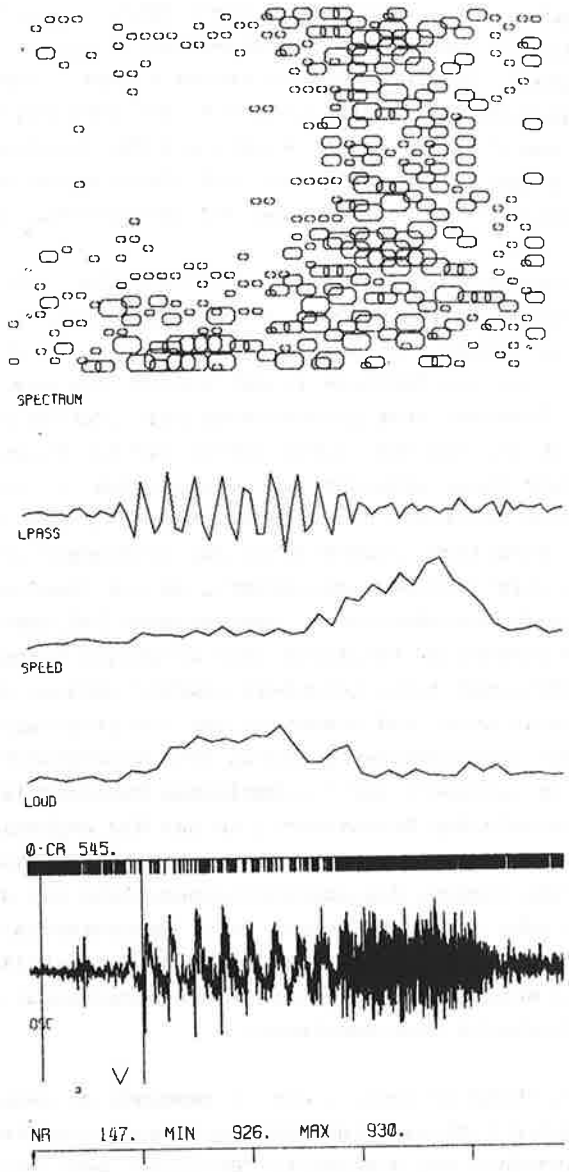


Abb. 2 Anfang der emphatischen Äußerung Abb. 1

für einen gewöhnlichen S-Laut sehr tiefe Anteile aufweist (und dadurch auffällt); normalerweise konzentrieren sich die Spektralanteile im oberen Frequenzbereich, dieser s-Laut enthält zugleich eine Behauchung.Mitten in der Aussprache der Kombination /a/ + /s/ "überfällt" den Sprecher ein Atemstoß, der der Äußerung - per natürlicher Kompetenz interpretiert und global-generisch zugeschrieben - den Anschein von Verwunderung und Überraschung gibt.

Die beschriebenen Strukturen liegen eindeutig unterhalb der linguistischen bedeutsamen Zeit- und Klangrelationen, z.T. in Minimalanteilen eines phonematischen Segmentes. Die Erfahrung ist nicht neu: LIEBERMAN, MICHAELS (1962) hatten Versuche durchgeführt, in denen der Grundton über verschiedene Zeitstrecken hinweg geglättet wurde. Erst ab ca. 1oo msec verschwindet für den Hörer der paraphonetische Gehalt einer Äußerung. Im Beispiel Abb. 2 sind es ca. 1o msec, innerhalb derer die nichtlinguistische Information kodiert ist. Bei linguistisch -phonetischen Untersuchungen werden solche Details berechtigterweise ausgefiltert, um die Schallstrukturen in der neutralen Realisierung zu analysieren. Daß durch die Minimalphänomene personale, situative oder affektive Besonderheiten ausgedrückt werden, muß nicht besonders erwähnt werden. Man kann jedoch noch weiter gehen und annehmen, daß die *Bedeutung* einer Aussage durch die Substrukturen geradezu erst konstituiert wird. Das steht etwas im Gegensatz zur herkömmlichen Paralinguistik (und zu einigen Positionen der Semantiker), in der die emphatischen, nichtneutralen phonetischen Formungen als zusätzliche, fakultative Elemente aufgefaßt werden, die strenggenommen nicht für das Verständnis eines Satzes oder Wortes notwendig sind. Andererseits wird durchaus eingeräumt, daß eine Wortbetonung, eine phonostilistische Variante oder die Wahl einer speziellen Intonationskontur den Inhalt ändert,unterstreicht oder umgekehrt.

Diese Art "tone of voice", der - gemessen am demonstrierten Beispiel - immer noch auf einer Art Makroebene angesiedelt ist, ist hier nicht gemeint. Die sogenannte "Betonung" oder eine besondere Pitch-Kontur sind in der beschriebenen Aussage gar nicht so entscheidend; es sind vielmehr die versteckten Details einer sonst geläufigen

phonetischen Grobstruktur. Die erste deutliche Markierung, die der Sprecher setzt und damit andeutet, daß seine Frage nicht ohne Hintergründigkeit ist, ist der große Sprung zwischen Teil 1 und Teil 2 des Grundtones (ein Abstand von ca. 8o Hz). Zusätzlich wird der folgende Laut behaucht, was möglicherweise ein bescheidener und sofort unterdrückte Anlauf zu einem Lachen gewesen ist. Der Rhythmus ist agogisch, besonders in der Verbindung von /s/ und /m/ sowie durch die Verlängerung von [ʂs], das zusammen mit dem auslautenden /u/ das längste Segment das Satzes bildet.

Die "eigentliche" Bedeutung der Frage ist, wenn man sie hört, sofort evident. Man kann sie paraphrasieren:"Ich kann mir gar nicht vorstellen, was für exotische Dinge mit dir in dem psychologischen Test angestellt worden sind; außerdem habe ich ein Wort nicht verstanden, wiederhole doch die angedeutete erheiternde Monstrosität.

Mit einem Minimum an Aufwand ist hier ein Maximum an Information gegeben worden. Die paraphonetische Einkleidung ist nicht in jedem Fall mit so drastischer bedeutungsbildender Funktion versehen, obwohl in jeder gesprochenen Äußerung die Möglichkeit dazu besteht. Die in der Sprache abgelagerte, allgemeine, vom Situationsbezug einigermaßen befreite "neutrale" Bedeutung wird im gesprochenen Text neu und erneut spezifiziert verwirklicht. Semantische und paraphonetische "Bedeutung" verhalten sich zueinander nicht immer gleichgerichtet. Paraphonetisch kann der abstrakte Bedeutungsgehalt kontrastiert werden. Nicht nur der Inhalt kann 'zerstört' werden, auch die akustische Form unterliegt dem Wechsel. Die phonetischen Mittel, die dazu zur Verfügung stehen, sind die gleichen, die auch der Phonemverwirklichung eigen sind. Die nicht-linguistischen Intentionen des Sprechers führen nicht dazu, daß neue phonetische 'features' auftreten, sondern dazu, daß die sprachge-

wohnte Kombination der gleichen akustischen Parameter durchbrochen wird. (Die entgegengesetzte Auffassung lautet sinngemäß: Mit Hilfe vieler anderer paralinguistischer Elemente übt die Intonation ihre expressive Funktion aus.) Das kann z.B. durch Überlagerung vokalischer Spektren mit konsonantischen Anteilen geschehen. So gesehen sind linguistische und paraphonetische 'Materie' ein- und derselbe Stoff; wissenschaftlich abtrennbar sind einerseits die *Funktion*, andererseits die *strukturellen Veränderungen*, die mit der Funktion verbunden sind.

Die der Arbeit zugrundeliegende Hypothese ist:
Die Funktionsänderung - der Wechsel zwischen linguistisch-phonematischer und paraphonetischer Belastung der akustischen Manifestationen - wird mittels kurzfristiger Änderungen von 'Normal' - Strukturen durchgeführt. Sie besteht im Oszillieren zwischen der sprachzeichenbedingten und einer anderen Schallkonfiguration. Die paraphonetischen Markierungen sind keine substanzimmanenten und von der phonetischen Laut-Struktur getrennten, eigenen "Register" des Schalles (mit der Ausnahme des Stimmklanges). Die paraphonetischen Strukturen lösen während kurzer Zeitspannen die sprachlich notwendigen Konstellationen ab, ersetzen die alte Zusammenstellung durch eine neue, indem sie sie bis zur völligen Auflösung verändern, und werden ihrerseits von phonematischen und intonematischen Wiederaufnahmen abgelöst, bis diese erneut verschwinden und neuen paraphonetischen Markierungen Platz machen. Die Ausgangslage, die Startposition, sind die Laut-Strukturen, deren Parameterarrangement geändert und in eine neue Gestalt übergeführt wird (die allerdings instabil ist und häufig nicht einmal das zeitliche Ausmaß eines Segmentes erreicht). Welche Kombinationstypen auftreten und welche Richtung die Parameterverschiebung einschlägt, ist möglicherweise für einen Ausdrucks- oder Bedeutungstyp kennzeichnend; "Ärger" wird durch eine andere Kombination ausgedrückt als "Freude".

Bevor an 25 ausgewählten emphatischen Sequenzen die Kombinatorik akustischer Parameter untersucht werden soll, eine weitere Demonstration des Analysematerials. Abb. 3 ist die

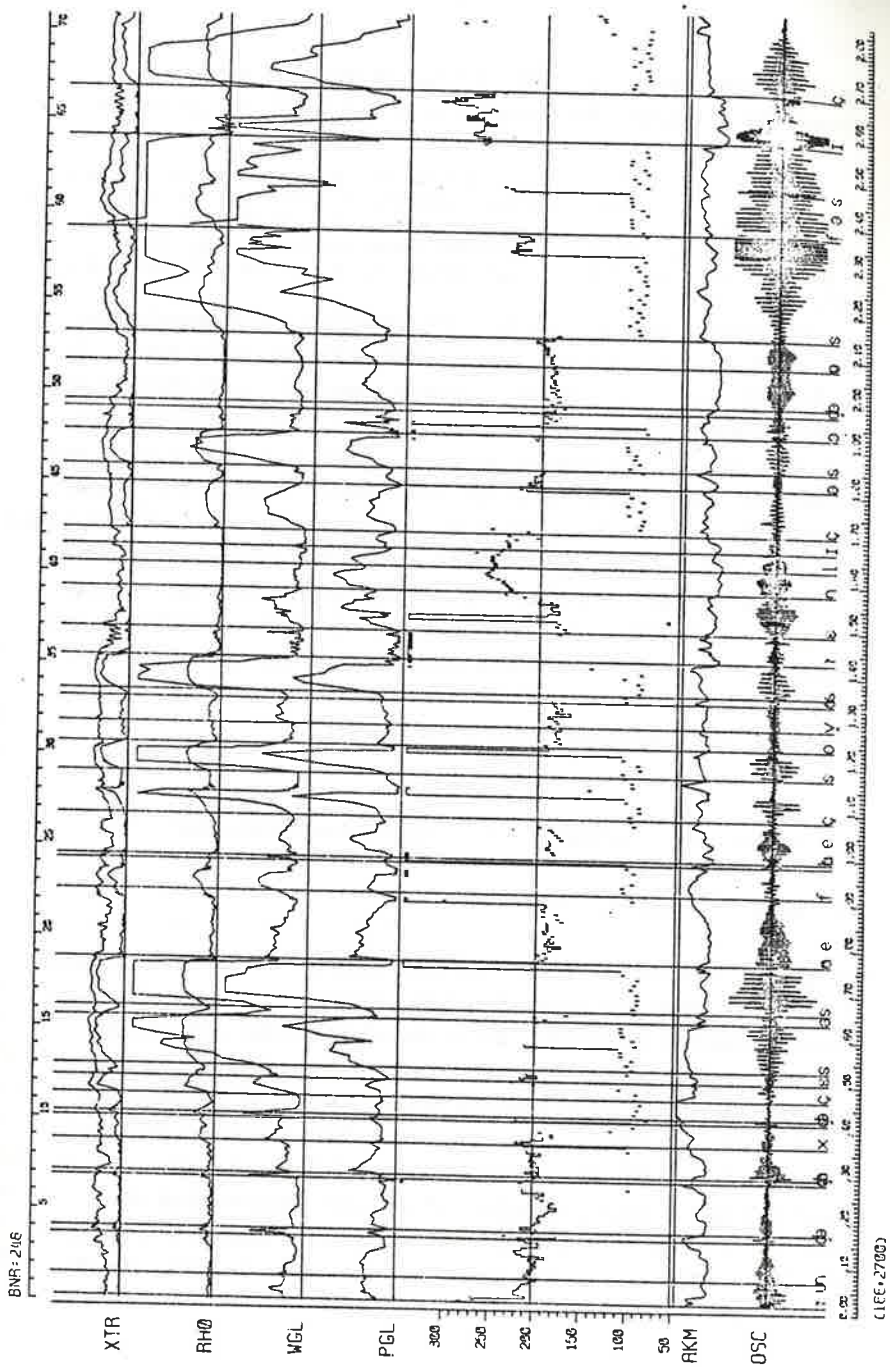


Abb. 3 Akustische Parameter einer emphatischen Äußerung (Sprecherin)

Registrierung eines Satzfragmentes, in dem die paraphonetischen Markierungen über mehr als ein Segment verteilt sind.

Es handelt sich um den Endteil des Satzes:"Und da dachte ich, es sei vielleicht so etwas ähnliches oder was weiß ich..."; *er wird ironisch, belustigt gesprochen; paraphrasiert:"Vielleicht ist dieser Versuch genauso lächerlich wie der andere;* *zu befürchten ist, daß er noch ärger wird ... es ist kaum auszumalen."*

Die Aussprache ist phonetisch reduziert, umgangssprachlich gefärbt, der Endteil wird verschliffen zu: [ɔgəɒsfəsɪç/ʔə]. Die Silben /er/+/va/ werden zu einem einzigen, vokalisch aufgelösten R-Laut zusammengezogen; /vae/ wird zu einem stimmlosen Reibelaut mit nachfolgendem schwachtonigem /e/ umgebildet. Der Satzakzent liegt auf /ich/, das sehr hoch (265-3oo Hz, Frauenstimme) und intensiv ausgesprochen wird; Grundton steigen-fallend. In die beiden S-Laute vor /f/ und nach /fe/ werden stimmhafte Passagen eingefügt (im Bild nicht transkribiert, aber im Meßprotokoll sichtbar); kleine Unterbrechungen des stimmlosen Kontinuums (ca. 23o Hz) mit hohem Schalldruck (Oszillogramm geclippt) und hoher Geschwindigkeit; mit Eintreten der Stimmhaftigkeit sinkt gesetzmäßig der Rauschanteil.

2. *Statistische Analyse der Parameterkombinationen*

2.1 *Material und Methoden*

Material[1]: Dyadische, freie Interaktion, aufgezeichnet in normalhalligem Raum mit separaten Mikrofonen für jeden Sprecher (Beyer Dynamic mit Richtwirkung; NAGRA Recorder 9,5 cm/s).
Phonetische Methoden: Computerunterstütze Segmentation und Transkription; digitale Signalanalyse,PDP-11/5o (Anlage und Software des Münchner Instituts für Phonetik und sprachliche Kommunikation, Leitung: Prof. Dr. H.G. Tillmann).
Selektion und Klassifikation: 245 auditiv abgegrenzte Segmente aus 25 emphatischen Äußerungen beider Interaktionspartner. Vergleichssample: 244 Segmente aus neutralen Sequenzen der gleichen Aufnahmen.

Die zugehörigen Kurvenplots wurden unter auditiver Kontrolle in sieben Kurventypen klassifiziert

- ↗ Kurvenanstieg
- → gleichbleibend und/oder Ø
- ↘ Kurvenabfall
- ∿ Anstieg und Abfall im Segment
- ∪ Abfall und Anstieg im Segment
- ↧ abruptes Sinken oder Beendigung im Segment
- ↥ abruptes Steigen oder Anfang im Segment

(jeweils ohne Rücksicht auf die Plazierung innerhalb des Segmentes)

Absolute Höhendifferenzen, z.B. Hoch- oder Tiefton, Frequenz-oder Lautstärkemessungen wurden nicht verwertet, weil die gesuchten Veränderungen als dynamische Entwicklungen auftreten; mehr relativ im Verlauf als absolut in der Maßzahl.

Zur Verdeutlichung der Klassifizierung einige Beispiele: Eine Kombination ↥ (F_o) + ↗ (PGL) + →(WGL) + ↧ (RHØ) symbolisiert eine im Segment abrupt einsetzende Stimmhaftigkeit mit Lautstärkeanstieg bei gleichbleibender Geschwindigkeit und abrupter Abnahme des Geräuschanteils.↗ + ∿ + ∿ + ↥ bezeichnet einen steigenden F_o, An- und Abschwellen der Lautstärke und der Geschwindigkeit sowie einen abrupt einsetzenden Geräuschanteil.

Der Algorithmus für die Klassifikation bestand in drei Entscheidungen:

a) Parameterkonfiguration notwendig für Laut oder Intonation?
(In Abb. 4 ist das Wort "was", gesprochen vom gleichen Sprecher wie die Aufzeichnung Abb.2, analysiert. Dieses Beispiel entpricht der neutralen Aussprache: Das Segment /v/ mit der Kombination F_o → PGL ↧ , WGL ↧ , RHØ ↘ wurde als neutral eingestuft und in das Kontrollsample eingeordnet.)

b) Koartikulatorische Phänomene?
(Das Segment /a/ in Abb.4 ist ein solcher Fall, in dem der Anstieg der Geräuschhaftigkeit nicht durch Emphase, sondern durch den nachfolgenden Konstriktiv bewirkt wird; es gehört daher in die neutrale Kategorie.)

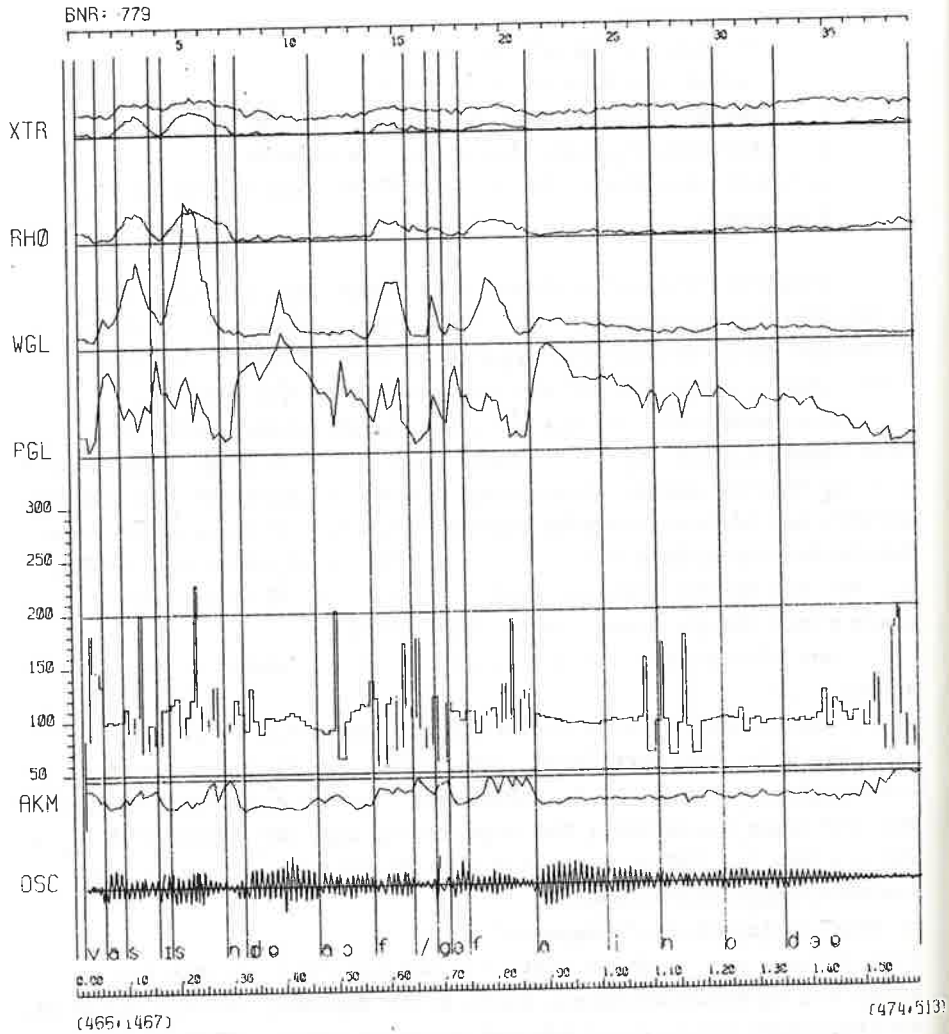


Abb. 4 Neutrale Äußerung
(Sprecher)

c) Wenn a) und b) nicht zutrafen, wurde über die Zugehörigkeit zu den sieben Kurventypen entschieden, wobei das gesamte Segment beurteilt wurde.

In Abb. 1 treten sowohl neutral als auch emphatisch gesprochene Segmente auf:

	/m/ neutral	/a/ emphatisch
F_0	↘	↗
PGL	↘	~
WGL	↘	↗
RHØ	→	→

Methodologisches Prinzip der Klassifizierung: Die auditiv-impressionistische Entscheidung über "emphatisch" oder "neutral" basiert auf dem Grundsatz der Auffälligkeit sowie dem Rekurs auf die empirische Anschauung (begründete Behauptung; warranted assertion). Die Analyse beruht mithin darauf, daß bestimmte Änderungen "auffällig", "ungewöhnlich" erscheinen, sie belegt nicht die Auffälligkeit an sich. Mit diesem Prinzip lassen sich sehr gut typische Merkmale herausarbeiten; beispielsweise werden koartikulatorische Erscheinungen sehr genau ausgesondert - denn diese sind sehr "vertraut", unauffällig, gewohnt. Für einen deutschen Hörer wäre im Gegenteil gerade auffällig, wenn eine koartikulatorische Verbindung unterlassen wird, wenn z.B. /z/ nach stimmlosen Lauten vollständig stimmhaft gesprochen würde (F_0 → statt ↗). Der auditive Eindruck wird zugleich an der Meßkurve geprüft; ebenfalls mit den Kriterien der Plausibilität. Beispielweise zeigt /m/ in Abb. 1 die vertrauten, normalen Kurvenbilder; alle Verläufe sind gleichgerichtet und RHØ ist erwartungsgemäß nahe O.

3. Ergebnisse

Die 244 Segmente der neutralen Aussprache enthielten 44 offene Vokale, 43 geschlossene Vokale, 25 stimmhafte Klusile, 9 stimmlose Klusile, 2o stimmhafte Konstriktive, 52 stimmlose Konstriktive, 32 Nasale und 15 Liquide. Die emphatischen Segmente lassen sich nicht eindeutig in diese Lautgruppen einordnen, weil z.B.stimmlose Laute mit stimmhaften Elementen durchsetzt waren, Vokale von frikativen Anteilen überlagert wurden usw.

Alle Kurventypen und Kombinationen wurden nach der Verteilungshäufigkeit aufgelistet und verrechnet.

3.1 Korrelationen

Tabelle 1: Korrelationsmatrizen der Kombinationen neutrale vs. emphatische Aussprache
(kleine Ziffern = neutral; signifikante Unterschiede von r sind unterstrichen; α = .o1)

4 Parameter über Kurventypen 1-7

	F_o	PGL	WGL	RHØ
F_o	-			
PGL	-.24 -.4o	-		
WGL	o.o -.25	.75 .98	-	
RHØ	.51 .7o	.o3 .3o	.65 .46	-

6 Kurventypen über 4 Parameter (↑ + ↓ zusammengefaßt)

	↗	→	↘	⇝	↺	↕
↗	-					
→	-.61 -.98	-				
↘	.o8 .35	-.83 -.19	-			
⇝	-.16 .82	-.7o -.91	.21 -.23	-		
↺	.45 .45	-.84 -.29	.51 .84	.73 -.11	-	
↕	-.99 -.65	.62 .78	-.11 .45	-.99 -.97	-.65 .34	-

Emphatisch vs. neutral 4 Parameter über Kurventypen 1-7

(emphatisch)	F_o 1-7	PGL_{1-7}	WGL_{1-7}	$RHØ_{1-7}$
F_o 1-7	.87			
PGL_{1-7}	-.32	.83		
WGL_{1-7}	-.18	.8o	.89	
$RHØ_{1-7}$	.57	.17	.71	.98
(neutral)				

Emphatisch vs. neutral 7 Kurventypen über 4 Parameter

neutral	↗	→	↘	⇝	↺	↕	↕
↗	.58						
→	-.71	.93					
↘	-.55	-.2o	.7o				
⇝	.94	-.8o	.39	.98			
↺	-.4o	-.49	.87	-.28	.22		
↕	-.85	.78	-.46	-.9o	-.56	.66	
↕	-.99	.53	.o3	-.97	-.67	.95	.85

Die 4 Parameter korrelieren hoch; die 7 Kurventypen 1-7 treten ebenso häufig in den emphatischen wie in den neutralen Passagen auf. Daraus kann geschlossen werden, daß es keine paralinguistischen features per se gibt; die paraphonetische Funktion ist nicht substanzimmanent.

In der Kombinatorik unterscheiden sich jedoch beide Gruppen. Die Kopplung innerhalb der emphatischen und innerhalb der neutralen Sequenzen ist unterschiedlich bei F_o + (WGL,RHØ) sowie PGL + (WGL,RHØ) und WGL + RHØ; aber nicht bei F_o + PGL.

Die Beziehungen zwischen emphatischer und neutraler Aussprache sind konträr (negativ) bei F_o + PGL; kaum unterschiedlich bei F_o + WGL oder PGL + RHØ; sonst positiv

Auch die Kurventypen treten sowohl bei emphatischen und neutralen Passagen gleichermaßen auf (Ausnahme: ↺); d.h. Verlauf oder Richtung eines Parameters absolut betrachtet sind keine paraphonetischen Markierungen. (Anzumerken ist jedoch, daß es

sich hier um Kurvenverläufe innerhalb eines Segmentes handelt; bei suprasegmentalen Strukturen kann die Richtung des Pitches durchaus paralinguistische Informationen vermitteln.)

Die Kombinationen der Kurventypen innerhalb der emphatischen oder neutralen Äußerungen unterschieden sich bei den Kopplungen ↗ + (→ ∼ ↴ ↨), → + (↖ ∼ ↶), ↘ + (↶ ↴ ↨), ∼ + ↶, ↶ + (↴ ↨). Die Kombinationen der Kurventypen hängen zwischen emphatischen und neutralen Sequenzen folgendermaßen zusammen:
↗ + (→ ↘ ↶ ↴ ↨), → + (∼ ↶), ↘ + ↘, ∼ + (↴ ↨), ↶ + (↴ ↨) = konträr;
→ + ↘, ↘ + ↨, ∼ + ↶, ↶ + ↶ = schwach oder nicht;
↗ + (↗ ∼), → + (→ ↴ ↨), ↘ + (↘ ∼ ↶), ∼ + ∼, ↴ + (↴ ↨), ↨ + ↨ = positiv bzw. gleichgerichtet.

3.2 Faktorenanalyse

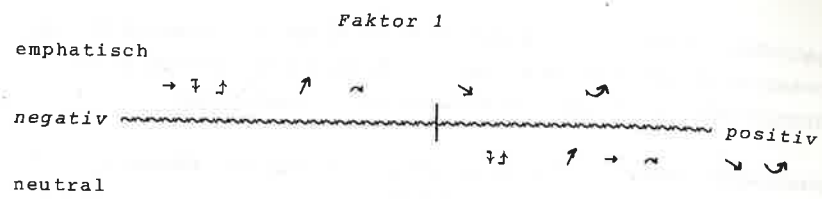
Zentroidmethode ohne Rotation; Abbruch nach 3. Faktor (Matrix noch nicht erschöpft); ↑ + ↓ zusammengefaßt

Tabelle 2: Faktorenladungen (kleine Ziffern = neutral)

	↗		→		↘		∼		↶		↴ ↨	
F_1	-.49	.26	-.63	.37	.02	.91	-.26	.42	.51	.92	-.61	.20
F_2	.66	.73	-.99	-.99	.42	-.23	.73	.99	.71	.21	-.89	-.72
F_3	-.46	.18	-.40	-.89	.41	.67	.24	-.85	.07	.66	.66	.20

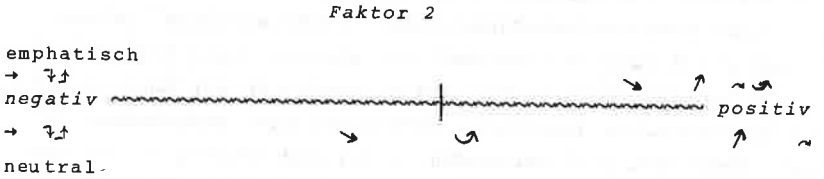
Die Faktoren lassen sich relativ leicht interpretieren:

F_1: Genereller Faktor; erster qualitativer Faktor; On-Off-Muster; Kontrastfaktor.



Emphatische Sequenzen unterschieden sich durch den Faktor 1 von neutralen Passagen in der Art eines Kontrasteffektes. Die Kurventypen sind untereinander nicht vertauscht, sondern insgesamt verschoben (negative Richtung bei emphatischen Segmenten). Ausnahmen bilden der untypische Kurvenverlauf sinkend-steigend und die Bewegung nach unten. Es treten keine strukturellen Veränderungen auf; der Faktor 1 bildet den 'externen', nicht-relationalen Unterschied zwischen emphatischen und neutralen Sequenzen ab. Wenn ein Sprecher die Parameter konträr verwendet, markiert er damit global eine Emphase (oder Neutralität).

F_2: Flüssigkeit, dynamischer Faktor.

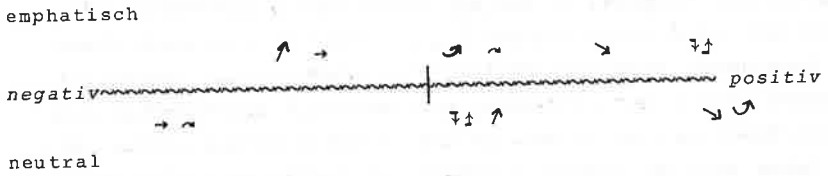


Dieser Faktor ist in gewissem Maße trivial; er zeigt nur an, daß Sprechen ein dynamischer Vorgang ist. Quasistationäre sowie abrupte Parameterbewegungen sind hoch negativ geladen, sowohl im emphatischen wie im neutralen Sprechen. Alle gleitenden, flüssigen und geschmeidigen Parameterverläufe laden positiv (Ausnahme: Sinkende Bewegung in der neutralen Aussprache). Zwischen beiden Arten können allenfalls steigende und steigend-fallende Kurven diskriminieren; quantitativ gesehen sind in neutralen Segmenten Fallen und Fallen-Steigen der Kurve weniger

hoch geladen. Abruptes, unverbundenes Sprechen mit krassem Wechsel der Parameterverläufe könnte vermutlich eher ein Zeichen einer Sprechpathologie sein als ein Merkmal des Emphatischen.

F3: Spezieller Faktor; zweiter qualitativer Faktor, Strukturkontrast (Anti- bzw. Super-Neutralität).

Faktor 3



Fast alle internen Relationen der Kurventypen sind in emphatischen Äußerungen durchbrochen. Konstant bleiben nur Kurvenverläufe der Typen → und ↘ ; eine Mittentendenz haben in emphatischen Sequenzen sinkend-steigende Bewegungen (von positiv kommend) und steigend-sinkende Verläufe (von negativ kommend). Eine Dezentralisierungstendenz haben abruptes Steigen und Sinken (nach positiv strebend) und Steigen (nach negativ strebend). Die Antinomien der Kurventypen werden in der Emphase gegenüber der neutralen Aussprache vertauscht oder aufgehoben. In neutralen Sequenzen sind Antipoden ∼ (-) und ↶ (+); → (-) und ↘ (+). In der emphatischen Aussprache sind konträr: ↗ (-) und ↘ (+); → (-) und ↕ (+). Eine mittlere Ladung haben steigende sowie abrupt sinkende/steigende Bewegungen (in emphatischer Aussprache sind diese Kurventypen Antipoden); weiterhin Steigen-Sinken und Sinken-Steigen, die in der neutralen Aussprache Antipoden sind. Der Faktor 3 trennt strukturell emphatische und neutrale Äußerungen (demgegenüber hatte Faktor 1 nur eine 'äußerliche' Anzeigefunktion).

3.3 *Probabilistische Klassifikation der Kopplung Parameter-Kurventyp*

Für die quantitative Beschreibung der Kombination von Parametern mit Kurventypen und die Berechnung der Wahrscheinlichkeit, mit der sich beide koppeln, wurde auf die Klassifikationsvorschläge von ALTMANN,LEHFELDT (198o, S. 295 ff) zurückgegriffen.

	$n_{ij}>E_{ij}$	$n_{ij}=E_{ij}$	$n_{ij}<E_{ij}$
$n_{ij}>0$	P (bevorzugt)	A (aktuell)	M (marginal)
$n_{ij}=0$	-	V (virtuell)	I (unzulässig)

Wenn $E \leq 3o$, Berechnung P (Poissonverteilung)
Wenn $E > 3o$, Berechnung z (Normalverteilung)
Wenn $P > .o5$ Klassifikation A oder V
Wenn $P \leq .o5$ Klassifikation P
Wenn $n_{ij} \leq E_{ij}$ und $P > .o5$ Klassifikation A oder V
Wenn $n_{ij} \leq E_{ij}$ und $P \leq .o5$ Klassifikation M oder I
Wenn $z > 1.645$ bei $\alpha = .o5$ Klassifikation P
Wenn $z \leq -1.645$ bei $\alpha = .o5$ Klassifikation M oder I
Wenn $z \leq 1.645$ und $z > -1.645$ bei $\alpha = .o5$ Klassifikation A oder V

Tabelle 3: Kombinatorik der Parameter mit Kurventypen (kleine Buchstaben = neutrale Aussprache)

	↗	→	↘	∼	↶	↕	↨
F_o	M A ⇐	P P	A P ⇐	M M	A P ⇐	P P	P P
PGL	P P	M M	P A ⇐	P P	P P	A M ⇐	M M
WGL	P A ⇐	A M ⇐	A A	P P	A A	M M	M M
RHØ	A M ⇐	P P	M M	A M ⇐	A M ⇐	M P ⇐	P A ⇐

In 12 von 28 möglichen Kombinationen weicht die emphatische Aussprache von der neutralen ab. F_o wird weniger häufig steigend, fallend oder fallend-steigend verwendet als in den neutralen Segmenten. Ein Heben oder Senken der Stimme innerhalb eines Lautes ist also nicht typisch für emphatisches Sprechen. Ein gleichbleibender segment-interner Pitch oder abruptes Steigen/Sinken sind in beiden Fällen gleichhäufig (bevorzugt). Das ist eine Eigenschaft, die vielleicht vom 'intrinsic pitch' von Vokalen abhängt (ANTONIADES,STRUBE 1981). Gleitendes und abruptes Sinken des Pegels tritt im emphatischen Sprechen viel häufiger auf, die Geschwindigkeit steigt häufiger an oder bleibt in mehr Fällen konstant als im neutralen Sprechen. Am meisten verändert sich in der Emphase die Geräuschhaftigkeit: Sie steigt gleitend oder abrupt viel häufiger, schwillt öfter an oder ab als in neutralen Segmenten; jedoch sinkt sie weniger oft in abrupter Art.

Tabelle 4.1. Kombination der Kurventypen von F_o und PGL
(kleine Buchstaben = neutrale Aussprache)

F_o / PGL	↗	→	↘	~	∽	↧	↥	S emphatisch/ neutral	r
↗	A v	A A	A A	A v	A v	A v	A P⇐	3.27/3.82	.76
→	A A	A A	A v	A v	V v	A v	A A	1.98/o.95	.57
↘	A v	A A	P A⇐	A A	A A	P A⇐	M A⇐	9.14/3.76	.82
~	A A	P A⇐	A A	A A	A A	A A	A M⇐	1o.27/22.47	.9o
∽	A A	A A	A A	A A	A v	A A	A A	1.11/2.37	.51
↧	V v	A A	A v	A v	V v	A A	P v	2.73/1.25	.4o
↥	V v	A A	A v	V v	V v	V A	A A	o.53/1.29	.48
S em	1.o7	11.45	8.42	2.79	1.51	6.41	4.93		
S ne	2.94	25.41	9.41	5.o1	5.99	2.75	6.32		
r	.77	.9o	.62	.88	.51	.72	.82		

Die Kombination der F_o-Kurven mit der Art des Pegelverlaufes ist in 6 Fällen beim emphatischen Sprechen anders als beim neutralen. Mit sinkendem Pegel sinkt der F_o häufiger stetig oder abrupt ab, beginnt aber weniger häufig abrupt als im neutralen Sprechen. Ein An- und Abschwellen des Pegels ist häufiger mit einem gleichbleibendem F_o verbunden oder aber mit einem abrupt steigenden Pitch. In neutralen Segmenten ist dagegen das abrupte Steigen des Pitches häufiger mit steigendem bzw. sinkendem Pegel verbunden.

Tabelle 4.2 Kombination der Kurventypen F_o und WGL
(kleine Buchstaben = neutrale Aussprache)

F_o / WGL	↗	→	↘	~	∽	↧	↥	S emphatisch/ neutral	r
↗	A v	A A	A P⇐	A A	P v	A v	A A	2.24/2.5	.57
→	A A	A A	A A	A v	A A	A A	A A	6.o5/4.72	.85
↘	A v	A A	A A	A A	A A	A A	A P⇐	5.5/3.32	.72
~	A A	A A	A A	A A	I A⇐	A A	A A	9.32/21.32	.73
∽	V A	A A	A A	A A	A A	A A	A A	1.21/o.79	.42
↧	V v	A A	A v	A v	V v	A v	A v	1.35/1.13	.37
↥	V v	A A	A v	V v	V v	A P⇐	V A	1.86/1.29	.35
S em	1.21	9.17	6.19	2.48	2.23	6.54	3.63		
S ne	2.57	24.23	7.72	5.o9	4.31	3.86	6.58		
r	.72	.83	.86	.92	-.28	.87	.83		

Die vier unterschiedlichen Kombinationen entstehen durch jeweils selteneres Koppeln von steigender Geschwindigkeit mit fallendem Pitch, sinkender Geschwindigkeit mit abrupt einsetzendem Grundton, abrupt steigender Geschwindigkeit mit abrupt fallendem F_o ; ansteigend-fallende Geschwindigkeit und fallend-steigender Pitch sind im Gegensatz zur neutralen Aussprache im Emphatischen statistisch unzulässige Kombinationen. Bevorzugt tritt dagegen in der Emphase die Kopplung von steigender Geschwindigkeit mit fallend-steigendem Grundton auf.

Tabelle 4.3 Kombination der Kurventypen F_o und RHØ (kleine Buchstaben = neutrale Aussprache)

F_o RHØ	↗	→	↘	∼	↺	↴	↕	S emphatisch/ neutral	r
↗	V v	A A	A A	A v	A v	A v	A A	2.98/1.29	.69
→	A P⇐	A A	A P⇐	A A	A A	A M⇐	M M	9.23/15.63	.86
↘	A v	A A	A A	A A	V A	A A	A P⇐	2.19/2.48	.67
∼	A I⇐	A P⇐	P M⇐	M A⇐	A A	A A	A A	7.42/14.62	.57
↺	V A	A A	A v	V v	V A	A v	A A	2.31/1.o	.14
↴	V v	A A	A A	A A	V v	A A	P A⇐	2.5/2.36	.89
↕	A A	A I⇐	A A	A v	A v	A P⇐	A P⇐	1.62/2.o7	.48
S em	1.35	9.35	7.61	2.73	1.62	4.67	2.1		
S ne	2.94	2o.35	1o.o	4.o7	4.35	3.21	3.1		
r	.57	.93	.69	.7o	.71	.63	.7o		

Die hier auftretenden Kombinationsunterschiede bestätigen im Detail, was bereits global durch den Faktor 3 ausgedrückt worden ist: Die kombinatorische Struktur von Grundton und Geräuschanteil ist in vielen Fällen in emphatischen Segmenten konträr zum neutralen Sprechen. In neutralen Passagen sind nicht kombinierbar:

- steigender Pitch und Ansteigen/Absinken von RHØ (⇔ aktuell im emphatischen Sprechen);
- gleichbleibender Grundton und abrupt ansteigender Geräuschanteil (⇔ aktuell in neutralen Segmenten).

Bevorzugt werden in neutralen Sequenzen kombiniert:

- steigender Pitch und stabiler RHØ (⇔ aktuell im emphatischen Sprechen);
- gleichbleibender Grundton und Ansteigen/Absinken des Geräuschanteils (⇔ aktuell)
- sinkender F_o und stabiler resp. fehlender Geräuschanteil (⇔ aktuell);
- abrupt sinkender Pitch mit abrupt steigendem RHØ (⇔ aktuell);
- abrupt steigender Grundton mit gleitend sinkendem oder abrupt steigendem Geräuschanteil (⇔ aktuell).

Im emphatischen Sprechen werden häufiger als im neutralen Sprechen miteinander kombiniert:

- sinkender Grundton und Ansteigen/Absinken von RHØ (⇔ marginal);
- abrupt sinkender Pitch mit stabilem Geräuschanteil (⇔ marginal);
- abrupt steigender Grundton und abrupt sinkender RHØ (⇔ aktuell).

Weniger häufig wird der steigend-fallende Grundton mit einem ansteigendem/sinkenden Geräusch gekoppelt (⇔ aktuell).

Tabelle 4.4. Kombination der Kurventypen PGL und WGL (kleine Buchstaben = neutrale Aussprache)

PGL / WGL	↗	→	↘	∼	↺	↯	↕	S emphatisch/ neutral	r
↗	P P	V A	A V	A M⇐	A V	V V	A A	5.o/3.9	.94
→	A A	A A	A A	A M⇐	A A	A A	V A	5.77/3.82	.93
↘	I A⇐	A A	P P	M M	A P⇐	A V	A V	9.29/6.o6	.93
∼	M M	A A	A M⇐	P P	M M	A I⇐	V I	17.94/52.1	.91
↺	A V	V V	A P⇐	A M⇐	P P	A V	V V	1.95/1.9	.78
↯	V V	A V	A V	I V	V V	P P	V V	3.34/1.13	.94
↕	V A	A V	A V	A I⇐	A V	A A	A P⇐	o.69/1.83	-.26
S em	5.39	1.8	1o.22	17.o5	2.43	3.o8	o.53		
S ne	3.9	o.76	6.o4	51.75	2.57	1.11	1.83		
r	.95	.79	.81	.96	.38	.74	.51		

Folgende Kombinationen des Pegels unterscheiden sich im neutralen vs. emphatischen Sprechen:

- sinkender Pegel und Anstieg/Sinken der Geschwindigkeit (⇔ marginal in neutralen Segmenten);
- steigend-sinkender Pegel und steigende, gleichbleibende oder fallend-steigende Geschwindigkeit (⇔ marginal im neutralen Sprechen) sowie abrupt steigende Geschwindigkeit (⇔ unzulässig im neutralen Segment);
- abrupt sinkender Pegel und steigend-fallende Geschwindigkeit (⇔ unzulässig).

Seltener sind im emphatischen Sprechen Kombinationen von:

- steigendem Pegel mit sinkender Geschwindigkeit (unzulässig ⇔ aktuell);
- sinkendem Pegel mit fallend-steigender Geschwindigkeit (⇔ bevorzugt);
- fallend-steigendem Pegel mit sinkender Geschwindigkeit (⇔ bevorzugt);

- abrupt steigendem Pegel und abrupt steigender Geschwindigkeit (⇔ bevorzugt)..

Tabelle 4.5 Kombination der Kurventypen PGL und RHØ (kleine Buchstaben = neutrale Aussprache)

PGL / RHØ	↗	→	↘	∼	↺	↯	↕	S emphatisch/ neutral	r
↗	A V	V V	A P⇐	A A	A A	A V	V V	2.93/1.41	.85
→	A A	A A	A A	P A⇐	A A	M V	A A	13.41/18.23	.89
↘	A P⇐	A V	M A⇐	P M⇐	A A	V V	V V	4.91/2.34	.49
∼	A M⇐	A A	P M⇐	M P⇐	A A	A A	V A	9.17/2o.97	.37
↺	V A	A V	A V	A A	V V	A V	A V	2.77/2.24	.92
↯	P A⇐	A V	A V	A A	V A	A P⇐	V A	2.o6/2.51	.o6
↕	V A	A V	A P⇐	A M⇐	A V	P A⇐	V P	2.o6/2.75	.o7
S em	3.42	1.6	1o.82	11.17	2.56	2.41	o.79		
S ne	3.o4	1.25	5.o8	31.72	2.82	1.11	1.o		
r	.45	.37	.73	.88	.83	.33	.21		

In neutralen Sequenzen werden häufiger kombiniert:

- steigender Pegel mit sinkendem Geräuschanteil (⇔ aktuell);
- sinkender Pegel und steigender Geräuschanteil (⇔ aktuell) oder sinkender RHØ (⇔ marginal) bzw. abrupt steigende Geräuschhaftigkeit (⇔ aktuell);
- abrupt sinkender Pegel und abrupt sinkender RHØ (⇔ aktuell);

Häufiger als in neutralen Sequenzen sind in emphatischen Äußerungen gekoppelt:

- steigender Pegel und abrupt sinkender Geräuschanteil (⇔ aktuell) oder steigend-fallender RHØ (⇔ marginal);
- sinkender Pegel und steigend-fallender RHØ (⇔ marginal);

- steigend-fallender Pegel und stabiler (⇔ aktuell), sinkender (⇔ marginal) oder abrupt steigender Geräuschanteil (⇔ marginal);
- abrupt sinkender Pegel und abrupt steigender RHØ (⇔ aktuell).

Tabelle 4.6 Kombination der Kurventypen WGL und RHØ (kleine Buchstaben = neutrale Aussprache)

WGL / RHØ	↗	→	↘	⌒	∽	↴	↕	S emphatisch/ neutral	r
↗	P P	A A	A A	M A⇐	A V	A V	A V	2.06/1.15	.84
→	A A	P P	A P⇐	A M⇐	A A	A V	A A	11.16/22.3	.88
↘	V P	A A	A A	A M⇐	A A	A V	V V	3.53/1.68	.50
⌒	A A	A M⇐	A I⇐	P P	A I⇐	A V	A A	12.4/26.54	.96
∽	A A	A V	A V	A A	V A	A V	A V	9.41/1.83	.26
↴	A A	A A	A A	A A	A V	P P	V A	1.5/2.69	.85
↕	A V	A A	P A⇐	M A⇐	A A	A V	P P	2.44/1.27	-.12
S em	3.16	7.7	3.78	14.3	1.57	1.07	1.07		
S ne	1.8	6.34	7.2	30.9	1.81	1.13	0.95		
r	.50	.94	.56	.99	.73	.47	.37		

Im neutralen Sprechen sind nicht kombinierbar:
- sinkende Geschwindigkeit bzw. fallend-steigende Geschwindigkeit und steigend-fallender Geräuschanteil (⇔ aktuell).

Häufiger als im emphatischen Sprechen werden in der neutralen Aussprache gekoppelt:
- sinkende Geschwindigkeit und gleichbleibender Geräuschanteil (bevorzugt ⇔ aktuell);
- steigend-fallende Geschwindigkeit und gleitend oder abrupt steigende Geräuschhaftigkeit (marginal ⇔ aktuell).

In emphatischen Segmenten sind häufiger verbunden:
- gleichbleidende Geschwindigkeit und steigend-fallender RHØ (⇔ marginal);
- sinkende Geschwindigkeit und abrupt steigende Geräsuchkomponente (bevorzugt ⇔ aktuell);
- steigend-fallende Geschwindigkeit und stabiler (⇔ marginal) oder sinkender (⇔ marginal) Geräuschanteil.

4. *Diskussion*

Das analysierte Material wurde nicht mit phonetischer Intention und nicht eigens für den vorliegenden Zweck hergestellt. Die Sprecher wurden nicht instruiert, sie waren auch keine professionellen Sprecher (etwa Schauspieler). Es wurden keine phonetische Prädispositionen getroffen (Induzieren von Standardaussprache, bewußt emphatisches Sprechen usw.). Die Auswahl der Samples war zufällig; die Typisierung wurde ohne vorgefertigtes Klassifikationsschema direkt aus den Kurvenplots entwickelt. Die Strukturunterschiede wurden bewußt nicht absolut (etwa nur eine Sprechweise, neutral *oder* emphatisch), sondern im Vergleich herausgearbeitet. Die Beschränkung auf die Segmentebene war beabsichtigt, um genau jene Besonderheiten herauszulösen, die eine phonematisch bedingte phonetische Form zur paraphonetischen werden lassen.

4.1 *Ergebnisdiskussion unter methodologischem Aspekt*

Die impressionistische Beurteilung von Kurvenbildern ist auch angesichts der modernen Methoden der digitalen Signalanalyse immer noch der Prüfalgorithmus per se. Jedes signalphonetische Analyseergebnis und jedes Analyseprogramm wird vom Phonetiker durch die optische (und auditive) Inspektion kontrolliert. Solche routinemäßigen Überwachungen stützen sich auf die intuitive Kenntnis von normalen, möglichen und abweichenden Parameterkonstellationen. Die probabilistischen Auflistungen in den Tabellen 4.1 - 4.6 beschreiben aus diesem Grunde nicht nur die phonetischen

Unterschiede von neutralen oder emphatischen Sprechmustern, sondern zugleich einige der methodologischen Implikationen bei der Interpretation phonetischer Kurven. Beispielsweise sind solche Kombinationen, die als unzulässig im statistischen Sinne klassifiziert wurden (Buchstabe I), für den Phonetiker Anhaltspunkte dafür, ob ein Programmfehler auftritt, ob emphatische Anteile im Signal erkennbar sind und welche Fraktionen des Signals überarbeitet werden müssen. Solche prototypischen Kombinationen sind im Anhang noch einmal zusammengestellt. Für den untrainierten Hörer existieren diese Bewertungsmuster nur vor-prädikativ, jedoch automatisiert und durch lange Erfahrung sedimentiert. Man kann die Vermutung aussprechen, daß für die Identifizierung von Emphase gerade im alltäglichen Kommunizieren die Parameterkombination entscheidend ist. Für die wissenschaftliche Analysetechnik sind sie ganz sicher ausschlaggebend: Korrekturprogramme für die Stilisierung und Aufbereitung natürlicher Sprachaufnahmen arbeiten teilweise mit Entscheidungen über Proportionen zwischen den einzelnen Parametern. Synthetische Sprache kann u.a. dadurch natürlicher klingen, daß Kombinationen eingefügt werden, die nicht dem Standardspektrum von Phonemen entsprechen. Bei linguistisch-phonematischen Analysen muß nun gerade umgekehrt verfahren werden: Die Materialien müssen so präpariert werden, daß die dem Phonem zugehörigen Konstellationen pronunciert und eindeutig auftreten. Dazu ist jedoch das Vorwissen nötig, welche Konfigurationen prototypisch für Sprachlaute sind. Die Vorweg-Interpretation, bevor die akustische Analyse durchgeführt werden kann, stützt sich auf die intuitive Kenntnis über mögliche und außergewöhnliche Kombinationen. Die Vorprüfung des Materials wird in der Regel ja ohrenphonetisch vorgenommen, unterstützt durch Probeanalysen und Probeplots, die ihrerseits anhand der Kombinatorik eingeschätzt werden. Wissenschaftsphilosophisch gesehen ist das lege-artes-Prozedere der Phonetik eine Bestätigung dafür, wie stark wissenschaftliches Arbeiten (wenn es lebensweltliche Bedeutung haben soll) von alltagsweltlichen Kenntnissen und Zugriffsalgorithmen abhängt.

4.2 Strukturelle Unterschiede zwischen neutralem und emphatischem Sprechen

In der Literatur werden z.T. sehr ausführlich die typischen Kopplungen phonetischer Parameter beschrieben, vorwiegend auf der Grundlage sprechphysiologischer Erfahrungen. FANT (1974, S. 2oo) leitet die Tatsache, daß eine Verdopplung des F_o zugleich eine Verdopplung der Intensität hervorruft (Steigerung des Pegels um ca. 3 dB), aus der stimmphysiologischen Notwendigkeit ab, durch die bei einer Anspannung der Stimmlippen zur Tonerhöhung zugleich Atemdruck und Impuls steigen müssen. Die physiologischen Ursachen sollen hier einmal außer Acht gelassen werden; statistisch gesehen ist die Kombination "steigender Pitch + steigender Pegel" nur dann prototypisch (bevorzugt), wenn der Grundton *abrupt* steigt. In der emphatischen Aussprache ist diese Kopplung nur bei abrupt steigendem Pitch + abrupt steigendem Pegel bevorzugt. Dieses Ergebnis paßt noch besser zu FANT's Aussage: Eine impulsartige Erhöhung der Tonhöhe zieht die abrupte Steigerung der Intensität nach sich, was stimmphysiologisch plausibel ist. Die gleitende Grundtonerhöhung ist in der neutralen Aussprache nur virtuell mit gleitender Pegelerhöhung gekoppelt, im emphatischen Segment statistisch aktuell. Mit anderen Worten: Nicht immer muß eine Grundtonerhöhung mit Intensitätsanstieg verbunden sein; offenbar gibt es physiologische Kompensationsmechanismen, die bei einer langsamen Tonerhöhung der Druckwelle entgegenarbeiten.

In der Arbeit von CARLSON, ERIKSON, GRANSTRÖM,LINDBLOM, RAPP (1974) über neutrale und emphatische Betonungsmuster im Schwedischen findet sich eine Bestätigung dafür, daß nicht die Richtung des Grundtones (z.B. die "Deklination") oder die absolute Tonhöhe Markierungen des Emphatischen sind. Die Autoren fanden, daß der einzige Unterschied zwischen neutralen und emphatischen Patterns des F_o lediglich in einer Gipfelbildung (Peak) des Grundtones besteht (was dem Kurventyp "⌐"

entspricht). Jedoch verfolgten die Autoren nicht weiter ihre Hypothese, daß die Peaks durch emphatische Sprechweise entstanden sind. Mit den vorliegenden Ergebnissen läßt sich diese Hypothese nicht verifizieren; die Konzentration eines Parameters auf einen Kurventyp (F_o auf Peakbildung) läßt sich statistisch nicht klar belegen bzw. kann keine spezielle Kurvenbildung eines einzigen Parameters als typisch für Emphase angesehen werden. Die Dynamik und Richtung eines Kurvenverlaufes innerhalb des Segmentes ist keine eindeutige Markierung für Emphase vs. Neutralität. Die hohen Korrelationen von Grundton, Pegel, Weglänge und Geräuschanteil zwischen beiden Sprechweisen sind weiterhin ein Beleg dafür, daß paraphonetische Informationen nicht an bestimmte Parameter gebunden sind.

Der Unterschied zwischen neutralen und emphatischen Segmenten drückt sich in der Kombination der Parameter sowie deren Kurventyp aus. Alle Parameter mit Ausnahme von F_o und Pegel sind bei emphatischen Sequenzen anders kombiniert als bei neutralen Äußerungen. Aus der faktorenanalytischen Ordnung kann abgeleitet werden:

- Es besteht kein Unterschied in der Flüssigkeit der Parameter zwischen neutralen und emphatischen Äußerungen (F2).
- Emphatisches Sprechen wird global durch entgegengesetzte Parameterkopplung angezeigt (F1). Abrupt sinkende Geräuschhaftigkeit ist z.B. typisch für neutrale Aussprache, untypisch (marginal) für Emphase.
- Strukturelle Unterschiede (F3), d.h. Austausch und Unterlaufen der phonematischen Klanggestalt durch Kombinationsänderung, innerhalb des Segmentes tragen am meisten zur paraphonetischen Markierung bei. Sinkend-steigende Kurvenbewegungen sind typisch für neutrale Äußerungen (im emphatischen Sprechen kommt dieser Typ auch vor, ist jedoch nicht kennzeichnend); während abruptes Steigen oder abruptes Sinken von Parametern typisch für Emphase ist.

Emphatisches Sprechen wird vor allem durch die Art des Verlaufs von Geräuschbeimischungen angezeigt. Im neutralen Sprechen sind Kombinationen von steigender Intensität + steigender Geschwindigkeit + *steigender* Geräuschhaftigkeit prototypisch, während im emphatischen Segment in solchen Fällen der Geräuschanteil *abrupt sinkt*. Bei sinkender Intensität + sinkender Geschwindigkeit *steigt* im neutralen Sprechen die Geräuschhaftigkeit, im emphatischen bleibt sie konstant oder nahe Null. Das ist besonders dann der Fall, wenn stimmlose Laute stimmhaft werden bzw. wenn vokalische Elemente in Konsonanten eingeschoben werden. Für welche expressive Gattung welcher Kombinationstyp vorrangig verwendet wird (welcher Affekt durch welche Konfigurationen angezeigt wird), ist in dieser Arbeit nicht untersucht worden.

Fußnote:

1 Material des Konstanzer Projektes "Analyse unmittelbarer Kommunikation und Interaktion als Zugang zum Problem der Entstehung sozialwissenschaftlicher Daten", finanziert von der Fritz-Thyssen-Stiftung; Leitung: Prof. Th. Luckmann/Prof.Dr. P. Gross.

ANHANG

Probabilistische Prototypen (Klassen P und I) der Parameter- und Kurventyp-Kombinationen

α = o.o5 E = emphatisch N = neutral - = unzulässig

	Kurventyp ↗	→	↘	~	↘↗	↗↘	↕
F_o		E N	N		N	E N	E N
PGL	E N		E	E N	E N		
WGL	E			E N			
RHØ		E N				N	N

RHØ	WGL ↗	→	↘	~	↘↗	↗↘	↕
↗	E N						
→		E N	N				
↘	N						
~			-N	E N	-N		
↘↗							
↗↘						E N	
↕			E				E N

WGL	PGL ↗	→	↘	~	↘↗	↗↘	↕
↗	E N						
→							
↘	-E		E N		N		
~				E N		-N	-N
↘↗			N		E N		
↗↘				-E		E N	
↕				-N			N
RHØ							
↗			N				
→			E				
↘	N		E				
~			E	N			
↘↗							
↗↘	E					N	
↕			N			E	N

PGL	F	→	↘	~	↺	↧	↕
↗							N
→							
↘							
~		E				E	
↺		E					
↧							
↕							E
WGL							
↗			N		E		
→							
↘							N
~					-E		
↺							
↧							
↕							N
RHØ							
↗							
→	N		N				
↘							N
~	-N	N	E				
↺							
↧							E
↕		-N				N	N

Beispiele:

	neutral							emphatisch						
FØ	↕	()	()	()	()	()	()	→	→	↧	↕	()	()	()
PGL	↗	↗	↗	↘	~	↺	↕	~	↺	~	↕	↘	↧	↗
WGL	↘	↗	↗	↘	~	↺	↕	~	↺	~	↕	↘	↧	↗
RHØ	↘	→	↘	↗	~	()	↕	~	()	~	↧	→	↧	↧

LITERATUR

ALTMANN, G., LEHFELDT, W. 1980 Quantitative Phonologie. Bochum.

ANTONIADES, Z., STRUBE, H.W. 1981 Untersuchungen zum "intrinsic pitch" deutscher Vokale. Phonetica 38, 277-290.

BUGENTAL, D.E. 1974 Interpretation of naturally occurring discrepancies between words and intonation: Modes of inconsistency resolution. Journal of Personality and Social Psychology 30, 125-133.

CARLSON, B., ERIKSON, Y., GRANSTRÖM, B., LINDBLOM, B., RAPP, K. 1974 Neutral and emphatic stress patterns in Swedish. In: FANT, G. (Ed.): Speech Communication, Vol. 2, 209-217. Stockholm.

FANT, G. 1974 Analysis and synthesis of speech processes. In: MALMBERG, B. (Ed.): Manual of Phonetics, 173-277. Amsterdam, London, New York.

LAVER, J., TRUGDILL, P. 1979 Phonetic and linguistic markers in speech. In: SCHERER, K.R., GILES, H. (Eds.): Social Markers in Speech, 1-33. Cambridge.

IZARD, C.E. 1965 Affect, awareness, and performance. In: TOMKINS, S. S., IZARD, C.E. (Eds.): Affect, Cognition, and Personality. New York.

LIEBERMAN, P., MICHAELS, S.B. 1962 Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech. Journal of the Acoustical Society of America 34, 922-927.

LOVEDAY, L. 1981 Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of english and japanese politeness formulae. Language and Speech 24, 71-88.

MARTIN, H. 1977 The prosodic and paralinguistic analysis of dramatic speech. Ann. Arbor, University of Michigan Phonetics Lab.

MARTIN, H. 1981 The prosodic components of speech melody. The Quarterly Journal of Speech 67, 81-99.

't HART, J. 1974 Discriminability of the Size of Pitch Movements in Speech. IPO Annual Progress Report 9, 56-63.

UMEDA, N. 1981 F_o rule for discourse. Journal of the Acoustical Society of America 69, S82.

UMEDA, N. 1982 "F_o declination" is situation dependent. Journal of Phonetics 1o, 279-29o.

Linguo-statistical Studies of Siberian Languages in the USSR

Project Report

Yuri A. Tambovtsev, Novosibirsk

The statistical studies of Siberian languages began in autumn 1973, so that in autumn 1983 they will celebrate their tenth anniversary. Our group of linguo-statistical studies started with phonostatistics: for practical purposes (especially for publishing) it was necessary to know the frequency of occurrence of different phonemes of Siberian native languages. The investigations were held at the Computing Centre of the Novosibirsk State University with the help of specialists in programming and speech recognition of the Laboratory of Technical Cybernetics.

The first language computed was Mansi (Vogul). The Northern and Konda dialect texts were transcribed by Mansi native speakers. Then this material, containing about half a million of phonemes, was fed to a computer; the same procedure was applied to the other languages computed by the group of experimental linguistics of the Novosibirsk State University. Our first computer was a M-222, then a computer of the third generation was used. At the present time we work with the computers EC-1022 and EC-1033 (with a memory of 500 and 800 K respectively). The programming languages were "Epsylon" and "Fortran", now we use "PL/1".

The group of the Novosibirsk University directed by Y.A. Tambovtsev has computed the following Finno-Ugric languages: Khanty (Ostyak), Udmurt (Votyak), Komi-Zyryan, Mari (Cheremis), Karelian, Finnish, Mordva (Erzya) and Saame (Lopari). In addition to the Finno-Ugric family, languages of the Turkish, Paleo-Asiatic and Tungus-Manchurian families were computed: Khakas, Altay, Yakut, Kazakh, Ket, Eskimo, Koryak, Itelmen, Nanay, Oroch, Orok and Japanese. Every time the largest sample possible of a language was fed to the computer, but unfortunately some of them proved to be not large enough (the smallest sample contained more than 10.000 phonemes), so that the computing results of these samples should be considered preliminary. The

material of these languages will be added later.

We collected the following frequency data:

1) Frequency of occurrence of phonemes;
2) frequency of occurrence of certain phonemes in certain positions (especially in word-initial and word-final position);
3) frequency of combinations of two phonemes;
4) frequency of occurrence of certain dyads (combinations of two phonemes) in certain positions (especially in word-initial and word-final position);
5) frequency of occurrence of triads (combinations of three phonemes).

It should be stressed that all investigations were based on phonemes but not on graphemes, since in a language one letter may or may not correspond to one sound.

Near the end of the first stage of our study - i.e. the investigations in the field of phonemic statistics - it is planned to proceed with investigations in the field of lexical statistics, and then with grammatical statistical investigations of the same or enlarged material of the languages in question.

At the present time our group collects material on Mongol, Buryat, Tibetan, Nivkh and some other languages of Asia. The aim of the group is to continue to compute other languages of Asia and the Far East. After that it is planned to analyze and compare statistically as many languages of Siberia, Asia and the Far East as possible.

THE REPEAT RATE OF PHONEME FREQUENCIES AND THE ZIPF-MANDELBROT LAW

P. Zörnig, Dortmund, G. Altmann, Bochum

1. In a previous work (Altmann and Lehfeldt 1980: 151-166) following a proposal of Sigurd (1968) it was assumed that the relative frequencies of phonemes (or letters) ranked according to their magnitude are geometrically distributed; using an approximate method it was determined that the repeat rate is

$$R_K = \sum_{j=1}^{K} p_j^2 \approx 2/K \qquad (1)$$

where K is the number of phonemes (letters) in the inventory and p_j is the relative frequency of the j-th phoneme. This curve can be found as R_1 marked in Fig. 1 together with the observed Rs of 63 data points. It has been found that the sum of squared deviations of the observed data from this curve yields

$$W = \sum_{K=13}^{74} \sum_{i=1}^{n_K} (R_{iK} - 2/K)^2 = 0.0082 \qquad (2)$$

where n_K is the number of languages with K phonemes (letters) and K takes values between 13 to 74.

Though the result was "optically" quite satisfactory it has been recommended to search for a better theoretical curve since the geometric distribution was not adequate for the ranked phoneme frequencies.

2. In linguistics it is generally accepted that the ranked frequencies of linguistic entities are distributed according to the Zipf-Mandelbrot law

$$p_j = \frac{A}{(B+j)^C} \qquad (3)$$

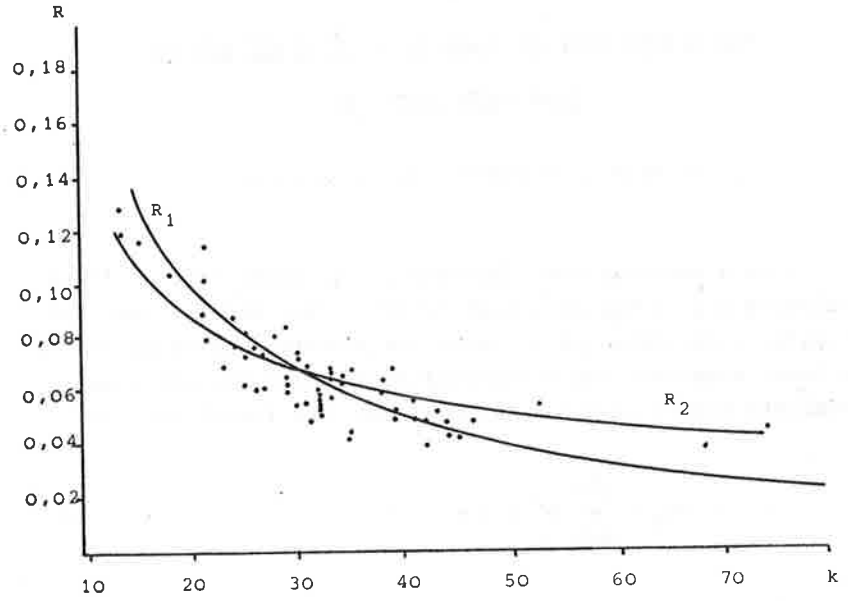


Fig. 1. Repeat rates of 64 languages and two theoretical curves.

where j is the rank of the entity (j = 1,2,...,K), B and c are certain constants and $A = [\sum_{j=1}^{K}(B+j)^{-c}]^{-1}$. Since (3) is usually used for word frequencies K is assumed to be infinite, but since we are considering phoneme inventories with maximally 74 phonemes and want to show that R depends on K we must use a finite and variable K.

Using (3) we obtain for the repeat rate

$$R = \sum_{j=1}^{K} p_j^2 = A^2 \sum_{j=1}^{K} (B+j)^{-2c} \qquad (4)$$

or

$$R = \left[\sum_{j=1}^{K} (B+j)^{-c}\right]^{-2} \left[\sum_{i=1}^{K} (B+i)^{-2c}\right]. \qquad (5)$$

The computation of sums in (5) presents considerable difficulties and therefore we use an approximation. Fig. 2 illustrates the relation

$$\sum_{j=1}^{K} (B+j)^{-c} \approx \int_{1}^{K} \frac{dx}{(B+x)^c} = \begin{cases} \ln \frac{B+K}{B+1} & \text{for } c = 1 \\ \frac{(B+K)^{1-c} - (B+1)^{1-c}}{1-c} & \text{for } c \neq 1 \end{cases} \qquad (6)$$

for any nonnegative B and c. Replacing the sums in (5) by the

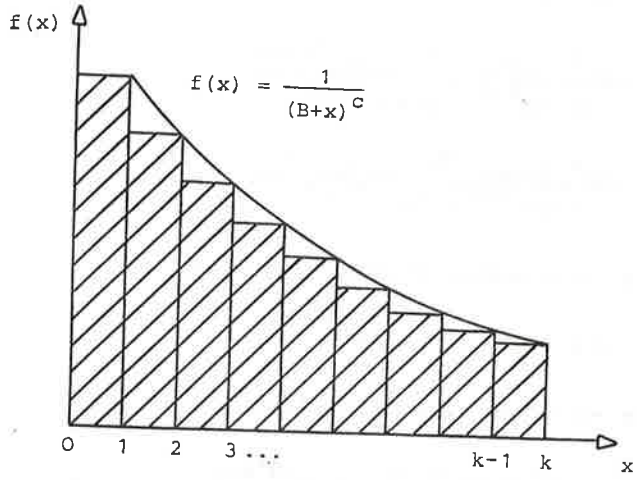


Fig. 2. $\sum_{j=1}^{k} \frac{1}{(B+j)^c}$ is the area with hatching. The integral in (6) is the area under the curve.

corresponding approximations in (6) we obtain

$$\frac{(1-c)^2[(B+K)^{1-2c} - (B+1)^{1-2c}]}{(1-2c)[(B+K)^{1-c} - (B+1)^{1-c}]^2} \quad \text{for } c\neq 1,\ c\neq 0.5 \qquad (7a)$$

$$\frac{(1-c)^2}{[(B+K)^{1-c} - (B+1)^{1-c}]^2} \ln \frac{B+K}{B+1} \quad \text{for } c=0.5 \qquad (7b)$$

$$\frac{(B+1)^{-1} - (B+K)^{-1}}{\left(\ln \frac{B+K}{B+1}\right)^2} \quad \text{for } c=1 \qquad (7c)$$

The constants B and c can be estimated in several ways. We content ourselves with the follwoing simple method: we choose two points, let us say K = 15 and K = 40 and find in Fig. 1 the approximate values of R_{15} and R_{40}. We obtain

$$R_{15} \approx 0.11, \qquad R_{40} \approx 0.05.$$

By means of these values we simultaneously solve the equations

$$0.11 = \frac{(1-c)^2[(B+15)^{1-2c} - (B+1)^{1-2c}]}{[(B+15)^{1-c} - (B+1)^{1-c}]^2}$$

$$0.05 = \frac{(1-c)^2[(B+40)^{1-2c} - (B+1)^{1-2c}]}{[(B+40)^{1-c} - (B+1)^{1-c}]^2}$$

by means of a trial-and-error method and obtain

$$B \approx 0.59, \qquad c \approx 0.99.$$

Putting these values in (7a) we obtain

$$R_K = \frac{-0.0001[(0.59+K)^{-0.98} - 1.59^{-0.98}]}{[(0.59+K)^{0.01} - 1.59^{0.01}]^2}.$$

This curve can be called R_2. It is evidently better than $R_1 = 2/K$ since the sum of squared deviations is now W = 0.006738. This number can be reduced yet further if B and c are more exactly estimated and if a better approximation of the Zipf-Mandelbrot formula is used (cf. e.g. Zörnig 1983).

In the course of the iterative correction of the fitting we ascertained that the approximation improved as c approached 1. Assuming that c = 1 we can use formula (7c). The approximation will be even better if we take e.g. B = 0.61, i.e.

$$R_K = \frac{1.61^{-1} - (0.61+K)^{-1}}{(\ln \frac{0.61+K}{1.61})^2}. \qquad (8)$$

This curve is represented as R_2 in Fig. 1. The sum of the squared deviations is now W = 0.006738.

The fact that a simpler, theoretically better grounded curve (R_2 or R_3) yields better results is sufficient reason to give it preference over a competing curve and to accept it for the given data. The problem is that B cannot be theoretically derived for the time being. No approaches from other domains of linguistics are known to us. With any additional data B must be estimated anew until a theoretical value is found.

Assuming that the deviations of the observed $\hat{R}$s from the theoretical curve remain normal and in each point K equal we can compute a confidence interval for the theoretical curve. In analogy to Altmann/Lehfeldt (1980: 160-161) we compute the variance of $\hat{R}$ from the multinomial distribution of the relative frequencies as

$$V(\hat{R}) = \frac{4}{N}\left(\sum_{j=1}^{K} p_j^3 - R^2\right) \qquad (9)$$

Here $p_j = A/(B+j)$ (since c = 1) and according to (5) and (6)

$$\sum_{j=1}^{K} p_j^3 \approx \left[\int_1^K \frac{dx}{B+x}\right]^{-3}\left[\int_1^K \frac{dx}{(b+x)^3}\right] = \frac{1}{2}\,\frac{[(B+1)^{-2} - (B+K)^{-2}]}{(\ln \frac{B+K}{B+1})^3}. \qquad (10)$$

Inserting (8) and (10) in (9) we obtain

$$V(\hat{R}) \approx \frac{4}{N}\left\{\frac{1}{2}\,\frac{[1.61^{-2} - (0.61+K)^{-2}]}{(\ln \frac{0.61+K}{1.61})^3} - \frac{[1.61^{-1} - (0.61+K)^{-1}]^2}{(\ln \frac{0.61+K}{1.61})^4}\right\} \qquad (11)$$

where N is the number of languages. The confidence band

$$R_3 + z_{\alpha/2}\sqrt{V(\hat{R})}, \quad R_3 + z_{1-\alpha/2}\sqrt{V(\hat{R})}$$

is shown in Fig. 3. The values of $(K, \hat{R}, R_3)$ are shown in Table 1 (for sources cf. Altmann, Lehfeldt 1980: 154-156).

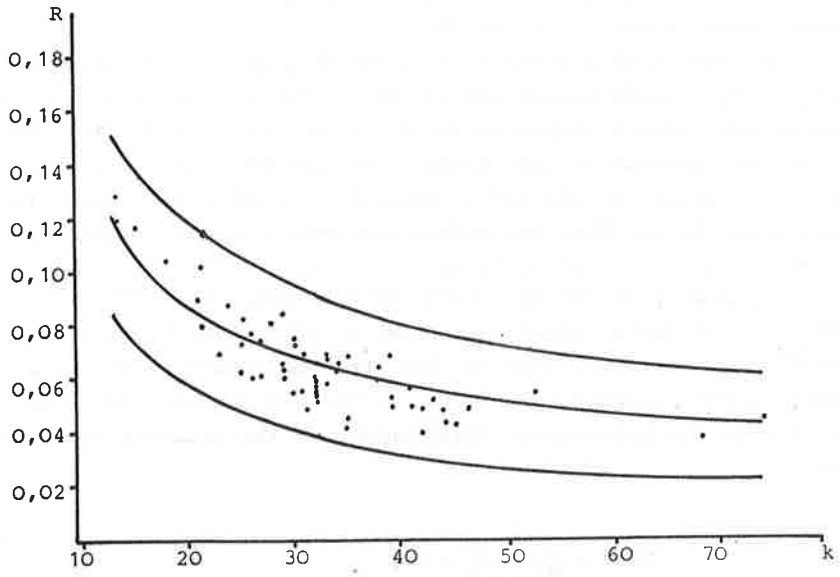


Fig. 3. Confidence band

Table 1. Repeat rate for 63 languages

	Language	K	$\hat{R}$	R
1	Hawaiian	13	0.120716	0.120192
2	Hawaiian L	13	0.131031	
3	Samoan	15	0.118269	0.107946
4	Hawaiian	18	0.105342	0.094720
5	Pilipino	21	0.102547	0.085238
6	Pilipino	21	0.116018	
7	Kaiwa	21	0.080747	
8	Sea-Dayak L	21	0.091596	
9	Estonian L	23	0.070229	0.080255
10	Swahili L	24	0.087816	0.078086
11	French L	24	0.079699	
12	Albanian L	25	0.063719	0.076039
13	Indonesian L	25	0.083864	
14	Chamorro	25	0.074272	
15	Dutch L	26	0.077732	0.074163
16	English L	26	0.062183	
17	Rumanian	27	0.062323	0.072419
18	Spanish L	27	0.074637	
19	Haussa L	27	0.105588	
20	Dutch L	28	0.082022	0.070792
21	Serbocroatian L	29	0.063378	0.069270
22	Bulgarian L	29	0.067229	
23	German L	29	0.072007	

Table 1. (Cont.)

	Language	K	$\hat{R}$	R
24	Indonesian (text)	29	0.085824	
25	Indonesian (lex.)	29	0.060762	
26	German L	30	0.073938	0.067844
27	Gujarati	30	0.055911	
28	Italian L	30	0.074775	
29	Italian	31	0.067565	0.066503
30	Ikrainian L	31	0.049725	
31	Russian L	31	0.057713	
32	Amer. English	32	0.054875	0.065241
33	Hungarian	32	0.053986	
34	Hungarian L	32	0.052090	
35	Khasi	32	0.062487	
36	Latvian L	32	0.056568	
37	Russian L	32	0.056568	
38	German	33	0.059241	0.058961
39	Georgian	33	0.070258	
40	Georgian	33	0.069390	
41	Ostyak	33	0.062470	
42	Ostyak	33	0.066969	
43	Ostyak	34	0.066856	0.062922
44	Ostyak	34	0.064061	
45	Czech	35	0.046491	0.061854
46	Czech L	35	0.043964	
47	French	35	0.070591	
48	Marathi	38	0.060408	0.058961
49	Bengali	38	0.065865	
50	Hungarian	39	0.052800	0.058087
51	English	39	0.050495	
52	Armenian L	39	0.070748	
53	Russian	41	0.050003	0.056456
54	Polish	42	0.050928	0.055693
55	English	42	0.040990	0.055693
56	Gujarati L	43	0.055151	0.054962
57	English	44	0.043749	0.054260
58	Slovak	44	0.048530	
59	Swedish	45	0.043407	0.053587
60	Ukrainian	46	0.049402	0.052940
61	Hindi	52	0.054557	0.049528
62	Burmese L	68	0.039244	0.043081
63	Vietnamese	74	0.047003	0.041298

REFERENCES

ALTMANN, G., LEHFELDT, W., Einführung in die quantitative Phonologie. Bochum, Brockmeyer 1980

SIGURD, B., Rank-frequency distribution for phonemes. Phonetica 18, 1968, 1-15

ZÖRNIG, P., Zwei allgemeine Näherungsformeln. (im Druck)

ZÖRNIG, P., Zwei neue Summenformeln. (im Druck)

CURRENT BIBLIOGRAPHY

ABBREVIATIONS

IJPs	International Journal of Psycholinguistics
JPsyR	Journal of Psycholinguistic Research
JVLVB	Journal of Verbal Learning and Verbal Behavior
L&S	Language and Speech
Sistema	Sistema i struktura jazyka v svete marksistsko-leninskoj metodologii. Kiev 1981.
Symposium 1979	Symposium: Mathematical Processing of Cartographic Data (Tallinn, Academy of Sciences of the Estonian S.S.R. Division of Social Sciences, December 18-19, 1979), Summaries. Tallinn 1979.
Symposium 1981	Symposium: Processing of Dialectological Data (Tallinn, November 23-25, 1981), Summaries. Tallinn 1981.

GENERAL

1. ALEKSEEV, P.M.: O nelinejnyx formulirovkax zakona Cipfa [On non-linear formulations of Zipf's law]. Voprosy kibernetiki 41, 1978, 53-65.
2. AXUNDOV, A.: Matematičeskoe jazykoznanie [Mathematical Linguistics]. Baku 1979.
3. GLADKIJ, A.V. (otv. red.) et al.: Matematičeskaja logika i matematičeskaja lingvistika [Mathematical Logic and Mathematical Linguistics]. Kalinin: Kalininskij gosudarstvennyj universitet 1981, 171 pp.
4. HUG, M: La statistique linguistique en France. Pottier, B. (Ed.): Les sciences du langage en France au XXème siècle. Paris: Legras 1980, 371-405.
5. KOMAROVA, L.I.: Verojatnost' kak mera vozmožnogo v jazyke [Probability as a measure of what is possible in language]. Sistema, 102-114.
6. KOVAL'ČENKO, I.D. (red.): Količestvennye metody v gumanitarnyx naukax [Quantitative Methods in the Humanities]. Moskva: Izdatel'stvo Moskovskogo universiteta 1981, 206 pp.
7. KRÁLIK, J.: Nové sovětské příspěvky z kvantitativní lingvistiky [New Soviet contributions on quantitative linguistics]. Slovo a slovesnost 43, 1982, 58-62.
8. LESOXIN, M.M., LUK'JANENKOV, K.F., PIOTROVSKIJ, R.G.: Vvedenie v matematičeskuju lingvistiku [An Introduction to Mathematical Linguistics]. Minsk: Nauka i technika 1982, 263 pp.
9. MURAVICKAJA, M.P., SLIPČENKO, L.L.: Simmetrija v lingvističeskix sistemax [Symmetry in linguistic systems]. Sistema, 70-84.
10. PEREBYJNIS, V.S.: Teoretyčni ta prykladni problemy strukturno-matematyčnoji linhvistyky [Theoretical and applied problems of structural-mathematical linguistics]. Movoznavstvo 1981/4, 3-13.
11. REMMEL', M.: O vyčislitel'no-lingvističeskix issledovanijax v Institute jazyka i literatury [On computer-aided linguistic research in the Institute of Language and Literature]. Akademija nauk Èstonskoj SSR v 1973-1979 godax. Tallin 1981, 298-301.
12. ŠTĚPÁN, J.: Kniha o statistických metodách v české gramatice [A book on quantitative methods in Czech grammar].- Review of: Těšitelová, M.: Využití statistických metod v gramatice. Praha 1980, 219 pp. Slovo a slovesnost 43, 1982, 55-58.
13. VELIEVA, K., MAXMUDOV, M., PINES, B.: Kniga o matematičeskix metodax issledovanij v jazykoznanii [A book on mathematical methods in linguistic research]. - Review of:

Axundov, A.: Matematičeskoe jazykoznanie [Mathematical Linguistics]. Baku 1979. Izvestija Akademii Nauk AzSSR. Serija literatury, jazykov i iskusstva. Baku, 1980/2, 133-134.

PHONOLOGY

14. BAZUNOV, S.I., TARASOV, A.I., TIRASPOL'SKIJ, Ju.I., JAKUŠENKOV, G.A.: K voprosu o statistike osnovnogo tona [On the problem of the statistics of the fundamental tone]. Voprosy kibernetiki: Analiz i sintez reči v sistemax upravlenija. Moskva 1981, 75-80.

15. GROTJAHN, R.: Ein statistisches Modell für die Verteilung der Wortlänge. Zeitschrift für Sprachwissenschaft 1, 1982, 44-75.

16. KOHN, K.: Die Rolle der Sprachlauterwartung in der kategorialen Wahrnehmung. Phonetica 38, 1981, 309-319.

17. RICHTER, L.: Wpływ tempa mowy na czas trwania głosek w języku polskim [The effect of speech tempo on the duration of speech sounds in Polish]. Polonica 6, 1980, 19-35.

18. SATO, S., YOKOTA, M., KASUYA, H.: Statistical relationships among the first formant frequencies in vowel segments in continuous speech. Phonetica 39, 1982, 36-46.

19. WERKEN, A.: On the evolution of word-length in Dutch. Jones, A., Churchhouse, R.F. (Eds.): The Computer in Literary and Linguistic Studies. Cardiff 1976, 271-284.

MORPHOLOGY

20. BARTKOV, B.I.: Kvantitativnye metody issledovanija slovoobrazovatel'noj podsistemy sovremennogo anglijskogo jazyka [Quantitative methods in the analysis of the word-formative subsystem of modern English]. Affiksoidy, po-

luaffiksy i affiksy v naučnom stile i literaturnoj norme. Vladivostok 1980, 117-142.

21. KULEŠOVA, L.V.: Opyt opisanija kornej russkogo jazyka s pomošč'ju statističeskix metodov [An attempt to describe the root morphemes of Russian with the aid of statistical methods]. Količestvennye metody v gumanitarnyx naukax. Moskva 1981, 200-204.

LEXICOLOGY

22. ALEKSEEV, P.M.: K osnovam statističeskoj leksikografii [On the fundamentals of statistical lexicography]. Problema slova i slovosočetanija. Leningrad 1980, 93-105.

23. AZLAROV, T.A., MUCHAMEDXANOVA, R., AXMEDOVA, V.K.: Formuly dlja granic zakona Cipfa i ix primenenie pri izučenii častotnogo slovarja uzbekskogo jazyka [Formulas for the limits of Zipf's law and their application to the study of the frequency vocabulary of the Uzbek language]. Piotrovskij, R.G. (Ed.): Inženernaja lingvistika i prepodavanie inostrannyx jazykov s pomošč'ju TSO. Leningrad: GPI 1981, 140-148.

24. BABANAROV, A.: Častotnyj slovnik i avtomatičeskij slovar' dlja mašinnogo perevoda tureckix gazetnyx tekstov [A frequency list of dictionary entries and an automatic dictionary for the machine translation of Turkish newspaper Texts]. Inženernaja lingvistika i optimatizacija prepodavanija inostrannyx jazykov. Leningrad 1980, 126-136.

25. BEKTAEV, K.B. (Ed.): Častotnyj slovar' po romanu M.O. Auėzova "Put' Abaja" [Frequency Dictionary of M.O. Auėzov's Novel "Put' Abaja"]. Alma-Ata: Nauka 1979, 336 pp.

26. BEÖTHY, E., ALTMANN, G.: Das Piotrowski-Gesetz und der Lehnwortschatz. Zeitschrift für Sprachwissenschaft 1, 1982, 171-178.

27. DIETZE, J.: Aufbau einsprachiger Frequenzwörterbücher. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 1982, 294-299.

28. SCURTU, G.: Essai de statistique sur l'élément lexical d'origine latine de la langue française. Revue Roumaine de Linguistique 27, 1982, 413-421.

TEXT ANALYSIS

29. CONFORTOVÁ, H.: Slovesa v oblasti hospodářské češtiny z hlediska kvantitativního [Quantitative aspects of the Czech verbs from the sphere of economy]. Slovo a slovesnost 43, 1982, 125-133.

30. DARČUK, N.P.: Mesto količestvennyx metodov v stilističeskix issledovanijax [The place of quantitative methods in stylistic analyseś]. Sistema, 172-182.

31. GRIGOR'EVA, A.S.: O leksiko-morfologičeskoj statistike russkoj ėpistoljarnoj reči [Lexical-morphological statistics of the Russian epistolary language]. Piotrovskij, R.G. (Ed.): Inženernaja lingvistika i prepodavanie inostrannyx jazykov s pomošč'ju TSO. Leningrad: GPI 1981, 140-148.

32. JAKUBAJTIS, T.A.: Časti reči i tipy tekstov [Parts of Speech and Types of Texts]. Riga: Zinatne 1981, 248 pp.

33. JAKUBAJTIS, T.A., SKLJAREVIČ, A.N.: Verojatnostnaja atribucija tipa teksta po morfologičeskomu priznaku [Probabilistic Attribution of Types of Texts According to Their Morphological Features]. Riga: Institut ėlektroniki i vyčislitel'noj texniki Akademii Nauk Latvijskoj SSR, 1981, 67 pp.

34. KOZINCEVA, N.A., KOZINCEV, A.G.: O vyjavlenii obščix zakonomernostej v raznorodnyx tekstax [On the manifestation of general laws in heterogeneous texts]. Piotrovskij, R.G. (Ed.): Statistika reči i avtomatičeskij analiz teksta. Leningrad 1980, 64-71.

35. LINDELL, A., PIIRAINEN, I.T.: Untersuchungen zur Sprache des Wirtschaftsmagazins "Capital". Vaasa: Vaasa School of Economics 1980, 103 pp.

36. MALAXOVSKIJ, L.V.: Strukturnye i kvantitativnye xarakteristiki omonimičeskix rjadov v sovremennom anglijskom jazyke [Structural and quantitative characteristics of homonymous series in contemporary English]. Piotrovskij, R.G. (Ed.): Statistika reči i avtomatičeskij analiz teksta. Leningrad: Nauka, 1980, 145-212.

37. MIZUTANI, S., MATUMORI, T.: Lexical similarities among popular songs of the same subject matters (In Japanese). Mathematical Linguistics 13, 1982, 149-164.

38. MUXAMEDOV, S.A.: Statističeskij analiz leksiko-morfologičeskoj struktury uzbekskix gazetnyx tekstov [Statistical Analysis of the Lexical-Morphological Structure of Uzbek Newspaper Texts]. Avtoreferat dissertacii. Taškent 1980, 25 pp.

39. TOMIK, M.: Jazikot vo literaturnite dela na Blaže Koneski (statistička analiza) [The Language in the Literary Works of Blaže Koneski (Statistical Analysis)]. Skopje 1977, 297 pp.

40. XOVANOV, G.M.: Nekotorye voprosy količestvennogo povtorenija slova v tekste [Some problems of the quantitative word repetition in a text]. Andrjuščenko, V.M. (Ed.): Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva: Izdatel'stvo Moskovskogo universiteta, 1978, 41-58.

41. ŽILINSKENE, V.: Korreljacionnyj i klasternyj analiz častej reči litovskoj publikacii [Correlation and cluster analysis of parts of speech in Lithunian Publications]. Kalbotyra 32, 1981/1, 121-133.

PSYCHOLINGUISTICS

42. ARROYO, F.V.: Negatives in context. JVLVB 21, 1982, 118-126.

43. BATES, E., KINTSCH, W., FLETCHER, C.R., GIULIANI, V.: The role of pronominalization and ellipsis in texts: Some memory experiments. Journal of Experimental Psychology: Human Learning and Memory 6, 1980, 676-691.

44. BOCK, J.K.: Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. Psychological Review 89, 1982, 1-47.

45. BOCK, J.K., BREWER, W.F.: Comprehension and memory of the literal and figurative meaning of proverbs. JPsyR 9, 1980, 59-72.

46. BOCK, J.K., IRWIN, D.E.: Syntactic effects of information availability in sentence production. JVLVB 19, 1980, 467-484.

47. CAIRNS, H.S., HSU, J.R.: Effects of prior context on lexical access during sentence comprehension: A replication and reinterpretation. JPsyR 9, 1980, 319-326.

48. CARPENTER, P.A., DANEMAN, M.: Lexical retrieval and error recovery in reading: A model based on eye fixations. JVLVB 20, 1981, 137-160.

49. CARROLL, J.M.: Creating names for things. JPsyR 10, 1981, 441-455.

50. CIRILO, R.K.: Referential coherence and text structure in story comprehension. JVLVB 20, 1981, 358-367.

51. CIRILO, R.K., FOSS, D.J.: Text structure and reading time for sentences. JVLVB 19, 1980, 96-109.

52. CURRIE, K.L.: An initial "Search for Tonics". L&S 23, 1980, 329-350.

53. DELL, G.S., REICH, P.A.: Stages in sentence production: An analysis of speech error data. JVLVB 20, 1981, 611-629.

54. DREWNOWSKI, A., HEALY, A.F.: Missing '-ing' in reading: Letter detection errors in word endings. JVLVB 19, 1980, 247-262.

55. DRIZULE, V.: Statističeskaja optimizacija prepodavanija

jazykov i inženernaja lingvistika [Engineering linguistics and the statistical optimization of language teaching]. Izvestija Akademii Nauk Latvijskoj SSR, Riga, 1981/4, 140-141.

56. DUNLAP, G.L., HURTIG, R.R.: Effects of clausal structure and word frequency in sentence processing. JPsyR 10, 1981, 313-326.

56. EGEN, O.: Intonation and meaning. JPsyR 9, 1980, 23-39.

57. FEAGANS, L.: Children's understanding of some temporal terms denoting order, duration, and simultaneity. JPsyR 9, 1980, 41-57.

58. FRAUENFELDER, U., SEGUI, J., MEHLER, J.: Monitoring around the relative clause. JVLVB 19, 1980, 328-337.

59. FRAZIER, L., RAYNER, K.: Making and correcting errors during sentence comprehension: Eye movements and the analysis of structurally ambiguous sentences. Cognitive Psychology 14, 1982, 178-210.

60. FRENCH, P.: Semantic and syntactic factors in the perception of rapidly presented sentences. JPsyR 10, 1981, 581-591.

61. FREYD, P., BARON, J.: Individual differences in acquisition of derivational morphology. JVLVB 21, 1982, 282-295.

62. GHIGLIONE, R., BEAUVOIS, J.L.: Speech-markers and attitude toward language as an object. IJPs 8-1 (21), 1981, 51-73.

63. GLUCKSBERG, S., GILDEA, P., BOOKIN, H.B.: On understanding nonliteral speech: Can people ignore metaphors? JVLVB 21, 1982, 85-98.

64. GRAESSER, A.C., HOFFMAN, N.L., CLARK, L.F.: Structural components of reading time. JVLVB 19, 1980, 135-151.

65. HABERLANDT, K., BERIAN, C., SANDSON, J.: The episode schema in story processing. JVLVB 19, 1980, 635-650.

66. HIRST, W., BRILL, G.A.: Contextual aspects of pronoun asignment. JVLVB 19, 1980, 168-175.

67. IRWIN, D.E., BOCK, J.K., STANOVICH, K.E.: Effects of information structure cues on visual word processing. JVLVB 21, 1982, 307-325.

68. JAEGER, J.J.: Testing the psychological reality of phonemes. L&S 23, 1980, 233-253.

69. KEMPER, S.: Filling in the missing links. JVLVB 21, 1982, 99-107.

70. KIERAS, D.E.: Component process in the comprehension of simple prose. JVLVB 20, 1981, 1-23.

71. KOLERS, P.A., GONZALES, E.: Memory for words, synonyms, and translations. Journal of Experimental Psychology: Human Learning and Memory 6, 1980, 53-65.

72. LONGONI, A.M., PIZZAMIGLIO, L.: Aspects of verbal processing in relation to perceptual disembedding ability. JPsyR 10, 1981, 199-208.

73. McDONALD, J.L., CARPENTER, P.A.: Simultaneous translation: Idiom interpretation and parsing heuristics. JVLVB 20, 1981, 231-247.

74. McKOON, G., RATCLIFF, R.: Priming in item recognition: The organization of propositions in memory for text. JVLVB 19, 1980, 369-386.

75. McKOON, G., RATCLIFF, R.: The comprehension processes and memory structures involved in anaphoric reference. JVLVB 19, 1980, 668-682.

76. MEHLER, J., DOMMERGUES, J.Y., FRAUENFELDER, U.: The syllable's role in speech segmentation. JVLVB 20, 1981, 298-305.

77. MICHAM, D.L., CATLIN, J., vanDERVEER, N.J., LOVELAND, K.A.: Lexical and structural cues to quantifier scope relations. JPsyR 9, 1980, 367-377.

78. MOTLEY, M.T., BAARS, B.J., CAMDEN, C.T.: Syntactic criteria in prearticulatory editing: Evidence from laboratory-induced slips of the tongue. JPsyR 10, 1981, 503-522.

79. NEZWORSKI, T., STEIN, N.L., TRABASSO, T.: Story structure versus content in children's recall. JVLVB 21, 1982, 196-206.

80. OLÉRON, P.: Coreference of the personal pronoun and sentence meaning. IJPs 8-1 (21), 1981, 31-50.

81. OMANSON, R.C.: The relation between centrality and story category variation. JVLVB 21, 1982, 326-337.

82. RAPHAEL, L.J., DORMANN, M.F., LIBERMAN, A.M.: On defining the vowel duration that cues voicing in final position. L&S 23, 1980, 297-307.

83. ROTHKOPF, E.C.: Copying span as a measure of the information burden in written language. JVLVB 19, 1980, 562-572.

84. RUBIN, D.C.: 51 properties of 125 words: A unit analysis of verbal behavior. JVLVB 19, 1980, 736-755.

85. SCHOLL, D.M., RYAN, E.B.: Development of metalinguistic performance in the early school years. L&S 23, 1980, 199-211.

86. SCHWARZ, M.N.K., FLAMMER, A.: Text structure and title - effects on comprehension and recall. JVLVB 20, 1981, 61-66.

87. SHERBLOM, J., REINSCH, N.L. jr.: Persuasive intent as a determinant of phonemic choice. JPsyR 10,1981, 619-628.

88. SIMPSON, G.B.: Meaning dominance and semantic context in the processing of lexical ambiguity. JVLVB 20, 1981, 120-136.

89. SPENCER, N.J., WOLLMAN, N.: Lexical access for phonetic ambiguities. L&S 23, 1980, 171-198.

90. SPIRO, R.J.: Accomodative reconstruction in prose recall. JVLVB 19, 1980, 84-95.

91. TARTE, R.D.: The relationship between monosyllables and pure tones: An investigation of phonetic symbolism. JVLVB 21, 1982, 352-360.

92. VIPOND, D.: MICRO- and macroprocesses in text comprehension. JVLVB 19, 1980, 276-296.

93. VUCHINICH, S.: Logical relations and comprehension in conversation. JPsyR 9, 1980, 473-501.

94. WATERS, H.S.: "Class News": A single-subject longitudinal study of prose production and schema formation during childhood. JVLVB 19, 1980, 152-167.

95. YEKOVICH, R.R., THORNDYKE, P.W.: An evaluation of alternative functional models of narrative schemata. JVLVB 20, 1981, 454-469.

Historical Linguistics

96. ZAXAROVA, V.P.: Verojatnostno-statističeskaja xarakteristika refleksov praslavjanskix DJ i TJ v Lavrent'evskoj letopisi [Probabilistic-statistical characteristics of the reflexes of Common Slavonic DJ and TJ in the chronicle of Lavrentij]. Ėvoljucija i predystorija russkogo jazykovogo stroja. Gor'kij 1980, 105-109.

97. ZAXAROVA, V.P.: Verojatnostno-statističeskij analiz refleksov praslavjanskix dj i tj v povestvovatel'nyx pis'mennyx tekstax nižegorodcev XIX-XX vv. [Probabilistic-statistical analysis of the reflexes of Common Slavic dj and tj in narrative written texts of the inhabitants of Nižegorod in the 19th and 20th century]. Problemy istorii kul'tury Volgo-Vjatskogo regiona. Gor'kij 1981, 70-88.

98. ZAXAROVA, V.P.: Verojatnostno-statističeskaja xarakteristika refleksov praslavjanskogo v "Žitii" protopopa Avvakuma [Probabilistic-statistical characteristics of the Common Slavic reflexes in "The Life" of Protopope Avvakum]. Ėvoljucija i predystorija russkogo jazykovogo stroja. Gor'kij 1981, 66-71.

Language Variation

99. AMONOVA, F.: Ob ispol'zovanii lingvostatističeskix metodov v sravnitel'no-tipologičeskom issledovanii persidskogo i tadžikskogo jazykov (Na materiale slovoobrazovanija) [Application of linguistic-statistical methods in the comparative-typological analysis of Persian and Tadzhik (Based on word-formation)]. Problemy vostočnogo istočnikovedenija. Dušanbe 1980, 97-101.

100. KAŠEVIČ, V.B., JAXONTOV, S.E. (otv. red.): Kvantitativnaja tipologija jazykov Azii i Afriki [A quantitative typology of the languages of Asia and Africa]. Leningrad: Izdatel'stvo Leningradskogo universiteta, 1982, pp. 331.

101. KOTOVA, N.V.: Glottometrija i sopostavitel'nyj sintaksis rodstvennyx jazykov (Na materiale russkogo i bolgarskogo jazykov) [Glottometry and contrastive syntax of related languages (Based on Russian and Bulgarian)]. Vestnik Moskovskogo universiteta, Serija 9, Filologija 1981/4, 33-42.

Documentation

102. KRAUSE, J.: Computational linguistics in Western Germany: A survey. Symposium 1981, 17-25.

103. PALL, V., REMMEL, M.: Ethnogeographical (including dialectological) data processing at the Institute of Language and Literature, Academy of Sciences Estonian S. S.R. Symposium 1981, 87-92.

Dialectology

104. GIRDENIS, A.: Častota fonem severožemajtskogo narečija [Phonem frequency in the North-Zhemajtskij dialect]. Kalbotyra 32, 1981/1, 15-37.

105. GOEBL, H.: Éléments d'analyse dialectométrique (avec application à l'AIS). Revue de Linguistique Romane 45, 1981, 349-420.

106. GOEBL, H.: Dialektometrie - Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Österreichische Akademie der Wissenschaften, Philosophisch-Historische Klasse. Denkschriften, 157. Band, Wien 1982, pp. 123.

107. KRIKMANN, A.: Some aspects of proverb distribution. Symposium 1979, 28-44.

108. KURKINA, G.G.: Otnositel'naja častotnost' glasnyx v jazyke Kazymskix Xanty [Relative frequency of vowels in the language of the Kazym Khanty]. Zvukovoj stroj sibirskix jazykov. Novosibirsk 1980, 66-71.

109. KUZNETSOVA, E.L.: Computer-based interpretation of linguistic maps. Symposium 1979, 45-46.

110. LEKOMCEVA, M.I.: On interpretation of linguistic maps. Symposium 1979, 52-55.

111. LEKOMTSEVA, M.I.: Preliminaries to a model of boundary formation in dialectology. Symposium 1981, 26-35.

112. LIZANETS, P.N., PESTCHAK, M.M.: Automatic compiling of dialectological maps. Symposium 1979, 56-58.

113. LONN, V.: Die gegenseitigen Beziehungen der Mundarten der Insel Saaremaa auf der Grundlage der taxonomischen Klassifikation. Symposium 1981, 36-42.

114. MAMSUROVA, E.N.: Statistical isogloss method of map treatment. Symposium 1979, 59-62.

115. MAMSUROVA, E.N.: O metode statističeskix izogloss [The method of statistical isoglosses]. Lingvističeskaja geografija i problemy istorii jazyka: Materialy Šestogo Vsesojuznogo soveščanija po obščim voprosam dialektologii i istorii jazyka, posvjaščennogo 60-letiju Oktjabr'skoj Socialističeskoj Revoljucii. Nal'čik 1980, c.1, 144-150.

116. MURUMETS, S.: Towards automatic auditing of discrete maps of proverb distribution. Symposium 1979, 67-69.

117. MURUMETS, S.: On measuring interregional linguistic communication. Symposium 1981, 43-80.

118. NIIT, E.: The place of Estonian dialects in the Baltic prosodic area. Symposium 1981, 81-86.

119. PŠENIČNOVA, N.N.: Nekotorye sposoby gruppirovanija ob"ektov, primenjaemye v dialektologii [Some methods of grouping objects, applied in dialectology]. Problemy strukturnoj lingvistiki 1979. Moskva: Nauka 1981, 278-286.

120. PSHENICHNOVA, N.N.: Data statistical analysis as a basis for identification and interpretation of linguogeographical areas. Symposium 1979, 70-73.

121. PUTSCHKE, W.: Automatische Sprachkartographie: Verfahren, Anwendungen und Perspektiven. Symposium 1981, 93-140.

122. REMMEL, M.: On employing phonostatistical regularities in dialectology. Symposium 1981, 141-143.

123. ROLSHOVEN, J.: Quantitative Phonologie des Ampezzanischen. Kramer, J. (Ed.): Studien zum Ampezzanischen. Romanica Aenipontana 9, 1978, 59-176.

124. TAMBOVCEV, Ju.A.: Častotnye xarakteristiki glasnyx pervogo sloga mansijskogo jazyka [Frequency characteristics of first syllable vowels in the Mansi language]. Zvukovoj stroj sibirskix jazykov. Novosibirsk 1980, 66-71.

All contributions to the CURRENT BIBLIOGRAPHY should be sent to Prof.Dr. Werner Lehfeldt, Universität Konstanz, Fachgruppe Sprachwissenschaft, P.O.Box 5560, D 7750 Konstanz

ANNOUNCEMENT OF A PROJECT

Within the context of a DFG-supported project the Department of Speech and Communication of the University of Essen is compiling a Bibliography of Quantitative Linguistics. The project was started in May 1982. Its goal is a comprehensive documentation of all publications and of all earlier data-files concerned with quantitative linguistics. Data storage and processing are computer-based. The complete bibliography will be published in three formats:

- in book-form,
- on micro-fiches,
- as a computer processable data-bank.

The overall goal, i.e. an exhaustive documentation of all pertinent publications cannot be attained by a thorough scanning of all extant bibliographical sources alone; goal attainment, rather, depends on the active cooperation of the greatest possible number of colleagues working in the area of quantitative linguistics. This is why the project organizers urge all authors in the field to make pertinent publications and files available to one of the following contacts:

-Bibliographiesystem QL
R.Köhler/D.Krallmann
Universität Essen GHS
Fachbereich 3
Universitätsstr.12
D-4300 Essen 1

-Doz.Dr.J.Sambor
Instytut Języka Polskiego
Uniwersytet Warszawski
Krakowskie Przedmiescie 26/28
00-325 Warszawa
Poland

-Dr.R.Grotjahn
Seminar für Sprachlehrforschung
Ruhr-Universität Bochum
Postfach 102148
D-4630 Bochum

-Dr.B.Rieger
Germanistisches Institut
RWTH
Annuntiatenbach
D-5100 Aachen

-Prof.Dr.W.Lehfeld
Universität Konstanz
Philosophische Fakultät
Postfach 5560
D-7750 Konstanz

-Prof.Dr.M.V.Arapov
VINITI AN SSSR
Ul. Baltiskaja 14
125219 Moskva
USSR