QUANTITATIVE LINGUISTICS

Vol. 35

GLOTTOMETRIKA 9

edited

by

Klaus-Peter Schulz

Studienverlag Dr. N. Brockmeyer
Bochum 1988

CIP-Titelaufnahme der Deutschen Bibliothek

Glottometrika ... – Bochum: Studienverl. Brockmeyer.
Früher mehrbd. begrenztes Werk
ISSN 0932-7991

9 (1988)
(Quantitative linguistics; Vol. 35)
ISBN 3-88339-648-6

NE: GT

ISBN 3-88339-648-6

Querenburger Höhe 281, 4630 Bochum 1
Gesamtherstellung: Druck Thiebes GmbH & Co. KG Hagen

Contents

Reviews

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 1–47

Lautliche Strukturierung von Spracheinheiten*

K.-P. Schulz
G. Altmann
Bochum

1. Einführung

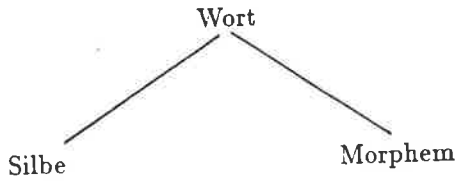
Die vorliegende Arbeit stellt sich das Ziel, einige statistische Methoden vorzustellen, die der Erfassung latenter Strukturierungen sprachlicher Einheiten dienlich sein könnten. Es handelt sich dabei nicht um Erscheinungen, die man unter den "Regeln" einer Grammatik oder Phonologie findet, sondern eher um Tendenzen, die eine gewisse Ausprägung aufweisen müssen, um als solche erkannt zu werden.

Die Erkennung von Tendenzen ist nur der erste Schritt unter die Oberfläche der Sprache. Jedoch erst wenn die Tendenzen bekannt sind, kann man Hypothesen über stochastische Regularitäten aufstellen. Diese wiederum sollten dazu anregen, Hypothesen deduktiv zu gewinnen, sie mit anderen Hypothesen in Verbindung zu bringen und schließlich, nach hinreichender Überprüfung, Gesetze zu etablieren.

Wir gehen davon aus, daß jede Sprache sehr typische, nach irgendeinem Muster gebaute Silben, Morpheme und Wörter hat. Hinter diesem Muster können sprachspezifische "Vorlieben" oder Restriktionen stecken, es kann sich aber auch um allgemeinere Muster handeln, die möglicherweise für alle Sprachen gelten. Die Datenbasis ist vorläufig zu gering, um an diese allgemeineren Muster heranzukommen, es sei denn, man begnügt sich mit vagen, verbal formulierten empirischen Generalisierungen, wie es in der Typologie oder in der Universalienforschung üblich ist.

* Diese Studie entstand im Rahmen des Projekts "Sprachliche Synergetik". Die Autoren bedanken sich bei der **Stiftung Volkswagenwerk** für die freundliche Unterstützung.

Weiter gehen wir davon aus, daß Spracheinheiten Subsysteme des Systems "Sprache" sind, wobei das Wort hierarchisch höher steht als Silbe und Morphem, da diese Komponenten des Wortes sind:



Morpheme und Silben stehen auf unterschiedlichen ontologischen Ebenen, gelegentlich fallen sie aber zusammen; in manchen Sprachen ist es die Regel (z.B. Chinesisch). Das System "Wort" übt auf seine Subsysteme eine Wirkung aus, wie es bei allen System/Subsystem Beziehungen der Fall ist (vgl. Bowler 1981 passim). Diese Wirkung nennt man Dominanz des Systems oder in synergetischen Begriffen "Versklavung" (vgl. Haken 1978). Die Größe, die diese Wirkung ausübt, bezeichnet Haken als Ordnungsparameter. Es kann gleichzeitig mehrere Ordnungsparameter geben, bisher wurde als solcher nur die Wortlänge erkannt, die in Form des Menzerathschen Gesetzes die Silbenlänge und die Morphemlänge steuert (vgl. Altmann 1980, Gerlach 1982, Geršić, Altmann 1980, Altmann, Schwibbe, Kaumanns, Köhler, Wilde 1987). Die Silben untereinander und die Morpheme untereinander bilden kollaterale (Sub)Systeme, die in bestimmten Bereichen miteinander konkurrieren oder kooperieren, jenachdem, ob sie die eigene Autonomie erweitern wollen oder ob sie der Aufrechterhaltung des (Sub)Systems dienen.

Im semantischen Bereich teilen sich die Morpheme untereinander die Erfassung der menschlichen Realität, wobei sich manche recht autonom, konkurrierend, verhalten und Bedeutungen anhäufen (dies bezeichnet man üblicherweise als die Diversifikation durch die Kraft des Sprechers), andere verhalten sich kooperativ, integrativ, und spezialisieren ihre Bedeutung (Unifikation durch die Kraft des Hörers). Beide Kräfte zusammen bezeichnet man als Zipfsche Kräfte.

Im lautlichen Bereich verhalten sich die Morpheme auch ähnlich: sie organisieren ihre Form nach übergeordneten Gesetzen - die wir noch nicht kennen - und greifen nur selten zu systemfremden Gestalten; die Dominanz des Systems ist hier offensichtlich. Gleichzeitig sorgen sie

für hinreichende gegenseitige Unterscheidung mit inherentem Autonomiedrang.

Die systemische Rolle der Silben liegt nur im lautlichen Bereich, wo sie sich ähnlich wie Morpheme verhalten.

Diese etwas anthropomorphische Darstellung sollte nicht den Eindruck erwecken, daß Spracheinheiten etwas selbständig "tun". Das alles wird von dem Idiolektträger, der auch ein Subsystem der Sprache darstellt, durchgeführt. Er ist derjenige, der die lautlichen Formen der Spracheinheiten erzeugt; dabei handelt er keineswegs chaotisch, sondern nach irgendwelchen Gesetzen, die er explizit nicht kennt, aber notgedrungen befolgen muß. Die Analogie mit der "Befolgung" von Naturgesetzen ist hier offensichtlich.

Die vorliegende Arbeit soll als Ansporn dienen, nach derartigen Gesetzen zu suchen.

Beim Aufbau der Spracheinheiten (Wort, Silbe, Morphem) werden aus n Lauten/Phonemen des Inventars immer jeweils k "gewählt" und permutiert, wobei die Laute in der Spracheinheit wiederholt auftreten dürfen. Kombinatorisch gesehen geht es also um Variationen mit Wiederholung, die die theoretisch mögliche Anzahl von Spracheinheiten der Länge k darstellen.

Diese theoretische Zahl unterliegt in der Empirie starken Beschränkungen. Am Beispiel semitischer Wurzeln zeigte Herdan (1966:194-198), wie man die Zahl der Einschränkungen aufgrund distributionaler Regeln ermittelt; jedoch stellte er fest, daß auch nach Abzug dieser Fälle nur ein Teil der restlichen theoretisch möglichen (zulässigen) Wurzeln verwendet wird. Wir nehmen an, daß diese Anzahl verwirklichter Einheiten (bzw. ihr Anteil an möglichen) keineswegs ein purer Zufall ist, sondern in allen Sprachen einem universalen Gesetz folgt, das wir noch nicht kennen. Die Unterschiede zwischen Sprachen lassen sich als Unterschiede zwischen den Parametern dieses Gesetzes ausdrücken. Das Auffinden dieses Gesetzes ist sicherlich nur eine Frage der Zeit. Wir wollen im weiteren nur seine Begleitumstände untersuchen.

Wenn sich eine Sprache diesem Gesetz beugt, dann muß sie zahlreiche kombinatorische Einschränkungen einführen, von denen ein Teil deterministischer Natur ist und manchmal in Lehrbüchern aufgeführt wird, z.B. "keine stimmhaften Verschlußlaute am Wortende" im Deut-

schen, während zahlreiche andere stochastischer Natur sind und die moderne Phonologie nicht weiter interessieren. Und gerade diese sind es, die das Geschehen der Konstruktion von Spracheinheiten steuern, Fluktuationen ermöglichen und die Entwicklung der Sprache in einem selbstregulierenden Prozess vorantreiben.

Im folgenden werden wir nur derartige stochastische Musterbildungen untersuchen.

2. Unabhängigkeit

Es wird allgemein akzeptiert, daß die Laute/Phoneme in einer Einheit nicht unabhängig voneinander plaziert werden, aber es gibt nur wenige Untersuchungen, die es versucht haben, eine der zahlreichen Unabhängigkeitshypothesen zu falsifizieren (vgl. Ross 1952, Krupa 1966, 1967a,b, 1971, Altmann 1964, 1973, Altmann, Lehfeldt 1980). Es wurden zahlreiche Daten gesammelt, jedoch lediglich für deskriptive Zwecke (vgl. Uhlenbeck 1949, 1950, Dubovska s.d., Odendal 1962, Greenberg 1950).

Die Problematik werden wir an einem sehr einfachen Beispiel darstellen, um Kontrollrechnungen zu ermöglichen.

Aus dem Wörterbuch der Hanunoo-Sprache (Philippinen) von Conklin (1953) wurden 2767 zweisilbige Grundwörter erhoben und die Anzahl der Kombinationen der beiden Vokale (erste und zweite Silbe) berechnet, wie in der Tabelle 2.1 dargestellt.

Tabelle 2.1.
Vokalkombinationen in zweisilbigen Grundwörtern des Hanunoo

		Vokal der zweiten Silbe			
		a	i	u	$\sum$
Vokal der ersten Silbe	a	698	224	425	1347
	i	221	143	151	525
	u	324	96	485	905
	$\sum$	1243	463	1061	2761

Unser Problem ist die Frage, ob die Verwendung des Vokals in der zweiten Silbe vom Vokal in der ersten Silbe abhängt bzw. umgekehrt

die Verwendung des Vokals in der zweiten Silbe vom Vokal in der ersten Silbe. Das bedeutet, es ist (statistisch) zu testen, ob derartige Zusammenhänge tatsächlich bestehen oder ob es sich dabei lediglich um zufällige Schwankungen handelt. Wir bezeichnen dann

mit n_{ij} die absoluten Häufigkeiten in den einzelnen Zellen der Tafel (i = 1,. . .,r; j = 1,. . .,s)

mit $n_{i.}$ die i-te Zeilensumme

$$n_{i.} = \sum_{j=1}^{s} n_{ij}$$

(i = 1,. . .,r)

mit $n_{.j}$ die j-te Spaltensumme

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

(j = 1,. . .,s)

und mit n die Gesamtzahl der Beobachtungen

$$n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$$

Dies läßt sich schematisch in einer Kontingenztafel darstellen (vgl. Tab.2.2).

Tabelle 2.2.
r × s-Kontingenztafel

	1	2	.....	s	$\sum$
1	n_{11}	n_{12}	.....	n_{1s}	$n_{1.}$
2	n_{12}	n_{22}	.....	n_{2s}	$n_{2.}$
:	:	:	:	:	:
:	:	:	:	:	:
$\sum$	$n_{.1}$	$n_{.2}$	.....	$n_{.s}$	n

Zu dieser Tafel der beobachteten Häufigkeiten korrespondiert eine Tafel der theoretischen Wahrscheinlichkeiten p_{ij} bzw. der erwarteten Häufigkeiten $m_{ij} = np_{ij}$ (vgl. Tab.2.2 a,b).

Tabelle 2.2a

Zu Tab.2.2. korrespondierende Tafel der theoretischen Wahrscheinlichkeiten p_{ij}

	1	2	.....	s	$\sum$
1	p_{11}	p_{12}	.....	p_{1s}	$p_{1.}$
2	p_{21}	p_{22}	.....	p_{2s}	$p_{2.}$
:	:	:	:	:	:
:	:	:	:	:	:
$\sum$	$p_{.1}$	$p_{.2}$	.....	$p_{.s}$	p

Tabelle 2.2b

Zu Tab.2.2 korrespondierende Tafel der erwarteten Häufigkeiten

	1	2	.....	s	$\sum$
1	m_{11}	m_{12}	.....	m_{1s}	$m_{1.}$
2	m_{12}	m_{22}	.....	m_{2s}	$m_{2.}$
:	:	:	:	:	:
:	:	:	:	:	:
$\sum$	$m_{.1}$	$m_{.2}$	.....	$m_{.s}$	n

Die Untersuchung auf Unabhängigkeit zwischen zwei Merkmalen (hier:zwischen Vokal in der ersten Silbe und Vokal in der zweiten Silbe) führt zu folgendem statistischem Testproblem:

$$H_o^u : p_{ij} = p_{i.}p_{.j} \text{ für alle } (i,j)$$

gegen

$$H_1^u : p_{ij} \neq p_{i.}p_{.j} \text{ für mindestens ein Paar } (i,j).$$

Das Testverfahren besteht nun darin, die erwarteten Häufigkeiten mit den tatsächlich beobachteten Häufigkeiten mit Hilfe einer Teststatistik zu vergleichen. Dabei stehen zur Auswahl

-die Pearsonsche χ^2-Statistik:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2.1}$$

-die Likelihood-Ratio-Statistik:

$$G^2 = 2 \sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{m_{ij}} \tag{2.2}$$

Beide Statistiken sind asymptotisch chiquadrat-verteilt mit (r-1)(s-1) Freiheitsgraden (vgl. u.a. Bishop, Fienberg, Holland 1975). Wegen der asymptotisch gleichen Eigenschaften beider Teststatistiken werden wir im folgenden der Einfachheit halber nur noch auf die χ^2-Statistik zurückgreifen.

Die erwarteten Häufigkeiten m_{ij} müssen geschätzt werden. Unter der Unabhängigkeitshypothese H_o^u ergibt sich

$$\hat{m}_{ij} = n_{i.}n_{.j}/n$$

als Maximum-Likelihood-Schätzer für m_{ij} (vgl. u.a. Bishop, Fienberg, Holland 1975). Setzt man dies in (2.1) ein, so erhält man als χ^2-Teststatistik für Unabhängigkeit:

$$\chi^2 = \sum_i \sum_j \left[\frac{(n_{ij} - (n_{i.}n_{.j}/n))^2}{n_{i.}n_{.j}/n} \right] = n \sum_i \sum_j \frac{n_{ij}^2}{n_{1.}n_{.j}} - n. \tag{2.3.}$$

Für Hanunoo ergibt sich

$$\chi^2 = 2767 \left[\frac{698^2}{1347(1243)} + \frac{224}{1347(463)} + \ldots + \frac{485^2}{905(1061)} \right]$$

$$-2767 = 171.27.$$

χ^2 ist chiquadrat-verteilt mit (3-1)(3-1) = 4 Freiheitsgraden.

Aus einer Tabelle für die Chiquadrat-Verteilung findet man $\chi^2_{4;0.99} = 13.28 (\chi^2_{4;0.95} = 9.49)$. Da dieser Wert vom Wert der Teststatistik übertroffen wird, lehnt man die Unabhängigkeitshypothese zum Niveau 0.01 (0.05) ab.

Der obige Test zeigt, daß die Plazierung der Vokale voneinander nicht unabhängig ist, sagt aber nichts über die Stärke der Abhängigkeit aus. Oft möchte man daher die Abhängigkeitsstruktur einer Kontingenztafel durch eine einzige reelle Zahl beschreiben. Diese wird meistens so normiert, daß sie nur Werte zwischen 0 und 1 annehmen kann, und zwar den Minimalwert 0 bei totaler Unabhängigkeit und den Maximalwert bei totaler Abhängigkeit. Da eine solche Zahl jedoch nicht alle Aspekte des Zusammenhangs beschreiben kann, sind im Laufe der Zeit viele solcher Assoziationsmaße vorgeschlagen worden. Die bekanntesten sind (vgl. Goodman, Kruskal 1954) der Pearsonsche Kontingenzkoeffizient, die Assoziationsmaße von Tschuprov und Cramér sowie die Lambda-Maße von Guttman und die Tau-Maße von Goodman und Kruskal. Diese werden im folgenden vorgestellt. Naheliegend ist es, Assoziationsmaße zu verwenden, die auf der χ^2-Teststatistik beruhen. χ^2 selbst eignet sich nicht gut, da seine Größe vom Stichprobenumfang n abhängig ist. Daher benutzt man oft den *Pearsonschen Kontingenzkoeffizienten* C, der definiert ist als

$$C = \left[\frac{\chi^2}{\chi^2 + n}\right]^{1/2}. \tag{2.4}$$

C liegt immer zwischen 0 und 1, nimmt als maximalen Wert aber nicht 1 an, sondern ist in einer r × s-Tafel höchstens gleich

$$\left[\frac{min(r,s) - 1}{min(r,s)}\right]^{1/2}.$$

(Für $r, s \to \infty$ wird der maximal mögliche Wert für C gegen 1 streben.) Deshalb verwendet man gelegentlich den *korrigierten Pearsonschen Kontingenzkoeffizienten*

$$C_{corr} = C\left[\frac{min(r,s)}{min(r,s) - 1}\right]^{1/2}, \tag{2.5}$$

der bei jeder Tafelgröße 1 als maximalen Wert hat. Selbstverständlich gilt immer $C < C_{corr}$.

Das *Tschuprovsche Assoziationsmaß* T ist definiert als

$$T = \left[\frac{\chi^2}{n[(r-s)(s-1)]^{1/2}}\right]^{1/2}. \tag{2.6}$$

T nimmt den Maximalwert 1 nur bei quadratischen Tafeln (r=s) an.

Das *Cramérsche Assoziationsmaß*

$$V = \left[\frac{\chi^2}{n(min(r,s) - 1)}\right]^{1/2} \tag{2.7}$$

hat dagegen stets 1 als maximal möglichen Wert.
Für quadratische Tafeln gilt immer T = V, für r ≠ s ist T < V.

Als Schätzungen für die Varianzen der oben angegebenen Assoziationsmaße ergeben sich (vgl. Hartung 1984; Bishop, Fienberg, Holland 1975; Kendall, Stuart 1967) mit

$$\frac{Var\chi^2}{n^2} = 4\sum_i\sum_j \frac{n_{ij}^3}{n_{i.}^2 n_{.i}^2} - 3\sum_i \frac{1}{n_{i.}}\left[\sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}}\right]^2$$

$$-3\sum_j \frac{1}{n_{.j}}\left[\sum_i \frac{n_{ij}^2}{n_{i.}n_{.j}}\right]^2 + 2\sum_i\sum_j \frac{n_{ij}}{n_{i.}n_{.j}}\left[\sum_k \frac{n_{kj}^2}{n_{k.}n_{.j}}\right]$$

$$\cdot\left[\sum_l \frac{n_{il}^2}{n_{i.}n_{.j}}\right] \tag{2.8}$$

falls $\chi^2 \neq 0$:

$$VarC = \frac{n^4}{4\chi^2(n+\chi^2)^3}\frac{Var\chi^2}{n^2}, \tag{2.9}$$

$$VarC_{corr} = \frac{n^4[min(r,s)]}{[min(r,s) - 1]\cdot 4\chi^2(n+\chi^2)^3}\cdot\frac{Var\chi^2}{n^2}, \tag{2.10}$$

$$VarT = \frac{1}{4(r-1)(s-1)T^2} \cdot \frac{Var\chi^2}{n^2}, \qquad (2.11)$$

$$VarV = \frac{1}{4[min(r,s)-1]^2V^2} \cdot \frac{Var\chi^2}{n^2}. \qquad (2.12)$$

Die Schätzungen für die Varianzen werden benötigt, um Konfidenzintervalle für die Assoziationsmaße anzugeben. Die Konfidenzintervalle bilden wir folgendermaßen: Sei F irgendein Kontingenzkoeffizient (C,C_{corr},T,V), dann wird das 95%-Intervall dargestellt durch die untere Grenze

$$\underline{F} = F + z_{0.025}(VarF)^{1/2}$$

und die obere Grenze

$$\bar{F} = F + z_{0.975}(VarF)^{1/2}.$$

Am Beispiel von Hanunoo (vgl. Tabelle 2.1) ergeben sich für die verschiedenen Assoziationsmaße folgende Werte:

$$C = \left[\frac{171.27}{171.27+2767}\right]^{1/2} = 0.2412$$

$$C_{corr} = 0.2414\left[\frac{3}{3-1}\right]^{1/2} = 0.2957$$

$$T = V = \left[\frac{171.27}{2767(3-1)}\right]^{1/2} = 0.1759.$$

Für das 95%-Konfidenzintervall von C erhält man näherungsweise mit

$$\frac{Var\chi^2}{n^2} = 0.001627$$

und

$$VarC = \frac{2767^4}{4(171.27)(2767+171.27)^3} \cdot 0.001672 = 0.005487$$

$$\underline{C} = 0.2414 - 1.96(VarC)^{1/2} = 0.0962$$

und

$$\bar{C} = 0.2414 + 1.96(VarC)^{1/2} = 0.3866.$$

Als 95%-Konfidenzintervalle für C_{corr}, T und V ergeben sich näherungsweise

[0.1179; 0.4735] für C_{corr}

und

[0.0636; 0.2883] für T und V.

Von den oben besprochenen Assoziationsmaßen (die allesamt Funktionen von χ^2 sind), wissen wir, daß sie den Wert 0 bei totaler Unabhängigkeit und den möglichen Maximalwert bei totaler Abhängigkeit annehmen. Eine wahrscheinlichkeitstheoretische Interpretation der Zwischenwerte ist jedoch nicht möglich. Dagegen erlauben die von Guttman vorgeschlagenen Lambda-Maße und die von Goodman und Kruskal entwickelten Tau-Maße eine solche Interpretation. Den Ausgangspunkt bei der Entwicklung dieser Maße bildet die Frage: Was nützt uns die Kenntnis des einen Merkmals bei der Vorhersage des anderen? Konkret: Für Hanunoo soll der Vokal der zweiten Silbe erraten werden, und zwar einmal bei Unkenntnis und einmal bei Kenntnis des Vokals der ersten Silbe. Man könnte dann fragen, um wieviel das Ergebnis des Ratens bei Kenntnis des ersten Vokals gegenüber dem Raten ohne Vorkenntnis verbessert wird. Unabhängigkeit nützt dabei sicher nichts, bei bekannten Abhängigkeiten erhöht sich dagegen die Vorhersagegenauigkeit. Nennen wir im folgenden die beiden Merkmale einmal A und B.

Die Lambda-Maße von Guttman

Die Lambda-Maße lassen sich dann folgendermaßen erklären: Es sei $V_{U(A)}$ die Wahrscheinlichkeit einer falschen Vorhersage von B bei Unkenntnis von A,

$V_{K(A)}$ die Wahrscheinlichkeit einer falschen Vorhersage von B bei Kenntnis von A.

Das Maß λ_B vergleicht nun diese beiden Wahrscheinlichkeiten miteinander. Es ist definiert als

$$\lambda_B = \frac{V_{U(A)} - V_{K(A)}}{V_{U(A)}}.$$

Es gilt

$$V_{U(A)} = 1 - max_j p_{.j},$$

denn die Wahrscheinlichkeit einer richtigen Prognose von B bei Kenntnis von A ist

$$max_j p_{.j}.$$

Kennt man A schon, hat man etwa A = i beobachtet, so wird das Auftreten von B durch die bedingten Wahrscheinlichkeiten $p_{ij}/p_{i.}$ beschrieben. Es gilt daher

$$V_{K(A)} = 1 - \sum_i p_{i.} max_j p_{ij}/p_{i.} = 1 - \sum_i max_j p_{ij}.$$

Daraus ergibt sich

$$\lambda_B = \frac{[1 - max_j p_{.j}] - [1 - \sum_i max_j p_{ij}]}{1 - max_j p_{.j}} = \frac{\sum_i max_j p_{ij} - max_j p_{.j}}{1 - max_j p_{.j}}.$$

Ersetzt man die theoretischen Wahrscheinlichkeiten durch ihre Schätzwerte so erhält man

$$\lambda_B = \frac{\sum_i max_j n_{ij} - max_j n_{.j}}{n - max_j n_{.j}}, \qquad (2.13)$$

wobei $max_j n_{.j}$ die größte Zahl der unteren Randsumme und $\sum_i max_j n_{ij}$ die Summe der größten Zahlen in den einzelnen Zeilen der Tabelle 2.1 ist.
Vertauscht man die Rollen der Merkmale A und B, wird also A mit bzw. ohne Kenntnis von B prognostiziert, so erhält man das Maß

$$\lambda_A = \frac{\sum_j max_i n_{ij} - max_i n_{i.}}{n - max_i n_{i.}} \qquad (2.14)$$

Sind die Lambda-Maße weder Null noch Eins, so ergibt sich als Schätzwert für die Varianzen (vgl. Hartung 1984, Kap. VII, Goodman, Kruskal 1963) Formel 2.15:
Var(λ_B)=

$$\frac{[n - \sum_i max_j n_{ij}]\,[\sum_i max_j n_{ij} + max_j n_{.j} - 2\sum_i^* max_j n_{ij}]}{[n - max_j n_{.j}]^3},$$

wobei mit $\sum_i^*$ die Summation über solche Zeilen bezeichnet wird, in denen $max_j n_{ij}$ in der gleichen Spalte liegt wie $max_j n_{.j}$.
Analog dazuFormel 2.16:
Var(λ_A)=

$$\frac{\left[n - \sum_j max_i n_{ij}\right]\left[\sum_j max_i n_{ij} + max_i n_{i.} - 2\sum_j^* max_i n_{ij}\right]}{[n - max_i n_{i.}]^3},$$

$\sum_i^*$ bezeichnet dabei die Summation über diejenigen Spalten, in denen $max_i n_{ij}$ in der gleichen Zeile wie $max_i n_{i.}$ liegt.

Mit Hilfe der obigen Varianzschätzungen lassen sich Konfidenzintervalle für λ_B und λ_A angeben.
Für Hanunoo ergeben sich z. B. die folgenden Werte:

$$\lambda_B = \frac{(698 + 221 + 485) - 1243}{2767 - 1243} = 0.1056$$

$$\lambda_A = \frac{(698 + 224 + 485) - 1347}{2767 - 1347} = 0.0423$$

$$Var(\lambda_B) = \frac{(2767 - 1404) + [1404 + 1243 - 2(698 + 221)]}{(2767 - 1243)^3} = 0.0003115$$

Daraus ergibt sich das 95%-Konfidenzintervall für λ_B als

$$\bar{\lambda}_B = \lambda_B + z_{0.975}[Var(\lambda_B)]^{1/2} = 0.1056 + 1.96(0.01765) = 0.1402$$

und

$$\underline{\lambda}_B = \lambda_B + z_{0.025}[Var(\lambda_B)]^{1/2} = 0.1056 - 1.96(0.01765) = 0.0710$$

d.h. [0.0710; 0.1402] für λ_B.
Analog erhält man [0.0015; 0.0830] als 95%-Konfidenzintervall für λ_A.

Die Tau-Maße von Goodman/Kruskal

Die Überlegung, wie man optimale Vorhersagen treffen könnte, führte zu den Lambda-Maßen. Dabei dürfte allerdings die Verteilung der vorhergesagten Klassen im allgemeinen nicht die Verteilung der dann tatsächlich aufgetretenen Klassen wiederspiegeln. Um das zu erreichen, müssen wir bei der Vorhersage des einen Merkmals, etwa B, wie folgt verfahren: Ist A unbekannt, so wird die Merkmalsausprägung $B = j$ mit Wahrscheinlichkeit $p_{.j}$ prognostiziert. Ist A bekannt, etwa $A = i$, prognostiziert man $B = j$ mit der Wahrscheinlichkeit $p_{ij}/p_{i.}$. Durch diese Vorgehensweise wird bei unbekanntem A die Verteilung von B und bei bekanntem A die gemeinsame Verteilung von A und B rekonstruiert. Damit ist die Wahrscheinlichkeit für eine falsche Vorhersage bei Unkenntnis von A

$$W_{U(A)} = 1 - \sum_j p_{.j^2}$$

und die Wahrscheinlichkeit für eine falsche Vorhersage bei Kenntnis von A

$$W_{K(A)} = 1 - \sum_i \sum_j p_{ij^2}/p_{i.}.$$

Der Vergleich dieser beiden Wahrscheinlichkeiten führt zu dem Assoziationsmaß τ_B. Es ist

$$\tau_B = \frac{W_{U(A)} - W_{K(A)}}{W_{K(A)}} = \frac{\sum_i \sum_j p_{ij}^2/p_{i.} - \sum_j p_{.j}^2}{1 - \sum p_{.j}^2}$$

beziehungsweise

$$\tau_B = \frac{\sum_i (1/n_{i.}) \sum_j n_{ij}^2 - (1/n) \sum_j n_{.j}^2}{n - (1/n) \sum_j n_{.j}^2}$$

$$= \frac{\sum_i \sum_j \frac{[n_{ij} - (n_{i.} n_{.j})/n]^2}{n_{i.}}}{n - (1/n) \sum_j n_{.j}^2}. \tag{2.17}$$

Vertauscht man die Rollen von A und B, so kommt man zu dem Maß

$$\tau_A = \frac{\sum_j (1/n_{.j}) \sum_i n_{ij}^2 - (1/n) \sum_i n_{i.}^2}{n - (1/n) \sum_i n_{i.}^2}$$

$$= \frac{\sum_i \sum_j \frac{[n_{ij} - (n_{i.} n_{.j})/n]^2}{n_{.j}}}{n - (1/n) \sum_i n_{i.}^2}. \tag{2.18}$$

Bei Abhängigkeit der beiden Merkmale A und B sind τ_B und τ_A asymptotisch normalverteilte Größen (vgl. Goodman, Kruskal 1963, 1972). Bezeichnet man den Zähler von τ_B mit v_B und den Nenner mit δ_B, also

$$v_B = \sum_i \sum_j \frac{[n_{ij} - (n_{i.} n_{.j})/n]^2}{n_{i.}} \tag{2.19}$$

und

$$\delta_b = n - \frac{1}{n} \sum_j n_{.j}^2, \tag{2.20}$$

so ergibt sich als Schätzwert für die Varianz von τ_B

$$Var\tau_B = (1/n^2) \sum_i \sum_j n_{ij} \left[\frac{\phi_{ij}^B - \phi^B}{\delta_B^2} \right]^2 \tag{2.21}$$

mit

$$\phi_{ij}^B = 2(\delta_B - v_B) n_{.j} + 2\delta_B n \frac{n_{ij}}{n_{i.}} - \delta_B n \sum_l \left[\frac{n_{il}}{n_{i.}} \right]^2 \tag{2.22}$$

und

$$\phi^B = 2(\delta_B - v_B)(n - \delta_B) + \delta_B(v_B + n - \delta_B). \tag{2.23}$$

Analog dazu ergibt sich als Schätzwert für die Varianz von τ_A (vgl. Hartung 1984, Kap. VII)

$$Var\tau_A = (1/n^2)\sum_i\sum_j n_{ij}\left[\frac{\phi_{ij}^A - \phi^A}{\delta_A^2}\right]^2, \qquad (2.24)$$

mit

$$v_A = \sum_i\sum_j \frac{[n_{ij} - (n_{i.}n_{.j})/n]^2}{n_{.j}} \qquad (2.25)$$

$$\delta_A = n - (1/n)\sum_i n_{i.}^2 \qquad (2.26)$$

$$\phi_{ij}^A = 2(\delta_A - v_A)n_{i.} + 2\delta_A n\frac{n_{ij}}{n_{.j}} - \delta_A n\sum_k\left[\frac{n_{kj}}{n_{.j}}\right]^2 \qquad (2.27)$$

$$\phi^A = 2(\delta_A - v_A)(n - \delta_A) + \delta_A(v_A + n - \delta_A) \qquad (2.28)$$

Mit den oben angegebenen Schätzwerten für die Varianz lassen sich approximativ Konfidenzintervalle für die Tau-Maße berechnen.
Für Hanunoo ergeben sich zum Beispiel

$$v_B = \left[\frac{698^2 + 224^2 + 425^2}{1347} + \frac{221^2 + 143^2 + 152^2}{515} + \frac{324^2 + 96^2 + 485^2}{905}\right]$$

$$-\left[\frac{1273^2 + 463^2 + 1061^2}{2767}\right] = 1097.95 - 1042.70 = 55.25,$$

$$\delta_B = 2767 - 1042.70 = 1724.30$$

und damit

$$\tau_B = \frac{v_B}{\delta_B} = \frac{55.25}{1724.30} = 0.0320.$$

Weiter ergibt sich

$$\phi_B = 2(1724.30-55.25)(2767-1724.30)+1724.30(55.25+2767-1724.30)$$

$$= 5373821.10.$$

Die Berechnung aller ϕ_{ij} ist langwierig, daher zeigen wir explizit nur ein Beispiel für $i = 1, j = 1$.

$$\phi_{11}^B = 2(1724.30 - 55.25)1243 + 2(1724.30)2767(698)/1347$$

$$-1724.30(2767)[(6982 + 2242 + 4252)/13472] = 7205901.80.$$

Die übrigen ϕ_{ij} lauten

$$\phi_{12}^B = 1244320.08$$
$$\phi_{13}^B = 4664406.14$$
$$\phi_{21}^B = 6587469.11$$
$$\phi_{22}^B = 2538515.32$$
$$\phi_{23}^B = 4682924.60$$
$$\phi_{31}^B = 5530003.40$$
$$\phi_{32}^B = 522264.21$$
$$\phi_{33}^B = 6620050.99.$$

Durch Einsetzen in Formel (2.21) erhält man

$$Var\tau_B = \frac{1}{2767^2}\left[\left[\frac{7205901.80 - 5373821.10}{1724.30^2}\right]^2 + \left[\frac{1244320.08 - 5273821.10}{1724.30^2}\right]^2 + \ldots + \left[\frac{6620050.99 - 5373821.10}{1724.30^2}\right]^2\right] = 0.0001617$$

sodaß das Konfidenzintervall für τ_B sich wie folgt ergibt:

$$\bar{\tau}_B = \tau_B + z_{0.975}[Var(\tau_B)]^{1/2} = 0.0320 + 1.96(0.012716) = 0.0570$$

und

$$\underline{\tau}_B = \tau_B + z_{0.025}[Var(\tau_B)]^{1/2} = 0.0320 - 1.96(0.012716) = 0.0071$$

Analog ergibt sich

$$\tau_A = 0.0314$$

und

$$[0.0029; 0.0599]$$

als 95%-Konfidenzintervall für τ_A.

Wann welches Assoziationsmaß verwendet wird, hängt von der jeweiligen Fragestellung ab. Die auf χ^2 basierenden Maße eignen sich beispielsweise zum Vergleich von Kontingenztafeln. Der Pearsonsche Kontingenzkoeffizient P wird wegen seiner Einfachheit häufig gewählt, eignet sich aber lediglich zum Vergleich von Tafeln mit gleicher Zeilen- und Spaltenzahl. Weil das Cramérsche V für alle Tafelgrößen auf das Intervall $[0, 1]$ normiert ist, ist auch V ein bevorzugtes Assoziationsmaß. Die Lambda- und Tau-Maße werden wegen ihrer guten Interpretierbarkeit gewählt. Aufgrund ihrer Asymmetrie sind sie aber nur interessant, wenn eine gerichtete Abhängigkeit der Variablen vermutet wird.

Für die Daten im Anhang wurden die obigen Maße berechnet und es ergaben sich Resultate, wie in der Tabelle 2.3 aufgeführt. Es handelt sich dabei um Abhängigkeiten der Vokale oder Konsonanten in der ersten und zweiten Silbe des Morphems (Wortes) .

Tabelle 2.3.
Unabhängigkeitstests und Kontingenzkoeffizienten
für einige Sprachen

Sprache	χ^2 FG	C	C_{corr}	T	τ_A	τ_B	λ_A	λ_B
Hanunoo V	171.27	.2414	.2957	.1759	.0314	.0320	.0423	.1056
	4	.0962	.1179	.0636	.0029	.0071	.0015	.0710
		.3866	.4735	.2883	.0599	.0570	.0830	.1402
Uraustro-	267.36	.3387	.3911	.2078	.0394	.0351	.0512	.0559
nesisch V	9	.2475	.2858	.1446	.0127	.0062	.0124	.0232
		.4299	.4964	.2710	.0660	.0641	.0900	.0887
Indone-	3185.5	.6418	.7176	.3957	.0837	.0737	.0872	.0186
sisch V	20	.6201	.6933	.3730	.0711	.0470	.0772	.0000
		.6635	.7418	.4185	.0963	.1004	.0972	.0383
Bare'e V	485.35	.3280	.3667	.1736	.0299	.0294	.0891	.0821
	16	.2451	.2740	.1244	.0192	.0209	.0653	.0563
		.4109	.4594	.2228	.0405	.0379	.1128	.1079
Angkola-	1915.7	.4338	.7850	.2407	.0427	.0415	.0776	.0292
Batak V	16	.3856	.4311	.2078	.0331	.0303	.0599	.0207
		.4820	.5389	.2737	.0522	.0526	.0953	.0377
Sundane-	6880.6	.6407	.6920	.3407	.0771	.0848	.0707	.0905
sisch V	36	.6135	.6627	.3162	.0676	.0767	.0566	.0762
		.6679	.7214	.3652	.0866	.0929	.0849	.1047
Hawaiisch	64.90	.2436	.2723	.1256	.0143	.0150	.0170	.0310
V[1)]	16	.1910	.2136	.0968	.0030	.0051	.0000	.0000
		.2961	.3311	.1544	.0256	.0250	.0537	.0802
Tahitisch	64.54	.2599	.2906	.1346	.0161	.0173	.0000	.0228
V[1)]	16	.2040	.2280	.1035	.0000	.0043	.0000	.0000
		.3158	.3531	.1656	.0355	.0304	.0000	.0695
Tuamotu	69.10	.2323	.2598	.1194	.0127	.0133	.0000	.0144
V[1)]	16	.1846	.2064	.0935	.0000	.0032	.0000	.0000
		.2801	.3131	.1454	.0281	.0235	.0000	.0543

Sprache	χ^2 FG	C	C_{corr}	T	τ_A	τ_B	λ_A	λ_B
Hawaiisch	26.26	.1577	.1821	.0922	.0099	.0072	.0000	.0000
KK	9	.1089	.1252	.0626	.0000	.0000	.0000	.0000
		.2071	.2391	.1218	.0415	.0370	.0000	.0000
Tahitisch	22.19	.1559	.1800	.0911	.0104	.0093	.0000	.0000
KK	9	.0966	.1116	.0556	.0000	.0000	.0000	.0000
		.2152	.2484	.1266	.0403	.0402	.0000	.0000
Tuamotu	41.71	.1825	.2107	.1071	.0126	.0119	.0514	.0585
KK	9	.1335	.1542	.0774	.0000	.0000	.0172	.0227
		.2314	.2672	.1369	.0314	.0356	.0856	.0944
Afrikaans	33.66	.2038	.2496	.1472	.0285	.0265	.0562	.0000
KK	4	.0000	.0000	.0000	.0000	.0000	.0000	.0000
		.4662	.5709	.3449	.0828	.1423	.1522	.0000
Arabisch	405.46	.3422	.4191	.2575	.0767	.0729	.1583	.0328
1-2 KK	4	.2501	.3063	.1790	.0494	.0408	.1183	.0000
		.4343	.5319	.3360	.1041	.1051	.1984	.0765
Arabisch	149.21	.2132	.2611	.1543	.0212	.0213	.0054	.0000
1-3 KK	4	.0578	.0708	.0365	.0000	.0000	.0000	.0000
		.3686	.4515	.2721	.0457	.0517	.0490	.0000
Arabisch	72.00	.1572	.1925	.1126	.0131	.0129	.0000	.0000
2-3 KK	4	.0000	.0000	.0000	.0000	.0000	.0000	.0000
		.3757	.4602	.2730	.0460	.0448	.0000	.0000

V -Vokale im Morphem; KK -Konsonantenklassen im Morphem;
[1]) Die χ^2 und die T -Werte findet man auch in Krupa (1971)

Leere Zellen

Bisher sind wir davon ausgegangen, daß alle Zellen einer Tafel besetzt sind. Was geschieht aber, wenn Nullen auftreten? Zunächst ist zu unterscheiden zwischen einer "zufälligen Null" (sampling zero) und einer "strukturellen Null" (structural zero). Zufällige Nullen ergeben sich aufgrund kleiner Zellwahrscheinlichkeiten und der Stichprobenvariabilität. Für sie gilt trotz $n_{ij} = 0$ immer $p_{ij} > 0$. Zufällige Nullen können im Prinzip dadurch verhindert werden, daß der Stichprobenumfang hinreichend groß gewählt wird. Bei einer strukturellen Null ist a priori bekannt, daß die entsprechende Zelle nicht besetzt ist, es gilt also $p_{ij} = 0$. Strukturelle Nullen liegen z.B. dann vor, wenn bestimmte Phonemkombinationen durch grammatische Regeln ausgeschlossen werden. Eine Schwierigkeit besteht oft darin, strukturelle Nullen als solche zu erkennen. Kontingenztafeln, die strukturelle Nullen enthalten, bezeichnet man als unvollständige Tafeln. Unvollständige Kontingenztafeln bedürfen einer besonderen Behandlung. Eine wichtige Rolle spielt dabei die (statistische) Hypothese der "Quasi-Unabhängigkeit".

Sei $S := \{(i,j) : p_{ij} > 0\}$ die Menge der Zellen einer Tafel, die keine strukturellen Nullen sind. Die Hypothese der Quasi- Unabhängigkeit lautet dann

$$H_0^{QU} = p_{ij} = \delta_{ij} \cdot p_{i.} p_{.j}$$

für alle (i,j)mit

$$\delta_{ij} = \begin{cases} 1 & \text{falls } (ij) \in s \\ 0 & \text{sonst} \end{cases}$$

Vor der Untersuchung empfiehlt es sich zu prüfen, ob die (unvollständige) Kontingenztafel separabel oder nicht-separabel (verbunden) ist. In einer zweidimensionalen Kontingenztafel nennt man zwei Zellen miteinander verbunden, wenn sie keine strukturellen Nullen enthalten und sich entweder in derselben Zeile oder in derselben Spalte befinden. Eine Menge von Zellen ohne strukturelle Nullen heißt verbunden, wenn jedes Paar von Zellen der Menge so durch eine Kette verbunden werden kann, daß jeweils zwei benachbarte Glieder der Kette miteinander verbunden sind. Eine unvollständige Tafel nennt man verbunden, wenn alle ihre Zellen, die keine strukturellen Nullen enthalten, eine verbundene Menge bilden.

Eine nicht verbundene unvollständige Kontingenztafel heißt separabel. Eine separable Tafel läßt sich nach geeigneter Umstellung von Zeilen und Spalten durch zwei oder mehrere Teiltafeln so darstellen, daß eine Teiltafel mit einer anderen keine besetzten Zeilen oder Spalten gemeinsam hat (s. Beispiel).

Beispiel:
Umordnung einer separablen Kontingenztafel

a) vor der Umordnung

	1	2	3	4	5
1	n_{11}	–	–	n_{14}	–
2	–	n_{22}	n_{23}	–	n_{25}
3	–	n_{32}	–	–	n_{35}
4	n_{41}	–	–	n_{44}	–

b) nach der Umordnung

	1	4	2	3	5
1	n_{11}	n_{14}	–	–	–
4	n_{41}	n_{44}	–	–	–
2	–	–	n_{22}	n_{23}	n_{25}
3	–	–	n_{32}	–	n_{35}

Wie man sieht, läßt sich eine separable Kontingenztafel so darstellen, daß sie aus sich nicht überlappenden Blöcken auf der Diagonalen besteht. Man spricht daher auch von blockdiagonalen Tafeln. Dabei kann jeder Block durchaus noch strukturelle Nullen enthalten.

Im Falle von separablen Tafeln läßt sich die Hypothese der Unabhängigkeit für jede Teiltafel mit Hilfe der jeweiligen χ^2- Statistik überprüfen und somit die Hypothese der Quasi- Unabhängigkeit der Gesamttafel mit der Summe der einzelnen χ^2- Statistiken der Teiltafeln (vgl. Goodman 1968).

Für unvollständige nicht-separable Kontingenztafeln lassen sich die ML-Schätzer in der Regel nicht in geschlossener Form darstellen.

Für diesen Fall geben Bishop und Fienberg (1969) einen Algorithmus, der eindeutige ML-Schätzer liefert (zur Existenz und Eindeutigkeit vgl. auch Fienberg 1970).

Der Algorithmus folgt dem Prinzip des "Iterative proportional fitting", bei dem unter dem Modell der Quasi-Unabhängigkeit die einzelnen Zellen sukzessive den Rändern angepaßt werden.

Als Startwert wählen wir

$$\hat{m}_{ij}^{(0)} = \delta_{ij}.$$

Im ersten Iterationszyklus ist der erste Schritt:

$$1.\ \text{Zyklus} = \begin{cases} \hat{m}_{ij}^{(1)} = \frac{\hat{m}_{ij}^{(0)} n_{i.}}{\sum_k \hat{m}_{ik}^{(1)}} \\ \text{und} \\ \hat{m}_{ij}^{(2)} = \frac{\hat{m}_{ij}^{(1)} n_{.j}}{\sum_k \hat{m}_{kj}^{(1)}} \end{cases}$$

Im v-ten Iterationszyklus ist

$$v\text{–ter Zyklus} = \begin{cases} \hat{m}_{ij}^{(2v-1)} = \frac{\hat{m}_{ij}^{(2v-2)} n_{i.}}{\sum_k \hat{m}_{kj}^{(2v-2)}} \\ \text{und} \\ \hat{m}_{ij}^{(2v)} = \frac{\hat{m}_{ij}^{(2v-1)} n_{.j}}{\sum_k \hat{m}_{kj}^{(2v-1)}} \end{cases}$$

Die Iteration wird solange fortgesetzt, bis für ein v gilt:

$|\hat{m}_{ij}^{(2v)} - \hat{m}_{ij}^{(2v-2)}| < \varepsilon$ für alle (i,j)

mit einer vorgegebenen Genauigkeit ε.

Für den Test auf Quasi-Unabhängigkeit sind die so erhaltenen ML-Schätzer für m_{ij} in die χ^2-Teststatistik einzusetzen, d.h. man benutzt Formel (2.1).

Für die Zahl der Freiheitsgrade g gilt:

$$g = (r-1)(s-1) - e,$$

wobei mit e die Anzahl der strukturellen Nullen in der Tafel bezeichnet wird. Zu beachten ist natürlich, daß die Zahl der Freiheitsgrade nicht verschwindet.

Die Entscheidung darüber, ob im Falle einer Null-Zelle eine strukturelle Lücke vorliegt, ist vom linguistischen Standpunkt sehr schwierig. Es gibt immer ausgesprochen kategorische Distributionsregeln für Phoneme, z.B. "keine mediae am Wortende im Deutschen", es gibt aber kein derartiges Gesetz, so daß man in der Sprache L nur von einer Tendenz sprechen darf. Nichts kann die Sprache hindern, zu beliebiger Zeit

ein Fremdwort zu übernehmen, dessen korrekte Aussprache eine aktuelle Distributionsregel bricht. Eo ipso stellen leere Zellen irgendwelche dissoziativen Tendenzen, nicht aber Gesetze dar. Man kann den obigen Test aber trotzdem verwenden, indem man unter der Bedingung, daß die leeren Zellen strukturelle Lücken darstellen, testet.

Versuchen wir auf diese Weise die Unabhängigkeit der Vokale im Sundanesischen (s. Tab. A4) zu testen, wo es 5 Null-Zellen gibt. Durch Iteration erhalten wir die unter der Hypothese der Quasi- Unabhängigkeit erwarteten Werte m_{ij} wie in Tabelle 2.4 angegeben.

Tabelle 2.4.
Erwartete Zellhäufigkeiten unter
Quasi-Unabhängigkeit für Sundanesisch

	a	e	ĕ	ë	i	o	u
a	1015.01	376.45	316.34	180.60	469.36	463.97	556.27
e	272.05	100.90	84.79	-	125.80	124.36	149.10
ĕ	503.90	186.89	157.05	89.66	233.01	230.34	276.16
ë	172.27	-	-	30.65	79.66	-	94.41
i	341.64	126.71	106.48	60.79	157.98	156.17	187.23
o	326.34	121.03	101.71	-	150.90	149.17	178.85
u	428.78	159.03	133.64	76.29	198.28	196.00	234.99

Der resultierende χ^2-Wert ergibt $\chi^2 = 5281.80$ mit (7-1)(7-1) - 5 = 31 Freiheitsgraden, was zu einer Ablehnung der Hypothese der Quasi-Unabhängigkeit führt.

Mit dem allgemeinen Unabhängigkeitstest kann man induktiv zu Generalisierungen gelangen, die zur Aufstellung von Hypothesen anregen können. Als erstes könnte man von dem zweidimensionalen Fall, den wir hier untersucht haben, zu mehrdimensionalen Fällen übergehen. Würde man z.B. die Abhängigkeiten der Vokale in dreisilbigen Wörtern prüfen, so hätte man eine dreidimensionale Kontingenztafel und es würde automatisch die Frage entstehen, wie sich die Abhängigkeit mit zunehmender Zahl der Dimensionen verändert. Außerdem läßt sich fragen, welche Abhängigkeiten überhaußt bestehen. Denn mit der Zahl der Dimensionen erhöht sich in starkem Maße die Zahl der möglichen Abhängigkeitshypothesen. So läßt sich zum Beispiel in einer dreidimensionalen Kontingenztafel untersuchen, ob totale Unabhängigkeit

vorliegt, d.h. alle drei Merkmale voneinander unabhängig sind, oder ob ein Merkmal unabhängig von den beiden anderen ist, ob zwei Merkmale bedingt unabhängig voneinander sind bei gegebenem dritten Merkmal usw.

Für das Formulieren und das Testen von Hypothesen in mehrdimensionalen Kontingenztafeln eignet sich besonders gut die Darstellung durch Log-lineare Modelle, auf die hier allerdings nicht näher eingegangen werden soll.

Wir wollen die Untersuchung vereinfachen und sequentiell vorgehen, indem wir schrittweise nur die Abhängigkeit der unmittelbar benachbarten Phoneme vom Wort-, Morphem-, und Silbenanfang bis zum Wort-, Morphem-, und Silbenende testen. Für semitische Sprachen wäre es sinnvoll, nur die Abhängigkeit zwischen dem ersten und zweiten Konsonanten, und dem zweiten und dritten Konsonanten der Wurzel zu testen, aber das Resultat würde zu keiner Hypothese führen, weil man dann nur zwei Werte und keine "anständige Kurve" bekommt. In anderen Sprachen kann man aber diese Untersuchung problemlos durchführen und sicherlich zu wertvollen Hypothesen gelangen.

Um das Verfahren zu illustrieren, haben wir 4602 indonesische Morpheme des Typs CVCVC (C = Konsonant, V = Vokal) ermittelt und die Interaktionen der benachbarten Phoneme ausgezählt. Unter Verwendung von (2.3) erhielten wir die Resultate wie in der Tabelle 2.5 angegeben.

Tabelle 2.5.
Test für Abhängigkeit zwischen benachbarten
Phonemen des Typs CVCVC in indonesischen Morphemen

Abhängigkeit zwischen dem	χ^2 FG	C	C_{corr}	T	V	τ_a	τ_b	λ_a	λ_b
1. und 2. Phonem	342.69 100	.263	.289	.086	.122	.004	.019	.003	.041
2. und 3. Phonem	376.69 100	.275	.301	.091	.128	.022	.003	.024	.004
3. und 4. Phonem	256.95 100	.230	.252	.075	.106	.003	.015	.007	.000
4. und 5. Phonem	581.57 80	.335	.367	.118	.159	.037	.009	.002	.006

Wie man sieht, wird die von der ersten zu der letzten Position steigende Abhängigkeit an der dritten Position unterbrochen und sinkt etwas, obwohl sie hoch signifikant bleibt. Die Maße T_b, L_a und L_b verhalten sich anders. Derartige Verläufe sind sicherlich nicht zufällig und verlangen eine Untersuchung auf breiter Basis.

Die gleiche Fragestellung, bezogen auf die Silben, führt sofort zu einer Generalisierung, die möglicherweise zu der Formulierung eines Geseztes führen kann. Stellt man eine Kontingenztafel auf, in der als Klassen die Zahl der Konsonanten vor und hinter dem Silbenkern (V) auftreten, so bekommt man z. B. für das Estnische nach Tuldava (1978) die in Tabelle 2.6 aufgeführten Daten.

Tabelle 2.6.
Silbentypen im Estnischen
nach Tuldava

	hinterer Kern			
vor Kern	.	.C	.CC	.CCC
.	13	96	89	1
C.	173	1345	1231	101
CC.	9	138	70	0
CCC.	1	6	5	15

Die Zahlen in der Tabelle sind folgendermaßen zu lesen: Es gibt 1231 Silben mit der Struktur CVCC, es gibt 1345 Silben mit der Struktur CVC usw.

Der Test für Unabhängigkeit ergibt $\chi^2 = 244.67$ mit 9 Freiheitsgraden, was sehr hoch signifikant ist. Betrachtet man die Tabelle aber näher, so sieht man, daß die Zahlen in den Zeilen oder Spalten einen recht regulären Verlauf aufweisen und womöglich die Existenz einer zweidimensionalen Wahrscheinlichkeitsverteilung signalisieren. Wir werden sie an einer anderen Stelle ableiten. Hier reicht uns die Erkenntnis, daß die kanonischen Silbentypen in einer Sprache mit starken Einschränkungen gebaut werden.

Der Unabhängigkeitstest ist natürlich auch für die Prüfung der Rigorosität grammatischer Regeln nützlich. Stellt man eine Kontingenztafel auf, die die Häufigkeiten der Übergänge zwischen Wortarten enthält, so kann man ein Charakteristikum der Texte, der Genres oder

auch Sprachen gewinnen.

3. Beurteilung einzelner Zellen

Jede einzelne Zelle der Kontingenzmatrix liefert einen Beitrag zu dem Gesamtchiquadrat des vorigen Kapitels, aber wir wissen nicht, ob die zwei betreffenden Laute sich signifikant assoziieren oder dissoziieren, welche "Qualität" ihre Interaktion hat usw. (vgl. Altmann 1973, Altmann, Lehfeldt 1980). Diese Probleme werden hier behandelt.

Den Test für einzelne Zellen führt man nach der Formel

$$z = n\frac{n_{ij} - (n_{i.}n_{.j})/n}{[(n_{i.}n_{.j})(n - n_{i.})(n - n_{.j})/n]^{1/2}} \tag{3.1}$$

durch (vgl. Řehák, Řeháková 1980). Dieser Test zeigt die Stärke einer Einzelinteraktion, d.h. deckt die isolierten Abhängigkeiten der Phoneme auf.

Illustrieren wir die obige Formel an einem Beispiel aus Hanunoo. Die Berechnung für die Kombination [a] – [i] ergibt sich dabei folgendermaßen:

$$z = 2767\frac{224 - [1347(463)/2767]}{[1347(463)(2767 - 1347)(2767 - 463)/2767]^{1/2}} = -0.14.$$

Das Resultat ist nicht signifikant. Wir können schließen, daß [a] und [i], in dieser Reihenfolge in Hanunoo keine irgendwie (positiv oder negativ) ausgezeichnete Interaktion bilden, da $P(X \geq \mid 0.14 \mid) = 0.8886$.

Die anderen Tests für Hanunoo findet man in der Tabelle 3.1. In den einzelnen Zellen sind die resultierenden z eingetragen.

Tabelle 3.1.
z-Tests für einzelne
Zellen in Hanunoo

	a	i	u
a	7.10	-0.14	-7.16
i	-1.02	7.44	-4.67
u	-6.72	-6.02	11.50

Um nun die einzelnen Interaktionen qualitativ zu beurteilen, benutzen wir die folgenden Kriterien:

Das Signifikanzkriterium mit Hilfe des Tests (3.1) und das Existenzkriterium aufgrund der Zellenhäufigkeit, und wählen folgende Bezeichnungen:

(A) Wenn $n_{ij} > 0$
  (1) $n_{ij} < E(n_{ij})$ signifikant, dann ist die Interaktion marginal (M)
  (2) $n_{ij} = E(n_{ij})$,d.h. nicht signifikant unterschiedlich, dann ist die Interaktion aktuell (A)
  (3) $n_{ij} > E(n_{ij})$ signifikant, dann ist die Interaktion präferiert (P).

(B) Wenn $n_{ij} = 0$
  (1) $n_{ij} = E(n_{ij})$,d.h. nicht signifikant unterschiedlich, dann ist die Interaktion virtuell (V)
  (2) $n_{ij} < E(n_{ij})$ signifikant, dann ist die Interaktion unzulässig (I).

Übersichlich kann man diese Klassifikation wie in der Tabelle 3.2 darstellen.

Tabelle 3.2.
Interaktionsarten

	$n_{ij} < E(n_{ij})$	$n_{ij} = E(n_{ij})$	$n_{ij} > E(n_{ij})$
$n_{ij} > 0$	M marginal	A aktuell	P präferiert
$n_{ij} = 0$	I unzulässig	V virtuell	–

Die Interaktionstabelle des Hanunoo kann man dann auch qualitativ darstellen (s. Tabelle 3.3).

Tabelle 3.3.
Qualitative Beurteilung der Vokalinteraktion in Hanunoo.

	a	i	u
a	P	A	M
i	A	P	M
u	M	M	P

Die Auswertung für die Vokalkombinationen in Morphemen in einigen Indonesischen Sprachen sind in den Tabellen 3.4 bis 3.8 aufgeführt.

Tabelle 3.4.
Qualitative Beurteilung der Vokalinteraktion in Angkola

	a	e	i	o	u
a	P	M	P	M	A
e	M	P	M	A	M
i	A	M	P	A	M
o	M	M	M	P	M
u	A	M	M	M	P

Tabelle 3.5.
Qualitative Beurteilung der Vokalinteraktion im Sundanesischen

	a	e	ĕ	ë	i	o	u
a	P	M	M	M	P	M	A
e	M	P	M	I	M	P	M
ĕ	M	P	P	M	A	A	A
ë	M	I	I	P	A	I	M
i	A	M	P	P	P	M	A
o	M	P	M	I	M	P	M
u	A	M	A	M	A	M	P

Tabelle 3.6.
Qualitative Beurteilung der Vokalinteraktion in Bare'e

	a	e	i	o	u
a	P	M	A	M	A
e	A	P	M	A	M
i	A	M	P	A	A
o	M	A	M	P	M
u	A	M	A	M	P

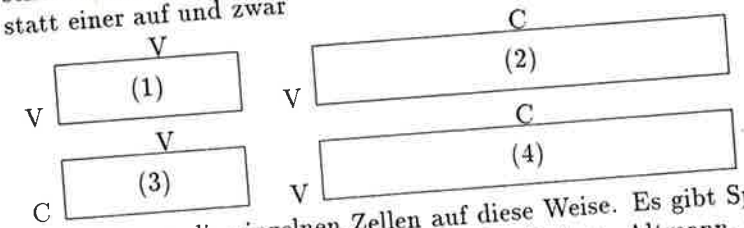
Tabelle 3.7.
Qualitative Beurteilung
der Vokalinteraktion
im Indonesischen

	a	e	i	o	u
a	P	I	P	I	P
e	M	P	I	P	I
i	A	I	P	I	A
o	M	P	M	P	I
u	M	I	A	I	P
ə	P	M	A	M	A

Tabelle 3.8.
Qualitative Beurteilung
der Vokalinteraktion
im Uraustronesischen

	a	e	i	u
a	P	M	A	M
e	M	P	M	M
i	A	A	P	M
u	M	M	M	P

Bei Untersuchungen dieser Art ist je nach Bedarf zu überlegen, ob man die Tests für Vokale und Konsonanten separat durchführt, weil die Interaktionen der Vokale mit Konsonanten so stark sind, daß sie auch starke Assoziationen zweier Konsonanten als nicht signifikant erscheinen lassen können. Das bedeutet, man stellt vier Kontingenztafeln statt einer auf und zwar



und testet die einzelnen Zellen auf diese Weise. Es gibt Sprachen, in denen Tafel (1) (z.B. Amerikanisch Englisch, s. Altmann, Lehfeldt 1980: 304) oder Tafel (4) (z.B. Hawaiisch) oder beide leer sind. In

solchen Fällen nimmt man definitorisch für alle VV- bzw. CC- Interaktionen die stärkste Dissoziation an und bezeichnet sie als "unzulässig".

Hier bietet sich die folgende Verallgemeinerung an. Das Ökonomieprinzip, das in der Sprache durchgängig gilt, führt dazu, daß sich in der Sprache Interaktionen mit ausgeprägten Assoziationen und Dissoziationen bilden, die dann den spezifischen Einheitenaufbau hervorrufen. Je mehr Phoneme eine Sprache hat, desto mehr leere Zellen kann sie zulassen, bzw. desto stärkere Präferenzen kann sie bilden. Mathematisch ausgedrückt kann man sagen, daß die Zahl der Lücken (oder Stärke der Präferenzen) proportional der Zahl der Phoneme ist. Diese Proportionalität sollte eine Invariante darstellen, d.h. einen Mechanismus bilden, der auch bei Veränderungen des Phoneminventars (innerhalb einer Sprache oder zwischen Sprachen) für Konstanz sorgt. Auf menschliche Sprache als ganze bezogen heißt es dann eher eine Proportionalität bezogen auf die Veränderung der Phonemzahl. Daher postulieren wir, daß die relative Veränderungsrate der Lückenzahl der relativen Veränderungsrate der Phonemzahl proportional ist, d.h.

$$\frac{dL}{L} = k\frac{dP}{P} \tag{3.2}$$

wobei L die Lückenzahl, P die Phonemzahl und k die Proportionalitätskonstante darstellt. Die Lösung dieser Differentialgleichung führt zum "Menzerathschen Gesetz",

$$L = aP^k, \tag{3.3}$$

das in diesem Bereich vorläufig nur schwach testbar ist, da dazu Daten aus vielen Sprachen nötig sind. Möglicherweise ergeben sich unterschiedliche Parameter für unterschiedliche Spielarten der Interaktionen z.B. für Phoneme direkt nebeneinander oder in Abständen i voneinander. Vorläufig müssen wir also auf die empirische Überprüfung dieser Hypothese verzichten.

4.Tests für die Diagonale

Betrachtet man die Tabellen 3.3 bis 3.9 , so stellt man fest, daß die Hauptdiagonale starke Präferenzen aufweist. Man kann daraus folgern, daß in indonesischen Sprachen eine starke Tendenz besteht, gleiche Vokale in die zweisilbigen Grundwörter zu plazieren, d.h. es gibt

hier eine tendenzielle Vokalharmonie. Diese Harmonie kann aber auch dann noch signifikant sein, wenn nicht alle Diagonalzellen präferierte Interaktionen aufweisen, sondern wenn sie als ganze signifikant stärker als die Erwartung ausgeprägt ist.

Dies führt zu dem Testproblem

$$H_0^{Dg} : \sum_i p_{ii} = \sum_i p_{i.}p_{.i}$$

gegen

$$H_1^{Dg} : \sum p_{ii} \neq \sum p_{i.}p_{.j}.$$

Um dies zu testen, benutzt man nach Altmann (1987) entweder den Likelihood-Ratio Test für den Trend in der Diagonale und berechnet die Größe

$$\chi^2 = \frac{n(n\sum n_{ii} - \sum n_{i.}n_{.i})^2}{\sum n_{i.}n_{.i}(n^2 - \sum n_{i.}n_{.i})} \tag{4.1}$$

mit 1 Freiheitsgrad, oder einen gleichwertigen Test nach

$$z = \frac{\sum_i n_{ii} - \sum_i \frac{n_{i.}n_{.i}}{n}}{\left[\frac{1}{n^2(n-1)}[\sum_i n_{i.}n_{.i}(n-n_{i.})(n-n_{.i}) + 2\sum_{i<i'} n_{i.}n_{.i}n_{i'.}n_{.i'}]\right]^{1/2}} \tag{4.2}$$

wobei $[\chi^2]^{1/2} = z$.

Für unser Beispiel in Hanunoo ergeben sich die Resutate wie folgt.

$$\sum n_{ii} = 698 + 143 + 485 = 1326$$

$$\sum n_{i.}n_{.i} = 1347(1243) + 515(463) + 905(1061) = 2872971$$

$$\sum n_{i.}n_{.i}/n = 2871971/2767 = 1038.2982$$

$$\frac{1}{n^2(n-1)}[\sum n_{i.}n_{.i}(n-n_{i.})(n-n_{.i}) + 2\sum_{i<i'} n_{i.}n_{.i}n_{i'.}n_{.i'}] =$$

$$\frac{1}{2767^2(2766)}[1347(1243)(2767-1347)(2767-1243)$$

$$+515(463)2252(2304) + 905(1061)1852(1706)$$

$$+2[1347(1243)515(463) + 1347(1243)905(1061)$$

$$+515(463)905(1961)]] = 583.9321.$$

Setzt man diese Zahlen in (4.1) ein, so erhält man

$$\chi^2 = \frac{2767[2767(1326) - 2872971]^2}{2872971(2767^2 - 2872971)} = 127.60$$

und mit (4.2) bekommt man

$$z = \frac{1326 - 1038.2982}{[583.9321]^{1/2}} = 11.91$$

Wie man sieht, ist $[127.60]^{1/2} = 11.30$, was mit dem z-Wert fast identisch ist. Die Übereinstimmung zwischen (4.1) und (4.2) ist nicht vollständig, weil (4.1) durch eine unvollständige Taylorentwicklung von Logarithmen entsteht.

Das Resultat zeigt, daß in Hanunoo eine starke Gesamttendenz nach identischer Vokalharmonie in zweisilbigen Grundwörtern besteht.

Die Stärke der Tendenz auf der Diagonale läßt sich mit dem als Cohens Kappa bekanntem Index K beschreiben (vgl. Cohen 1960). Es ist

$$K = \frac{n\sum_i n_{ii} - \sum_i n_{i.}n_{.i}}{n^2 - \sum_i n_{i.}n_{.i}} \tag{4.3}$$

Für Hanunoo bekommt man K = 0.1664. Dieser Wert läßt sich dahingehend interpretieren, daß die Tendenz in ein zweisilbiges Grundwort die gleichen Vokale zu setzen um 16.64% höher ist als zufällig zu

erwarten wäre. Als Schätzer für die Varianz von K ergibt sich (vgl. Fleiss, Cohen, Everitt 1969)

$$Var(K) = \frac{n^2 \sum_i n_{i.} n_{.i} + (\sum_i n_{i.} n_{.i})^2 - n \sum_i n_{i.} n_{.i} (n_{i.} + n_{.i})}{n(n^2 - \sum_i n_{i.} n_{.i})^2} \quad (4.4)$$

Für Hanunoo erhält man Var(K) = 0.000196, woraus sich ein 95%-Konfidenzintervall für K von (0.1390,0.1938) berechnet.
Dividiert man (4.3) mit $[Var(K)]^{1/2}$, so erhält man genau das Resultat in (4.2). Für die andere Sprachen ergeben sich die Resultate in Tabelle 4.1.

Tabelle 4.1.
Tests für die Diagonale einiger Sprachen
(Identitätsvokalharmonie)

Sprache	z	K	$(\underline{K},\bar{K})$
Hanunoo	11.90	0.1664	(0.1390, 0.1938)
Uraustronesisch	12.93	0.1720	(0.1460, 0.1981)
Indonesisch	18.36	0.1255	(0.1121, 0.1389)
Bare'e	18.99	0.1523	(0.1366, 0.1680)
Angkola	27.25	0.1591	(0.1477, 0.1705)
Sundanesisch	47.63	0.2143	(0.2055, 0.2232)
Hawaiisch	5.43	0.0851	(0.0544, 0.1157)
Tahitisch	4.50	0.0762	(0.0430, 0.1094)
Tuamotu	4.83	0.0701	(0.0417, 0.0986)

Krupa's Konsonantendissoziation

Während man bei den Vokalen oft eine deterministische und nicht selten eine stochastische "Harmonie" (= Assoziation) findet, scheint es bei den Konsonanten genau umgekehrt zu gehen. Wie Krupa (1967) gezeigt hat, weisen die Sprachen die Tendenz auf, in einem Stamm Konsonanten, die zu der gleichen Klasse gehören und die mit einem Vokal voneinander getrennt sind, zu meiden. Unter Anwendung des Tests für die Diagonale erhielten wir einige Resultate wie in Tabelle 4.2 dargestellt.

Tabelle 4.2.
Konsonantentrends in einigen Sprachen
(Daten nach Krupa)

Sprache	z	K	$(\underline{K},\bar{K})$
Hawaiisch	-3.95	-0.1031	(-0.1543, -0.0519)
Tahitisch	-4.35	-0.1222	(-0.1773, -0.0671)
Tuamotu	-5.47	-0.1397	(-0.1775, -0.0839)
Afrikaans	-5.75	-0.1434	(-0.1923, -0.0945)
Arabisch 1-2	-20.00	-0.2647	(-0.2906, -0.2388)
Arabisch 1-3	-10.61	-0.1381	(-0.1636, -0.1126)
Arabisch 2-3	-8.21	-0.1123	(-0.1392, -0.0855)

In allen Sprachen wurde die Artikulationsstelle als Klassifikationsgrundlage genommen (vordere, mittlere, hintere Konsonanten). Im Arabischen bezeichnet 1-2 den ersten und zweiten Konsonanten der Wurzel usw.

Aus den beiden Tendenzen kann man schließen, daß das, was die Sprache auf einer Seite fördert, unterdrückt sie in einem anderen Bereich, da sie mit ihren Mitteln sparsam umgehen will.

Eine oberflächliche Betrachtung dieses Phänomens der Phoneminteraktionen führt lediglich zu der Entdeckung irgendwelcher Tendenzen, zu induktiven Generalisierungen. Geht man aber von der Annahme der Selbstregulation in der Sprache aus, so muß man dieses Phänomen im Zusammenhang mit anderen Phänomenen der Sprache sehen. Im Augenblick ist noch nicht bekannt, was auf die Belastung oder Entlastung der Diagonale wirkt, warum Vokale zur Assoziation und Konsonanten zur Dissoziation tendieren, ob nur die Phonemzahl einen Ordnungsparameter darstellt, oder auch die mit ihr verbundene Wortlänge (Morphlänge), Vokabulargröße oder andere Größen (vgl. Köhler 1986). Um auf diese Fragen eine Antwort zu bekommen, muß man abwarten, bis Resultate aus vielen Sprachen vorliegen.

5. Spezielle Hypothesen in quadratischen Kontingenztafeln

In folgenden gehen wir davon aus, daß wir eine quadratische r × r-Tafel vorliegen haben. Außerdem nehmen wir an, daß die Katego-

rien der Spaltenvariablen mit den Kategorien der Zeilenvariablen übereinstimmen. Für diesen Fall lassen sich einige spezielle Hypothesen überprüfen.

DIE SYMMETRIEHYPOTHESE

$H_0^{sym}: p_{ij} = p_{ji}$, für alle $i \neq j$,

läßt sich mit der Teststatistik

$$\chi^2_{sym} = \sum\sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \qquad (5.1)$$

(vgl. Bowker 1948) testen. χ^2_{sym} ist χ^2-verteilt mit $r(r-1)/2$ Freiheitsgraden.

Für die Vokalkombinationen des Hanunoo erhalten wir

$$\chi^2_{sym} = \frac{(221-224)^2}{221+224} + \frac{(324-425)^2}{324+425}$$

$$+\frac{(96-151)^2}{96+151} = 25.89,$$

womit die Symmetriehypothese bei 3 Freiheitsgraden abzulehnen ist. Für einige andere Sprachen sind die Ergebnisse des Symmetrietests in Tabelle 5.1 dargestellt. Wie man sieht, ist Symmetrie als Konstruktionsprinzip unterschiedlich beliebt. Für die Vokalkombinationen wird die Symmetriehypothese in den hier untersuchten polynesischen Sprachen angenommen und in den indonesischen abgelehnt. Für die Konsonantenklassen besteht eine Symmetrie im Arabischen zwischen der zweiten und der dritten Position.

Tabelle 5.1.
Bowkers Test auf Symmetrie
für einige Sprachen

Sprache	χ^2_{sym}	FG
Hanunoo, V	25.89	3
Uraustronesisch, V	92.03	6
Bare'e, V	73.56	10
Angkola, V	163.97	10
Sundanesisch, V	572.23	21
Hawaiisch, V	15.43*)	10
Tahitisch, V	13.33*)	10
Tuamotu, V	16.58*)	10
Hawaiisch, C	21.26	3
Tahitisch, C	19.05	3
Tuamotu, C	24.04	3
Afrikaans, C	62.61	3
Arabisch 1-2, C	57.79	3
Arabisch 1-3, C	72.91	3
Arabisch 2-3, C	4.26*)	3

*) Führt zur Annahme der Symmetriehypothese ($\alpha = 0.05$)

Eine allgemeinere Hypothese in quadratischen Kontingenztafeln ist die der

MARGINAL-HOMOGENITÄT

$H_0^{MH}: p_{i.} = p_{.i}$, für i = 1,. . .,r .

Diese Hypothese läßt sich mit der folgenden Teststatistik Q überprüfen (vgl. Stuart 1955):

$$Q = d'V^{-1}d = \sum_{i=1}^{r-1}\sum_{j=1}^{r-1} V^{ij} d_i d_j, \qquad (5.2)$$

wobei mit V^{ij} die Elemente der inversen Matrix V^{-1} bezeichnet werden. Die Elemente V_{ij} der Matrix V (i,j = 1,. . .,r-1) sind definiert als

$$V_{ij} := \begin{cases} n_{i.} + n_{.i} - 2n_{ii} & \text{für i = j} \\ -(n_{ij} + n_{ji}), & \text{für i} \neq j \end{cases}$$

und die Elemente des Vektors d als

$d_i := (n_{i.} - n_{.i})$, für i = 1, ..., r − 1

Q ist χ^2-verteilt mit (r-1) Freiheitsgraden. Die r-te Zeile und Spalte bleiben beim Stuart-Test unberücksichtigt.

Gilt H_0^{sym}, so gilt auch H_0^{MH}, d.h. Symmetrie ist ein Spezialfall der Marginalhomogenität (vgl. Caussinus 1965).

Als Beispiel berechnen wir die Teststatistik Q für Hanunoo. Die Elemente der Matrix V ergeben sich folgendermaßen:

$$V_{11} = n_{1.} + n_{.1} - 2_{n11} = 1347 + 1243 - 2(698) = 1194$$

$$V_{12} = -(n_{12} + n_{21}) = -(224 + 221) = -445$$

$$V_{21} = V_{21} = -445$$

$$V_{22} = n_{2.} + n_{.2} - 2_{n22} = 515 + 463 - 2(143) = 692$$

$$d.h.\ V = \begin{bmatrix} 1194 & -445 \\ -445 & 692 \end{bmatrix}$$

Die Inversion ergibt dann

$$V^{-1} = \begin{bmatrix} 0.001102 & 0.000708 \\ 0.000708 & 0.001901 \end{bmatrix}$$

Die Elemente des Vektors d berechnen sich wie folgt:

$$d_1 = n_{1.} - n_{.1} = 1347 - 1243 = 104$$

$$d_2 = n_{2.} - n_{.2} = 515 - 463 = 52.$$

Damit ergibt sich für die Teststatistik Q:

$$Q = V^{11}d_1d_1 + V^{12}d_1d_2 + V^{21}d_2d_1 + V^{22}d_2d_2$$

$$= 0.001102(104)^2 + 2(0.000708)104(52) + 0.001901(52)^2$$

$$= 24.71.$$

Dieser Wert führt zu einer Ablehnung der Hypothese.
In Tabelle 5.2 sind die Testergebnisse für weitere Sprachen dargestellt.

Tabelle 5.2.

Tests auf Marginalhomogenität in einigen Sprachen

Sprache	Q	FG
Hanunoo, V	24.71	2
Uraustronesisch, V	87.07	3
Bare'e, V	44.01	4
Angkola, V	87.08	4
Sundanesisch, V	424.24	6
Hawaiisch, V	9.12*)	4
Tahitisch, V	4.08*)	4
Tuhamotu, V	4.60*)	4
Hawaiisch, C	19.22	2
Tahitisch, C	18.17	2
Tuhamotu, C	23.95	2
Afrikaans, C	62.43	2
Arabisch 1-2, C	57.78	2
Arabisch 1-3, C	72.60	2
Arabisch 2-3, C	2.82*)	2

*) Hypothese der Marginalhomogenität wird angenommen (α = 0.05)

Da Symmetrie Marginalhomogenität impliziert, wird die Hypothese der Marginalhomogenität selbstverständlich in den gleichen Fällen angenommen, in denen auch Symmetie gilt. Von den hier untersuchten Sprachen besteht für keine Marginalhomogenität ohne Symmetrie.

Eine weitere mögliche Hypothese in quadratischen Kontingenztafeln ist die der

PROPORTIONALITÄT DER DIAGONALZELLEN:

$H_0^{Prop}: \frac{p_{ii}}{p_{i.}p_{.i}} = \frac{p_{jj}}{p_{j.}p_{.j}}$, für i,j = 1,...,r.

H_0^{Prop} kann mit der von Durbin entwickelten Teststatistik V getestet werden (vgl. Mukherjee, Hall 1967). Es gilt

$$V = \sum_i n_{i.}n_{.i} \frac{[(n_{ii}/n_{i.}n_{.i}) - ((\sum n_{ii}/\sum n_{i.}n_{.i})]^2}{(\sum n_{ii}/\sum n_{i.}n_{.i})} \qquad (5.3)$$

V ist χ^2-verteilt mit r-1 Freiheitsgraden.

Für Hanunoo erhalten wir den Nenner von (5.3) als

$$(698 + 143 + 485)/[1347(1243) + 515(463) + 905(1061)] = 0.000462,$$

sodaß sich V folgendermaßen ergibt:

$$V = \frac{1}{0.000462}\left[1347(1243)\left[\frac{698}{1347(1243)} - 0.000462\right]^2 + 515(463)\left[\frac{143}{515(463)} - 0.000462\right]^2 + 905(1061)\left[\frac{485}{905(1061)} - 0.000462\right]^2\right] = 21.05$$

Dieser Wert führt zu einer Ablehnung der Hypothese. Testergebnisse für weitere Sprachen findet man in Tabelle 5.3.

Tabelle 5.3.
Tests auf Proportionalität der Diagonalzellen für einige Sprachen

Sprache	V	FG
Hanunoo, V	21.05	2
Uraustronesisch, V	68.93	3
Bare'e,V	31.28	4
Angkola, V	635.75	4
Sundanesisch, V	1819.91	6
Hawaiisch, V	5.42*)	4
Tahitisch, V	9.23*)	4
Tuhamotu, V	7.94*)	4
Hawaiisch, C	7.78	2
Tahitisch, C	0.51*)	2
Tuhamotu, C	2.39*)	2
Afrikaans, C	5.41*)	2
Arabisch 1-2, C	82.82	2
Arabisch 1-3, C	54.75	2
Arabisch 2-3, C	16.17	2

*) H_0^{prop} wird angenommen ($\alpha = 0.05$)

6. Ausblick

Die vorgestellten Methoden zur Auswertung von Tendenzen der Strukturierung sprachlicher Einheiten sollen nicht nur der Beschreibung von Tendenzen dienen, sondern vor allem als Ansporn zur Entdeckung von Gesetzen. Eine Tendenz wird durch einen Test entdeckt, aber das bestreben des Linguisten besteht darin, die Frage zu beantworten, *warum* es die gegebene Tendenz gibt. Die Antwort kann nicht durch induktive Anhäufung von Einzelfällen gegeben werden, sondern nur durch ein deduktiv gewonnenes Gesetz. Um dieses Gesetz abzuleiten, muß man die Bedürfnisse der Sprachträger in Betracht ziehen (vgl. Köhler 1986), die "Kräfte", die aus diesen Bedürfnissen entstehen, postulieren und in eine Gleichung zusammenstellen, deren Lösung eine Hypothese ergibt. Gilt diese Hypothese für alle Sprachen (unter Berücksichtigung der Randbedingungen) und wird sie in ein System von anderen Hypothesen eingebettet, dann kann man sie als Gesetz bezeichnen.

Das Bestreben der theoretischen lingusitischen Forschung besteht gerade in der Etablierung von Gesetzen, denn nur Systeme von Gesetzen bilden Theorien, und nur Theorien können unsere Warum-Fragen beantworten.

Literatur

Altmann, G. (1964). "*Kvantitatívne štúdie z indonézistiky.*" Bratislava (Diss.).

Altmann, G. (1973). "Probabilistische Klassifikation von Konsonantenverbindungen des Indonesischen." In *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 123, 98-116.

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

Altmann, G. (1987). "Tendentielle Vokalharmonie". In *Glottometrika* 8, Bochum, 104-112.

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie.* Bochum: Brockmeyer.

Altmann, G., Schwibbe, M., Kaumanns, W., Köhler, R., und Wilde, J. (1987). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Erscheint 1987.

Bishop, Y.M.M., Fienberg, S.E. (1969). "Incomplete Two-Dimensional Contingency Tables." In *Biometrics* 25, 119-129.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivarate Analysis. Theory and Practise.* Cambridge, Mass., London: MIT Press.

Bowler, A.H. (1948). "A Test for Symmetry in Contigency Tables." In *JASA* 43, 572-574.

Bowler, T.D. (1981). *General Systems Thinking. Its Scope and Applicability.* New York, North Holland.

Caussinus, H. (1965). "Contribution a l'analyse statistique des tableaux de correlation." In *Annales de la Faculté des sciences de l'Université de Toulouse*, 77-183.

Cohen, J. (1960)."A Coefficient of Agreement for Nominal Scales." In *Educational And Psychological Measurement* XX, 1, 37-46.

Conklin, H.C. (1953). *Hanunoo - English Dictionary.* Berkeley-Los Angeles: UCP.

Dubovská, Z. *Struktura indonézské morfémy.* Praha s.d. (diss.).

Fienberg, S.E. (1970). "Qusi-Independence and Maximum Likelihood Estimation in Incomplete Contingency Tables." In *JASA* 65, 1610-1616.

Fleiss, J.L. (1973). *Statistical Methods for Rates and Proportions.* New York, London, Sydney, Toronto: John Wiley & Sons.

Fleiss, J.L., Cohen, J., Everitt, B.S., (1969). "Large Sample Standard Errors of Kappa and Weighted Kappa." In *Psychological Bulletin* 72, 5, 323-327.

Gerlach, R. (1982). "Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie." In *Glottometrika* 4, Bochum, 95-102.

Geršić, S., Altmann, G. (1980). "Laut – Silbe – Wort und das Menzerathsche Gesetz." In *Forum Phoneticum* 21, Frankfurt/M., 115-123.

Goodman, L.A. (1968). "The Analysis of Cross-Classified Data: Independance, Quasi-Independance and Interactions in Contingency Tables with or without Missing Entries." In *JASA* 63, 1031-1091.

Goodman, L.A., Kruskal, W.H. (1954). "Measures of Association for Cross Classifications." In *JASA* 49, 732-764.

Goodman, L.A., Kruskal, W.H. (1959). "Measures of Association for Cross Classifications." Part II. In *JASA* 54, 123-163.

Goodman, L.A., Kruskal, W.H. (1963). "Measures of Association for Cross Classifications." Part III. In *JASA* 58, 310-364.

Goodman, L.A., Kruskal, W.H. (1972). "Measures of Association for Cross Classifications." Part IV. In *JASA* 67, 415-421.

Greenberg, J.H. (1950). "Patterning of Semitic Verbal Roots." In *Word* 6, 162-181.

Haken, H. (1978). *Synergetics: An Introduction.* Heidelberg, Berlin, New York: Springer.

Hartung, J. (1984). *Statistik.* München, Wien: Oldenbourg.

Hess, Z. (1975). *Typologischer Vergleich der romanischen Sprachen auf phonologischer Basis.* Frankfurt/M., Bern: P.Lang/H.Lang.

Kendall, M.G., Stuart, A. (1967). *The Advanced Theory of Statistics.* Vol II, 2nd Ed., London: Griffin.

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

Krupa, V. (1966). "The phonemic structure of bi-vacalic morphemic forms in Oceanic languages." In *Journal of the Polynesian Society* 75, 458-497.

Krupa, V. (1967a)."Dissociations of like consonants." In *Asian and African studies* 3, 37-44.

Krupa, V. (1967b). "On phonemic structure of morpheme in Samoan and Tongan." In *Beiträge zur Linguistik und Informationsverarbeitung* 12, 72-83.

Krupa, V. (1971). "The phonetic structure of the morph in Polynesian languages." In *Language* 47, 668-684.

Mukherjee, R., Hall, J.R. (1967). "A Note on the Analysis of Data on Social Mobility." In Glass, D.V., (ed). *Social Mobility in Britain.* London.

Odendahl, F.F. (1962). *Die struktuur van die Afrikaanse wortelmorfeem.* Kaapstad, Pretoria: H.U.A.M.

Placket, R.L. (1974). *The Analysis of Categorical Data.* London: Griffin.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* 2nd Edition. New York, London, Sydney, Toronto: John Wiley & Sons.

Řehák, J., Řeháková, B. (1980). "Analyse von Kontingenztafeln: Zwei Grundtypen von Aufgaben und das Vorzeichenschema." In *Glottometrika* 3, Bochum, 1-28.

Ross, A.C.S. (1950). "Philological Probability Problems." In *Journal of the Royal Statistical Society* 12, 19-41.

Stuart, A. (1955). "A test for homogenity of the marginal distributions in a two-way classification." In *Biometrika* 42, 412-416.

Uhlenbeck, E.M. (1949). *De structuur van het Javaanse morphem.* Bandoeng: Nix.

Uhlenbeck, E.M. (1959). "The structure of the Javanese morpheme." In *Lingua* 2, 239-270.

Anhang

A1. Uraustronesisch

	a	e	i	u
a	428	56	152	216
e	127	95	44	92
i	155	24	99	77
u	173	16	70	239

A2. Indonesisch

	a	e	i	o	u
a	806	0	268	0	400
e	97	100	0	50	0
i	294	0	133	0	111
o	130	73	1	168	0
u	374	0	131	0	294
e	648	4	198	1	267

A3. Bare'e

	a	e	i	o	u
a	401	143	197	180	220
e	166	220	74	152	65
i	128	44	133	109	91
o	216	171	128	342	166
u	182	88	100	73	238

A4. Sundanesisch

	a	e	ĕ	ë	i	o	u
a	1261	254	261	116	569	324	593
e	239	352	16	0	15	220	15
ĕ	387	230	336	9	225	225	265
ë	98	0	0	222	56	0	1
i	365	13	128	68	311	37	215
o	272	202	12	0	34	505	3
u	438	20	147	23	205	9	585

A5. Hawaiisch, V

	a	e	i	o	u
a	72	50	47	49	48
e	50	55	22	35	33
i	43	19	55	42	25
o	45	40	43	49	36
u	45	34	33	11	48

A6. Tahitisch, V

	a	e	i	o	u
a	68	47	54	48	51
e	37	45	18	24	29
i	48	16	49	36	16
o	34	34	35	40	22
u	45	23	25	10	37

A7. Tuamotu, V

	a	e	i	o	u
a	92	62	72	62	67
e	54	55	31	37	35
i	59	22	62	53	25
o	51	44	50	60	30
u	54	36	37	12	49

A8. Hawaiisch, C

	F	M	B
F	17	88	101
M	40	58	101
B	79	113	136

A9. Tahitisch, C

	F	M	B
F	37	122	62
M	72	94	87
B	40	72	33

A10. Tuamotu, C

	F	M	B
F	33	128	105
M	80	91	138
B	59	134	93

A11. Afrikaans

	F	M	B
F	34	224	46
M	93	201	74
B	22	73	10

A12. Arabisch 1-2

	F	M	B
F	0	284	158
M	334	526	574
B	252	783	146

A13. Arabisch 1-3

	F	M	B
F	18	258	174
M	322	676	478
B	290	685	232

A14. Arabisch 2-3

	F	M	B
F	65	313	160
M	325	673	472
B	198	471	165

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 49–58

Ein Modell für die Variabilität der Vokaldauer*

S. Geršić
Köln

G. Altmann
Bochum

1. Die Tatsache, daß die Vokaldauer beim Sprechen schwankt, ist hinreichend bekannt und kann durch Messungen nachgewiesen werden. Würde man aber eine vollständige Zufälligkeit der Fluktuation der Vokaldauer annehmen, so müßte ihre empirische Häufigkeitsverteilung der Normalverteilung folgen, d.h. einem Zufallsgesetz, bei dem die Abweichungen von einem Mittelwert in positiver und negativer Richtung gleich wären (Symmetrie), die Häufigkeiten nähmen in beiden Richtungen vom Mittelwert anfangend zunächst monoton ab, ab $x \pm \sigma$ verliefen sie konvex, und näherten sich der x-Achse asymptotisch.

Die Tatsache, daß dies nicht der Fall ist, signalisiert die Existenz einer Steuerung der Vokaldauer durch eine Interaktion von mehreren Faktoren, die zu einem ganz bestimmten Typ der Häufigkeitsverteilung führen muß. Diese Faktoren kooperieren oder konkurrieren miteinander und erzeugen dadurch einen synergetischen Mechanismus (vgl. Haken 1978), dessen Entzifferung zur Erklärung und Prognose der Lautdauer führen kann, d.h. zu einem Gesetz der Vokaldauerverteilung.

Betrachten wir als Illustration die Verteilungen der Vokaldauer bei phonologisch kurzen Vokalen eines Sprechers des Batschka-deutschen Dialekts, wie sie von Geršić (1971) gemessen wurden (vgl. Tabelle 1). Von einer Normalverteilung kann hier keine Rede sein. Steht aber ein

* Diese Studie entstand im Rahmen des Projekts "Sprachliche Synergetik". Die Autoren bedanken sich bei der **Stiftung Volkswagenwerk** für die freundliche Unterstützung.

Tabelle 1
Verteilung der Vokaldauer der phonologisch kurzen Vokale eines Batschka-Deutschen Sprechers (Geršić 1971:54)

Vokaldauer in Millisekunden x'	Häufigkeit $f_{x'}$		
	kurz nichtbetont	kurz halbbetont	kurz betont
20 – 39	133	4	7
40 – 59	286	18	17
60 – 79	283	50	28
80 – 99	178	64	54
100 – 119	86	43	48
120 – 139	46	15	46
140 – 159	21	22	25
160 – 179	8	9	16
180 – 199	15	2	10
200 – 219	6	6	4
220 – 239	1	0	1
240 – 259	3	1	2
260 – 279	3	0	0
280 – 299	2	0	0
300 – 319	0	0	1
320 – 339	2	0	1
340 – 359	1	0	0
360 –	0	1	0

Gesetz dahinter, dann müssen alle diese Verteilungen einem einzigen Kurventyp jedoch mit unterschiedlichen Parametern folgen.

2. Bei der Aufstellung eines Modells für diese und ähnliche Daten gehen wir davon aus, daß sich auf die Vokaldauer drei Arten von Kräften auswirken:

(a) Kräfte, die für die Innovation, Emotionalität, den Ausdrucksdrang des Sprechers u.ä. zuständig sind und zur Fluktuation,

d.h. zu einer Abweichung von der Norm führen. Von dieser Art sind z. B. die Bühlersche Ausdrucksfunktion (vgl. Bühler 1934), die Zipfsche Diversifikationskraft (vgl. Zipf 1949) u.a.

(b) Kräfte, die lokal oder global die Vokaldauer fixieren oder modifizieren, wie phonologische Normen (kurz, lang), suprasegmentale (Ton, Akzent) und kombinatorische Faktoren (Umgebung).

(c) Kräfte, die den Sprecher einschränken, die Fluktuation eindämmen, für Gleichgewichtung sorgen wie z. B. die Bühlersche Darstellungsfunktion, die Zipfsche Unifikationskraft, Druck oder Kontrolle der Sprachgemeinschaft u.a.

Die ersten zwei Kräftearten (a) und (b) kooperieren in dem Sinne, daß sie die Vokaldauer gestalten und konkurrierend gegen die dritte Kraft (c) wirken.

Um dies alles in eine Formel zu bringen, reicht es, wenn wir die Veränderung der Wahrscheinlichkeitsfunktion $f'(x)$ oder, was noch einfacher ist, den Differentialquotienten $D = f'(x)/f(x)$, und die ihn gestaltenden Faktoren in Betracht ziehen, denn wir nehmen an, daß die "Kenntnis" aller Abhängigkeitsverläufe, Fließgleichgewichte der Sprache usw. in dem Sprachträger irgendwie unbewußt gespeichert sind: er weiß nicht, daß er nach Gesetzen handelt, genauso wie ein Stein nicht weiß, daß er z. B. nach Naturgesetzen fällt; er kann diese Gesetze nicht erlernen, ebenso wie er nicht erlernen kann, wie er sich nach Newtonschen Gesetzen bewegen sollte. Er handelt lediglich gemäß diesen Gesetzen.

Die Modellierung der Wahrscheinlichkeitsfunktion können wir also durch Inbezugsetzung der o.e. Kräfte durchführen.

Auf D wirkt nach der obigen Aufzählung linear die Ausdruckskraft des Sprechers (S), die wir als Sx symbolisieren, wobei x die Lautdauer repräsentiert. In der gleichen Richtung wirkt die von Geršić getrennt behandelte Art der Betonung, die wir als Bx bezeichnen. Die Vokaldauer wird global und normativ durch die phonologische Längennorm bestimmt, die einen konstanten Beitrag F liefert. Wie Lindbloom (1983) gezeigt hat, wird die Vokaldauer auch von der Umgebung modifiziert, vor allem von der Quantität der folgenden Konsonanten, die auf die Dauer invers wirkt, sodaß wir sie in Form K/x einsetzen können.

Die eindämmende Kraft, die der Hörer (d.h. die gesamte Sprachgemeinschaft) ausübt (H), wirkt genauso linear, wie die Kraft des Spre-

chers, aber invers zu allen modifizierenden Faktoren.

Setzt man alle diese Faktoren zusammen, so erhält man die Differentialgleichung

$$D = \frac{f'(x)}{f(x)} = \frac{Sx + Bx + F + K/x}{Hx} \tag{1}$$

Durch Reparametrisierung von (1) kann man die Zahl der Parameter reduzieren. Schreibt man

$$\frac{S+B}{H} = a \qquad \frac{F}{S+B} = b \qquad \frac{K}{S+B} = c$$

dann erhält man

$$\frac{f'(x)}{f(x)} = \frac{a(x + b + c/x)}{x} \tag{2}$$

$$= a + \frac{ab}{x} + \frac{ac}{x^2}, \tag{3}$$

d.h.

$$\frac{d[f(x)]}{f(x)} = \left[a + \frac{ab}{x} + \frac{ac}{x^2}\right] dx,$$

deren Lösung

$$f(x) = N x^{ab} e^{ax - ac/x}, \quad 0 < x < R \tag{4}$$

ergibt. Einfachheitshalber werden wir schreiben

$$f(x) = N x^{A} e^{Bx + C/x}, \quad 0 < x < R \tag{5}$$

wo $A = ab$, $B = a$, $C = -ac$ und N die Normierungsgröße

$$N = \left[\int_0^R x^A e^{Bx + C/x} dx\right]^{-1} \tag{6}$$

ist. Soweit uns bekannt, wurde diese Verteilung in der statistischen Literatur noch nicht verwendet, daher sind ihre Eigenschaften nicht bekannt.

3. Die Anpassung dieser Verteilung an die Daten stellt uns vor das Problem der Parameterschätzung und der Berechnung der Wahrscheinlichkeiten in einzelnen Intervallen. Sowohl Schätzungs- als auch Anpassungsprobleme lassen sich recht einfach mit modernen numerischen und Optimierungstechniken lösen, und die Resultate sind in der Regel besser als bei klassischen Schätzungen (Methode der kleinsten Quadrate, Maximum likelihood u.a.).

Um Verteilung (5) anzupassen, transformieren wir die ursprüngliche Variable von Geršić auf $X = X'/20$, sodaß die Verteilungen nun in Intervallen $(1,2), (2,3), \ldots$ verlaufen. Wir wählen Anfangswerte für die Parameter so, daß die Kurven lediglich konkav verlaufen, ohne Rücksicht auf die anfängliche Anpassungsgüte. Die Intergrale von (5) für die einzelnen Intervalle berechnen wir numerisch mit Hilfe der erweiterten Trapez-Regel, indem wir jedes Intervall in 10 Teilintervalle aufteilen. Das Resultat wird mit Hilfe des Chiquadrat-Kriteriums mit den Daten verglichen und mit Hilfe des Optimierungsalgorithmus von Nelder und Mead (1964) solange verbessert, bis bei einem voreingestellten Abbruchskriterium das minimale Chiquadrat erreicht wird. Die Daten, die angepaßten Verteilungen, die Parameter und die Testresultate sind in der Tabelle 2 zu finden.

Von den ursprünglichen Daten von Geršić konnten nicht alle benutzt werden, da in einigen Kategorien zu wenige Fälle vorlagen. Bei allen anderen wurde eine gute Übereinstimmung mit der theoretischen Kurve (5) festgestellt, auch wenn die Daten nicht immer einen "glatten" Verlauf aufwiesen. Besonders markant ist bei der Sprecherin I in der Kategorie "lang betont" die Häufigkeit von $x = 18 (f_{18} = 11)$, da sie sehr oft ein emotionsbeladenes "nee" benutzte. Solche "pathologischen" Fälle lassen sich durch Zusammenfassung mehrerer Klassen teilweise ausgleichen. Die Klassenzusammenfassung wurde überall dort verwendet, wo die theoretischen Werte kleiner oder gleich 1 waren bzw. dort, wo die beobachteten Werte begannen, recht unregelmäßig zu verlaufen. In der Tabelle 2 sind die Zusammenfassungen angegeben. Von insgesamt 15 überprüften Verteilungen haben nur 2 eine signifikante Abweichung von dem Modell aufgewiesen, nämlich die Kategorie KN und KB beim Sprecher III.

Die Schätzungen zeigten üblicherweise mehrere Lokalminima, die gleichgute Anpassungen lieferten. So kann man z. B. bei dem Spre-

Tabelle 2
Anpassung der Verteilung (5) an die Daten
von Geršić (1971)
(a) Sprecherin I

	KN		KH		KB		LH		LB	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	133	135.04	4	4.09	7	4.05	0	0.17	0	0.10
2	286	302.46	18	22.42	17	19.31	1	1.21	0	0.97
3	283	257.84	50	45.01	28	38.50	5	3.62	2	3.65
4	178	162.80	64	52.82	54	48.62	5	6.76	9	8.41
5	86	93.65	43	44.79	48	46.72	8	9.52	13	14.34
6	46	52.56	15	30.62	46	37.47	12	11.11	19	19.96
7	21	29.51	22	17.96	25	26.45	13	11.37	23	24.03
8	8	16.74	9	9.41	16	16.98	12	10.55	36	25.94
9	15	9.62	2	4.51	10	10.13	10	9.07	35	25.74
10	6	5.61	6	2.02	4	5.70	8	7.34	24	23.87
11	1	3.32	0	0.85	1	3.06	4	5.66	13	20.95
12	3	1.99	1	0.34	2	1.58	2	4.19	18	17.57
13	3	1.21	1	0.13	0	0.79	3	3.00	10	14.17
14	2	0.74	-	-	0	0.38	1	2.08	9	11.05
15	0	0.46	-	-	1	0.18	2	1.41	3	8.38
16	2	0.29	-	-	1	0.08	1	0.94	5	6.20
17	1	0.18	-	-	-	-	0	0.61	3	4.48
18	-	-	-	-	-	-	2	0.39	11	3.18
A =		-3.9433		5.8000		4.8474		4.9950		5.8602
B =		-0.2706		-1.4079		-1.0787		-0.7319		-0.6732
C =		-12.5325		-2.1749		1.8922		-2.2835		-1.4661
χ^2 =		10.41		6.80		10.22		5.01		13.70
FG =		5		3		9		11		9
P =		0.06		0.08		0.33		0.93		0.13

KN - kurz nichtbetont, KH - kurz halbbetont,
KB - kurz betont, LH - lang halbbetont, LB - lang betont

cher II in der Kategorie "kurz betont" mit den Parametern $A = 1.80$, $B = -1.19, C = -27.65$ ein fast genauso gutes Resultat erzielen, wie mit denen in der Tabelle 2.

Die Minimierung des Chiquadrats konnte man also beträchtlich, aber nicht beliebig, verbessern:

(i) durch die Wahl der Parameter, die man als Anfangswerte eingab,

Tabelle 2
Anpassung der Verteilung (5) an die Daten
von Geršić (1971)
(b) Sprecher II

	KN		KH		KB		LH		LB	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	79	71.62	2	2.18	1	0.11	0	0.04	0	0.02
2	256	261.42	21	19.78	4	3.92	0	0.51	1	0.38
3	255	291.24	51	52.80	27	23.08	2	2.18	1	2.60
4	245	210.77	77	73.56	38	53.49	2	5.01	9	8.68
5	147	130.04	66	69.42	79	70.76	9	8.35	15	18.59
6	70	75.50	47	50.77	75	64.96	17	11.09	31	29.45
7	39	42.92	39	31.05	45	46.28	15	12.55	45	37.58
8	26	24.32	14	16.64	24	27.41	10	12.61	47	40.77
9	10	13.84	9	8.07	11	14.11	10	11.58	38	38.98
10	5	7.94	1	3.61	4	6.51	13	9.90	34	33.70
11	3	4.60	2	1.51	4	2.75	6	7.98	24	26.83
12	4	2.69	1	0.60	1	1.08	8	6.13	15	19.95
13	0	1.59	-	-	1	0.40	0	4.52	6	14.00
14	0	0.95	-	-	1	0.14	1	3.22	13	9.35
15	1	0.57	-	-	-	-	3	2.23	5	5.98
16	-	-	-	-	-	-	3	1.50	4	3.69
17	-	-	-	-	-	-	0	0.99	1	2.20
18	-	-	-	-	-	-	2	0.64	5	1.27
A =		-4.0766		6.2025		7.6338		5.7272		8.6288
B =		-0.3063		-1.4722		-1.6480		-0.7670		-1.0340
C =		-15.8054		-4.5498		-10.7622		-3.9418		-2.8974
χ^2 =		16.98		3.83		11.52		16.35		12.88
FG =		9		6		7		11		11
P =		0.049		0.70		0.12		0.13		0.30

(ii) durch die Klassenzusammenfasung, die dem Programm vorgegeben wurde und
(iii) durch das Abbruchskriterium der iterativen Optimierung.

Tabelle 2
Anpassung der Verteilung (5) an die Daten
von Geršić (1971)
(c) Sprecher III

	KN		KH		KB		LH		LB	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	256	250.14	21	18.64	7	3.00	0	0.51	0	0.00
2	416	428.48	63	70.05	35	27.27	3	2.61	0	0.00
3	302	296.94	94	89.70	59	64.31	2	5.94	0	0.04
4	178	160.27	72	69.45	64	74.02	9	9.08	0	0.25
5	77	81.30	42	41.09	56	55.64	20	11.05	1	0.96
6	40	41.05	21	20.66	41	31.72	12	11.60	3	2.49
7	10	21.02	5	9.34	10	14.93	8	10.99	4	4.84
8	7	10.98	6	3.91	6	6.10	8	9.66	6	7.53
9	5	5.86	0	1.55	1	2.24	6	8.02	9	9.85
10	6	3.19	1	0.59	1	0.76	9	6.37	18	11.19
11	2	1.77	-	-	-	-	2	4.88	12	11.31
12	3	1.00	-	-	-	-	1	3.62	7	10.37
13	-	-	-	-	-	-	5	2.62	6	8.75
14	-	-	-	-	-	-	2	1.86	7	6.88
15	-	-	-	-	-	-	1	1.29	6	5.08
16	-	-	-	-	-	-	1	0.89	6	3.56
17	-	-	-	-	-	-	1	0.60	1	2.37
18	-	-	-	-	-	-	2	0.40	1	1.51
A =		-4.3658		2.3277		6.2348		3.4675		13.3025
B =		-0.2919		-1.2703		-1.7741		-0.6062		-1.1833
C =		-12.0462		-6.8291		-6.0341		-2.9606		1.0766
χ^2 =		14.43		3.50		11.02		15.05		9.64
FG =		7		4		4		8		10
P =		0.04		0.48		0.03		0.06		0.47

Die mißlungene Anpassung in drei Fällen (KN II, KN III, KB III), von denen zwei eng an der 95% Signifikanzgrenze stehen, kann durch verschiedene Umstände verursacht werden:

(i) Niedrige Häufigkeiten in einzelnen Klassen, die wohl auch die irregulären Häufigkeitsverläufe in Klassen mit höheren x verursachen.
(ii) Die nicht objektivierbare Einteilung der Häufigkeiten in Intervalle. Bei einer anderen Intervallwahl hätte man möglicherweise Verteilugen mit anderen Parametern und besserer Übereinstimmung erhalten.
(iii) Meßgenauigkeit bei Verarbeitung gesprochener Texte ist ein Problem, das nicht leicht zu bewältigen ist. Es reicht die Einordnung von einigen wenigen Intervallgrenzfällen in das angrenzende Intervall, um die ganze Verteilung völlig zu verändern. Die Messungen müßten in der Zukunft grundsätzlich vollständig veröffentlicht werden, damit man die beste Intervalleinteilung sozusagen experimentell aus den Daten "herausschälen" kann, was besonders dann wichtig ist, wenn man noch keine Hypothesen hat. So wie man mit Optimierungsmethoden die Parameter einer fertigen Verteilung berechnen kann, so ließe sich mechanisch vom Computer auch die beste Intervalleinteilung machen.

Wir können konstatieren, daß die durch Ableitung (1)–(3) gewonnene Wahrscheinlichkeitsdichte das Zusammenspiel der Kräfte, die die Vokaldauer steuern, adäquat wiedergibt. Die Überprüfung an so wenigen Fällen aus nur einer Sprache ist sicherlich nicht genügend, um sie fest zu etablieren, aber die deduktive Herleitung mit Hilfe der Kräfte, die auch in anderen Bereichen der Sprache wirken, gibt ihr den Status eines Kandidaten auf eine Gesetzeshypothese, die nun weiterer Überprüfungen bedarf.

4. Würde man nun aus den Parametern den Anteil einzelner Kräfte berechenen wollen, so wäre dies im heutigen Stadium recht verfrüht. Erstens stehen nur diese drei Datenmengen aus einer Sprache zur Verfügung, was für eine Verallgemeinerung kaum ausreicht; zweitens war es rein technisch unmöglich, einige Verteilungen zu überprüfen (z. B. Diphtonge), da die Häufigkeitsintervalle sehr schwach belegt waren. Die Zusammensetzung verschiedener Stichproben (von unterschiedli-

chen Sprechern stammend) sollte man vermeiden. Die Homogenität gesprochener Sprache ist illusorisch. Man kann sie mit dem Popperschen Mückenschwarm vergleichen, der als ganzer in einer berechenbaren Richtung fliegt, aber die Bewegungen einzelner Mücken sind unberechenbar. Man müßte sehr viele Stichproben von einem Sprecher haben, um sagen zu können, ob seine Parameter im Durchschnitt den Parametern anderer Sprecher gleichen.

Unser Ziel bestand eher im Versuch zu zeigen, daß auch phonetische Erscheinungen durch Zusammenwirken verschiedener dimensionsloser Kräfte zustandekommen und daher theoretisch ableitbar und in eine künftige Sprachtheorie einbettbar sind.

Literatur

Bühler, K. (1934). *Sprachtheorie.* Frankfurt: Fischer.

Geršić, S. (1971). *Mathematisch-statistische Untersuchungen zur phonetischen Variabilität am Beispiel von Mundartaufnahmen aus der Batschka.* Göppingen: Kümmerle.

Haken, H. (1978). *Synergetics.* Berlin: Springer.

Lindbloom, B. (1983). "Economy of speech gestures". In MacNeilage, P. (Hrsg). *Speech production.* New York.

Nelder, I.A., Mead, R. (1965). "A simple method for function minimization". *Computer Journal* 7, 308-313 (8, 1965, 27).

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

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 59–104

Reihenfolgebeziehungen in der syntaktischen sprachtypologie

Wolf Thümmel
Göttingen

1. Syntaktische sprachtypologie ist nicht eine erfindung dieses jahrhunderts. Die morphologische spielte zwar lange zeit eine dominierende rolle. Doch schon Hans Conon von der Gabelentz hat 1861 – ganz im rahmen der damals and den universitäten noch nicht so recht wirksam gewordenen Beckerschen vorstellungen – sprachtypologische aussagen gemacht, die man nicht anders als syntaktisch nennen kann. Und sein sohn Georg hat bereits 1874/75 auf sprachen hingewiesen, in denen das objekt vor dem subjekt steht, auf sprachen also, nach denen man noch hundert jahre später ausschau hielt (Pullum hat 1981 einen katalog von dreißig derartigen sprachen zusammengestellt, von denen er allerdings zwölf (nämlich sieben OVS- und fünf OSV-sprachen) 1977 noch unter die rubrik "do not occur at all, contra various allusions in the literature on syntactic typology" fallen läßt). Die sechs schemata "der drei Hauptbestandteile des Satzes" hat Wundt später (1904:357ff.) in bezug auf ihre je "verschiedene" "psychologische Bedeutung" betrachtet (s. fig. 1).

1. S V O	**2. V O S**	**3. O V S**
Romulus condidit Romam	*Condidit Romam Romulus*	*Romam condidit Romulus*
1ª S O V	**2ª V S O**	**3ª O S V**
Romulus Romam condidit	*Condidit Romulus Romam*	*Romam Romulus condidit*

Figur 1
Wundts Greenberg-variation über S, V, O

Man kann sich des eindrucks nicht erwehren, daß die S-, O-, V-kombinationsschemata die syntaktische sprachtypologie bis zum heutigen tage beherrschen, besonders seit Greenberg (1963) sie als einen von fünf faktoren, die seine typologie der basisordnung ausmachen, erneut ins spiel gebracht hat (s. hierzu Grunig & Thümmel 1980: 146ff). – Diese Vorliebe für ausdrucksstücke, die S, O oder V zugeordnet werden, ist zumindest angesichts der folgenden umstände verwunderlich:

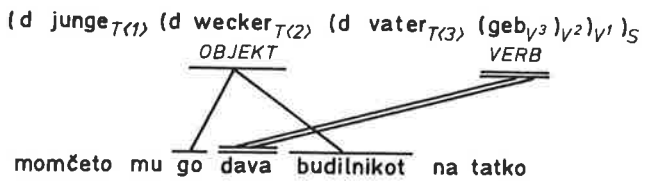
i. 'Subjekt' als universell anzuwendender begriff ist seit Keenans versuch (1976), dafür einige definitionsstücke zu nennen, keineswegs klarer geworden als vorher; das gegenteil ist offenbar der fall.

ii. Selbst wenn man die begriffe 'subjekt' und 'objekt' in bezug auf gewisse einzelsprachen einmal als bestimmt oder als bestimmbar voraussetzt, ergeben sich schwierigkeiten, sobald es um die reihenfolge der sie manifestierenden ausdrucksstücke geht. Man nehme etwa die makedonische übersetzung (2) des deutschen beispiels (1), an dem Bartsch & Vennemann (1982:138) ihre wortstellungstypologischen vorstellungen erläutern:

(1) (Weil$_1$) der$_2$ Junge$_3$ den$_4$ Wecker$_5$ dem$_6$ Vater$_7$ gibt$_8$
(2) (Začto$_1$) momče$_3$to$_2$ mu$_{6,7}$ go$_{4,5}$ dava$_8$ budilnik$_5$ot$_4$ na$_6$ tatko$_7$

Analysiert man (2) Bartsch & Vennemann folgend, so ergibt sich u.a. für O und V die schwierigkeit, daß man wegen der sog. pronominalantizipation ohne besondere vereinbarungen nicht mehr entscheiden kann, ob V vor O oder O vor V steht; fig. 2 veranschaulicht dies.

iii. In den mir bekannten sprachen sind die wenigsten (gesprochenen oder geschriebenen) satzmanifestationen so simpel, daß sie sich einzig als manifestation einer der sechs anordnungen von S, V, O beschreiben lassen. – Selbst wenn man noch – in ihrer universellen anwendbarkeit fragwürdigen – kategorien wie präpositionen, postpositionen, hilfsverben, attribute, genitiv u.ä. hinzunimmt, wird die zahl der in die typologie eingehenden syntaktisch beschreibbaren erscheinungen in nicht-gerechtfertigter weise reduziert. Syntaktisch beschreibbar ist etwa das besondere kombinatorische verhalten von "echter" im niederländischen, das offenbar unter keine der genannten und unter keine der sonst in



Figur 2
Das verb zwischen den teilen des objekts im makedonischen

der typologischen literatur herangezogenen kategorien fällt:

(3)(a) AAAA. *Echter* BBBB, *omdat* CCCC.
(b) AAAA. *Omdat* CC *echter* CC, BBBB.
(c) AAAA. *Omdat* CCCC, BB *echter* BB.
(d) *AAAA. BBBB, *omdat* CC *echter* CC.

AAAA, BBBB, CCCC steht für ausdrucksstücke, in denen je mindestens ein finites verb als teilstück vorkommt.

Ich werde meinen weg so einschlagen, daß reihenfolgebeziehungen aller segmentierbaren ausdrucksstücke, gleichgültig welchen abstrakten syntaktischen kategorien man sie zuweisen mag, grundsätzlich berücksichtigt werden. Neues über S-, V- und O-ordnungen sagen zu wollen, das typologisch fruchtbar wäre, hielte ich für einen kühnen wunsch. Ich gehe deshalb auch nicht auf neuere arbeiten über diese anordnungen wie die von Vennemann, von Krupa (1982), von Dezső (1982) von Hawkins (1983) oder von Ramat (1985) ein.

2. Einen von der tradition der SVO-typologie unabhängigen weg in der syntaktischen sprachtypologie ist Uspenskij gegangen. Seine grundidee ist, die syntaktisch zu vergleichenden sprachen auf ein étalonsystem zu beziehen, das die vier grundwortarten A (für: adjektiv), N (für: nomen), V (für: verb), Adv (für: adverb) sowie eine nukleuskonstruktion N;V umfaßt. Über diesem grundvokabular werden sieben sog. äquivalenzregeln formuliert.

(4)	(1)	AN ↔ N	*novyj dom, dom* ('neues haus, haus')
	(2)	NN ↔ N	*klass slov, klass* ('klasse von wörtern, klasse')
	(3)	NV ↔ V	*na bul'vare guljaet, guljaet* ('auf dem boulevard spaziert, spaziert')
	(4)	VN ↔ V	*postroil jatyk-étalon, ržët* ('konstruiert eine étalon-sprache, wiehert')
	(5)	AdvV ↔ V	*formal'no ograničit', ograničit'* ('formal beschränken, beschränken')
	(6)	NA ↔ A	*krasivyj licam, krasivyj* ('schön von angesicht, schön')
	(7)	AdvA ↔ A	*očen' xorošij, xorošij* ('sehr gut, gut')

Die konfiguration (4)(4) heißt bei Uspenskij 'kompletiv', die übrigen konfigurationen heißen 'attributiv'. Das étalon-system soll nach Uspenskij dem system für das russische entsprechen. Die beispiele in (4) gebe ich – hoffentlich – gemäß den intentionen von Uspenskij.

Im unterschied zu den bisher erwähnten bemühungen um syntaktische sprachtypologie sind es bei Uspenskij nicht die reihenfolgebeziehungen der "wörter" oder "wort"folgen, die in die typologie eingehen; mit dem étalon-system soll für jede sprache das "einheitliche funktionieren der wörter im satz" (Uspenskij 1965:145) beschrieben werden: Wenn Z, Y, X, W ... beliebige voneinander verschiedene morphologisch bestimmte wortarten einer sprache sind, so ist das funktionieren der elemente von Z innerhalb des satzes festgelegt durch sog. äquivalenzregeln der form (5):

(5)	(a)	(b)	(c)	
(1)	YZ ↔ Z	XZ ↔ Z	WZ ↔ Z	...
(2)	ZY ↔ Z	ZX ↔ Z	ZW ↔ Z	...
(3)	YZ ↛ Y,Z	XZ ↔ X,Z	WZ ↛ W,Z	...

An hand von entsprechungsregeln wie (6) lassen sich die systeme für die einzelnen sprachen – hier für das dänische – aus dem étalon-system ableiten:

(6)	(ER 1)	$N_{étalon}$	entspricht	$N_{dänisch}$
	(ER 2)	$V_{étalon}$	entspricht	$V_{dänisch}$
	(ER 3)	$A_{étalon}$	entspricht	$A_{dänisch}$
	(ER 4)	$Adv_{étalon}$	entspricht	$A_{dänisch}$

Den intentionen Uspenskijs gemäß läßt sich (7) als system für das dänische exemplifizieren:

(7)	(1)	AN ↔ N	*lille pige, pige* ('kleines mädchen, mädchen')
	(2)	NN ↔ N	*sprogets struktur, struktur* ('die struktur der sprache, die struktur')
	(3)	NV ↔ V	*sover i sengen, sover* ('schläft im bett, schläft')
	(4)	VN ↔ V	*køber en bog, sover* ('kauft ein buch, schläft')
	(5)	AV ↔ V	*sover hurtig, sover* ('schläft schnell, schläft')
	(6)	NA ↔ A	*dygdig i striden, dygdig* ('tapfer im kampf, tapfer')
	(7)	AA ↔ A	*dygdig træt, træt* ('ziemlich müde, müde')

An hand von (5) (1) und (2) ist ersichtlich, daß für das étalon-system und für die daraus durch "entsprechungsregeln" abgeleiteten systeme der zu klassifitierenden sprachen die reihenfolge auch bei Uspenskij eine rolle spielt, wenngleich eine ganz besondere. (1) und (2) in (5) repräsentieren konstruktionen, die man seit Bloomfield (1933) endozentrische nennen kann ((3) repräsentiert exozentrische konstruktionen), (1) steht für "attributive verknüpfungen", (2) für "kompletive verknüpfungen". Der unterschied soll formal dadurch ausgedrückt werden, daß diejenige komponente der konstruktion, die – um mit Bloomfield zu sprechen – reduktionsresultat ist, für attibutive verknüpfungen rechts, für kompletive links steht.

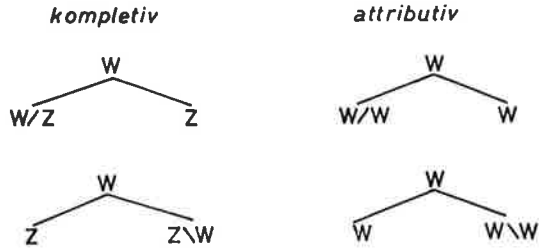
Was Uspenskij mit der unterscheidung zwischen attributiver und kompletiver verknüpfung erfassen will – die möglichkeit iterativer regelanwendung bei attributiven gegenüber der unmöglichkeit iterativer regelanwendung bei kompletiven verknüpfungen – wird durch die links-/rechts-normierung in (5) nicht oder zumindest nicht geschickt ausgedrückt.

Es gibt weitere formale schwächen in Uspenskijs vorschlag. Bedeutsam für den hier gegebenen zusammenhang ist, daß Uspenskij (1965:157) einen aufsatz von Bar-Hillel zitiert, der ihm – mit hilfe der von Bar-Hillel vorgestellten kategorialen syntax – in folgender weise hätte weiterhelfen können:

(a) Die unterscheidung zwischen kopf und modifikator (head and modifier) von endozentrischen konstruktionen hätte formal ausgedrückt werden können.

(b) Die reihenfolge von kopf und modifikator hätte – zumindest in einfachen fällen – expliziert werden können.

(c) Die unterscheidung zwischen "attributiven" und "kompletiven" hätte auf einfache weise expliziert werden können, z.b. wie in Fig.3.

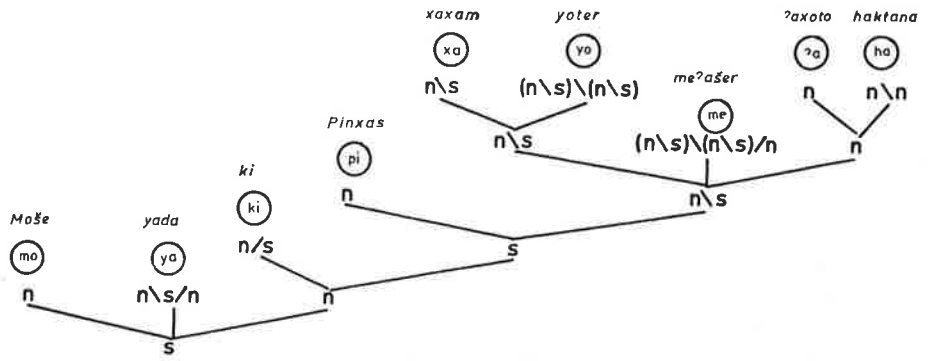


Figur 3
Uspenskijs unterscheidung zwischen "attributiven" und "kompletiven", ausgedrückt mit den mitteln einer kategorialen syntax

3. Da es mir im folgenden um die unterscheidung zwischen kopf und modifikator – oder zumindest um eine sehr ähnliche unterscheidung –

sowie um die reihenfolge von ausdrucksstücken, die den beiden begriffen zugeordnet werden, geht und da kategoriale syntaxen eigenschaften aufweisen, die für meine überlegungen entscheidend sind, gehe ich von einem beispiel aus, das Bar-Hillel in dem von Uspenskij zitierten, aber nicht nutzbar gemachten aufsatz gibt. Bar-Hillel exemplifiziert das funktionieren einer kategorialen syntax an dem satz (8), wofür er die struktur in fig. 4 anschreibt.

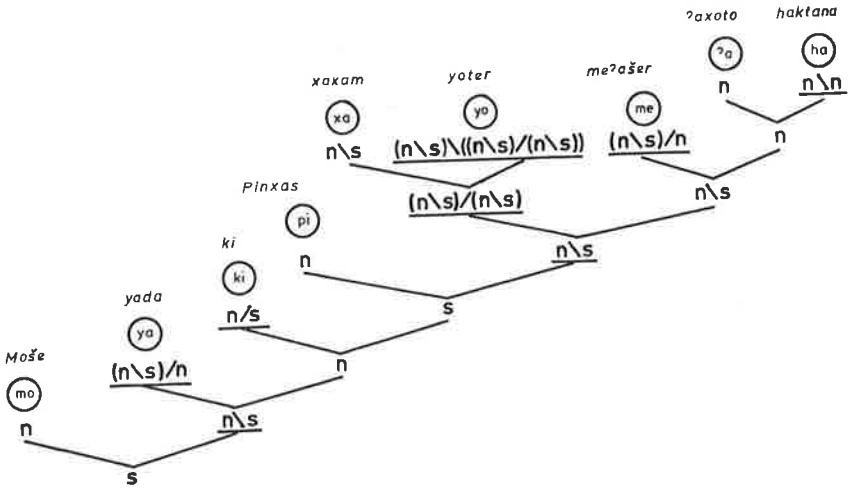
(8) Moše$_1$ yada$_2$ ki$_3$ Pinxas$_4$ xaxam$_5$ yoter$_6$ me?aser$_7$?axoto$_8$ haktana$_9$
'Moses$_1$ wußte$_2$, daß$_3$ Pinchas$_4$ klüger$_{5,6}$ ist als$_7$ seine$_8$ kleine$_9$ schwester$_8$.'



Figur 4
Die struktur, die Bar-Hillel dem ausdruck (8) zuordnet

In fig. 4 gibt es binäre und ternäre verzweigungen. Da ich mich im folgenden auf maximal binärverzweigende strukturen beschränken will, modifiziere ich fig. 4 derart, daß dem satz (8) die binärverzweigende struktur fig. 5 zugeordnet wird.

Will man die kopf-/modifikator-unterscheidung auf fig. 5 anwenden, so kann man sagen, daß die unterstrichenen symbole den modifikatoren, die jeweiligen nicht-unterstrichenen schwesternknotenetiketten den köpfen entsprechen.

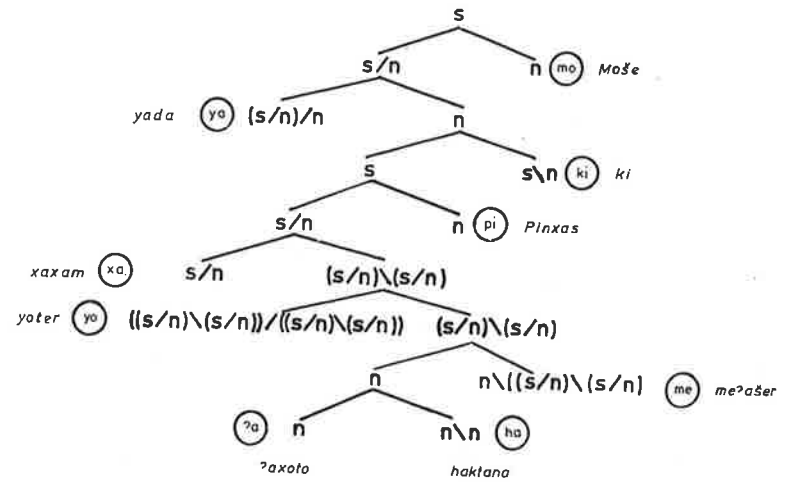


Figur 5
Binärverzweigende struktur für (8)

Die systeme, über denen die strukturen in fig. 4 und fig. 5 konstruiert sind, haben eine in verschiedenen hinsichten unerwünschte eigenschaft, die aber hier und im folgenden nicht stören soll: sie sind nicht fundiert, d.h., es gibt mindestens eine kategorie – n/s in fig. 5 –, die sowohl funktor als auch argument ist. Entsprechendes gilt für die figuren 6 und 9.

Fig. 5 ist eine von vielen in frage kommenden möglichen strukturen, die man dem satz (8) zuordnen kann. Die wahl der einen oder anderen struktur erfolgt – oder sollte wenigstens erfolgen – auf der grundlage umfassender empirischer untersuchungen nach einheitlichen methodologischen prinzipien. Ich will hier nicht über die breite palette der kriterien sprechen, die unter linguisten oder auch nur von einzelnen linguisten als empirische betrachtet werden, es genügt mir hier, zu sagen, daß fig. 4 – und trotz der vorgenommenen modifikation auch fig. 5 – nach bestimmten vorstellungen über semantik konstruiert worden sind, und daß distributionelle kriterien des zu bestimmenden typs

bei verwendung der gleichen formalen mittel und über dem gleichen basisvokabular (s, n) eher eine struktur wie fig. 6 nagelegen, ohne eine semantische beschreibung grundsätzlich auszuschließen.



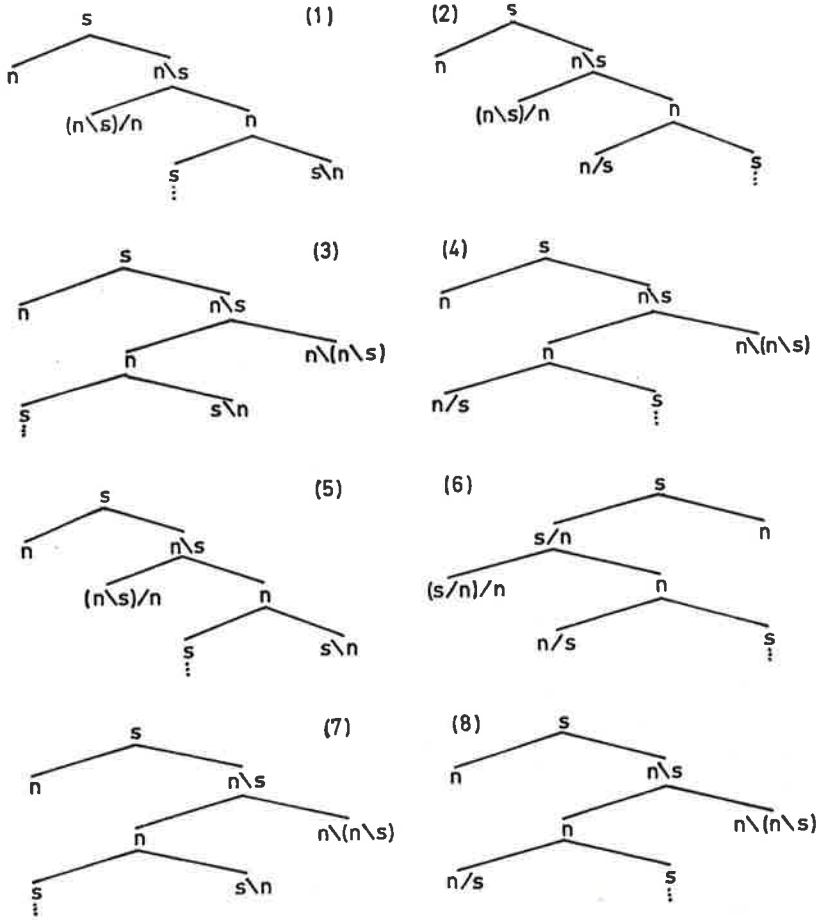
Figur 6
Alternative binärverzweigende struktur für (8)

Will man unter beibehaltung der in fig. 6 gegebenen hierarchisierung (d.h. der dominanzbeziehungen) auch (9) eine strukturbeschreibung zuordnen, so wird man eine der strukturierungen von fig. 7 wählen wollen.

(9) yada Moše ki (Pinxas xaxam yoter me?ašer ?axoto haktana)

Aber: keine der acht strukturierungen in fig. 7 erlaubt – ohne weitere vorkehrungen – die beschreibung von (9). Wenn man nun davon ausgeht, daß

(a) kategoriale syntaxen dominanzbeziehungen erfassen,
(b) die dominanzbeziehungen empirisch (und das heißt hier: distributionell) gerechtfertigt sind,



Figur 7
Acht kandidaten für eine strukturbeschreibung von (9)

(c) kategoriale syntaxen geeignet sind, die distributionell festlegbare unterscheidung zwischen kopf und modifikator formal zu treffen,
(d) eine kategoriale syntax, die (a) bis (c) respektiert, nicht unbedingt beliebige reihenfolgebeziehungen wiedergibt,

so liege es nahe, sich zu fragen, ob solche syntaxen überhaupt reihenfolgebeziehungen zwischen (oder wie man – wegen der fehlenden identitivität mißverständlich – auch sagt: die lineare ordnung von) teilen sprachlicher ausdrücke erfassen sollen.

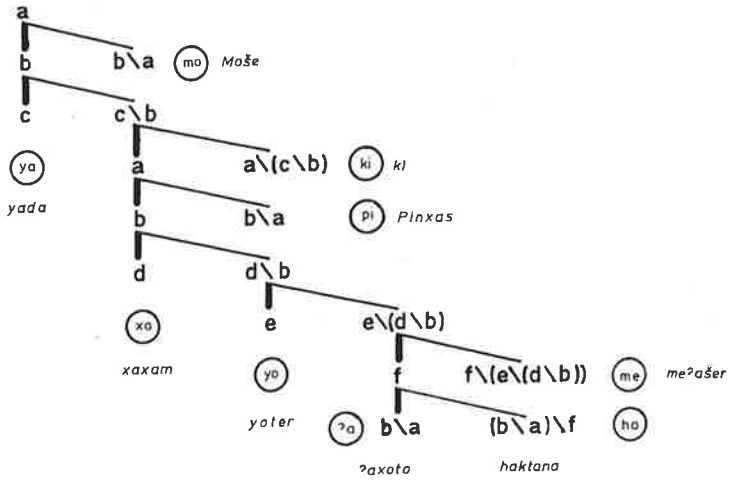
4. Gibt man besagtes ziel auf, so könnte man etwa so verfahren, wie es von etlichen autoren vorgeschlagen wurde (z.b. Staal 1967; Peterson 1979) oder die reihenfolgebeziehungen von den dominanzbeziehungen durch zwei sorten von regeln (den regeln der unmittelbaren dominanz einerseits und den regeln der linearen ordnung andererseits) trennen, wie Gazdar vorschlägt (Gazdar & Pullum 1982, Gazdar & Klein & Pullum & Sag 1985).

Ich will einen grundsätzlich anderen weg aufzeigen und in der formalen darstellung nicht nur an den reihenfolgebeziehungen festhalten, sondern sie zugleich zum ausdruck der kopf-/modifikator-unterscheidung nutzen: Ich will bäume so normieren, daß von zwei schwesternknoten der linke den kopf, der rechte den modifikator darstellt (s. fig. 8). Den linken knoten nenne ich einfach 'kopfknoten'. Ich nenne diese normierung 'rechtsnormierung'. Dabei soll vorausgesetzt sein, daß man die unterscheidung auf grund empirischer und formaler gegebenheiten treffen kann. Ich setze außerdem voraus, daß eine unterscheidung wie die zwischen kopf und modifikator auch für konstruktionen getroffen werden kann, die nach Bloomfield exozentrisch zu nennen wären (hierzu Clément & Thümmel 1981: 87ff).

Betrachtet man die strukturierung in fig. 6 als distributionell angemessen, läßt die beschränkung auf zwei basiskategorien (s, n) fallen und behält die hier verwendete schrägstrichnotation für die kopf-/modifikator-unterscheidung bei, so kann man für fig. 6 eine rechtsnormierte struktur der form fig. 9 anschreiben. Die fetten, senkrechten kanten verbinden die kopfknoten einer kopfknoten-hierarchie.

Mit der rechtsnormierung gibt es in der formalen darstellung – im baumgraphen – notwendigerweise reihenfolgebeziehungen, aber sie werden in einen neuen dienst gestellt: sie dienen nicht der beschreibung

Figur 8
Kopf-/modifikator-schema



Figur 9
Rechtsnormierte struktur für (8)

der linearen abfolge natürlichsprachlicher ausdrucksstücke, sondern zur unterscheidung zwischen kopf und modifikator.

Der vorteil, den die beibehaltung der reihenfolgebeziehungen in der

formalen darstellung hat, liegt vor allem darin, daß man nicht nur mit allen bekannten sätzen über kategoriale syntaxen, über die beziehungen zwischen kategorialen und kontextfreien syntaxen sowie über alle verwandten systeme in vertrauter weise rechnen kann, sondern auch die mit fig. 9 gegebene hierarchisierung nicht verändern muß.

Da mit der rechtsnormierung die schrägstrichnotation (oder eine äquivalente notation) überflüssig wird – jedenfalls für die unterscheidung zwischen kopf und modifikator –, entfallen auch die mit der eventuellen – in den bisherigen beispielen gegebenen – nicht-fundiertheit zusammenhängenden probleme. An stelle der knotenetikettierenden kategoriennamen können paare von typen und signaturen treten, wie sie in der theorie des Artikulators von Blanche-Noëlle Grunig (1981:51ff und speziell: 93ff) axiomatisch festgelegt sind, oder ähnliche formale gebilde (tupel von merkmalmengen) wie in der Generalisierten Phrasenstrukturgrammatik (s. hierzu Thümmel 1986:290ff).

5. Beim numerischen sprachenvergleich gibt man eigenschaften der zu vergleichenden sprachen als zahlen (meist als rationale zahlen) an. Diese zahlen bezeichnet man als die werte eines bestimmten indexes. Als syntaktischen index für den typologischen sprachenvergleich hat Andreev (1967), bezogen auf dependenzsyntaxen, ein maß der zentralisiertheit definiert. Eine neufassung dieses maßes findet sich bei Altmann & Lehfeldt (1973:120f). Ich werde – in dem von mir gewählten syntaktischen rahmen – ein maß der linksorientiertheit wenn auch nicht definieren, so doch skizzieren. Dieses maß weist gewisse ähnlichkeiten mit dem maß von Andreev auf, die ich hier nicht erörtern kann. Das maß der linksorientiertheit soll – grob gesprochen – angeben, wie weit die abstrakte rechtsnormierte darstellung der kopf-/modifikator-abfolge von der abfolge der sie manifestierenden tatsächlichen ausdrucksstücke abweicht.

Daß ich die folgenden erörterungen nicht auf dependenzsyntaxen, sondern auf kontextfreie syntaxen beziehe, hat vor allem zwei gründe: einen praktischen und einen theoretischen; der praktische: Ich selbst habe ausschnitte natürlicher sprachen bisher nicht mit dependenzsyntaxen, sondern mit kontextfreien syntaxen oder verwandten systemen zu beschreiben versucht; der theoretische: Ihre formalen eigenschaften machen dependenzsyntaxen nur in eingeschränkter weise tauglich für

den generellen sprachenvergleich (s. das theorem der instabilität des (elementaren) dependenziellen artikulators in der theorie des Artikulators von Blanche-Noëlle Grunig 1981: 494ff).

Gegeben sei eine terminale kette $\varphi = \sigma_1 \widehat{} \ldots \widehat{} \sigma_n$ samt rechtsnormierter strukturbeschreibung s (in baumform). s läßt sich in endlich viele teilstrukturen $t_1, \ldots, t_p$ zerlegen, wie in fig.10 für die struktur in fig. 9 veranschaulicht.

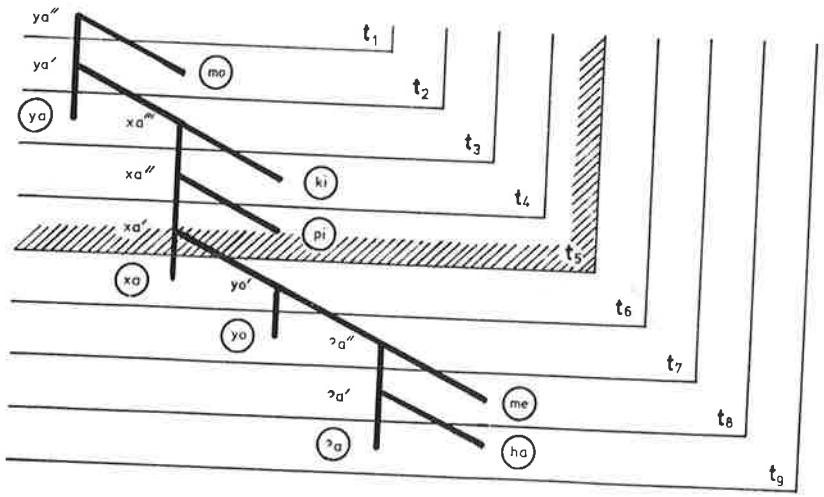
Es gibt in jeder teilstruktur $t_j(s)$ eine folge von finalknoten $\zeta_{o_j}, \ldots, \zeta_{q_j}$ (mit $1 \leq j \leq p; q = j-1$), wobei ich unter 'finalknoten' jeden $t_j(s)$ terminierenden knoten ζ_{i_j} verstehe (mit $0_j \leq i_j \leq q_j$).

Der grund, weshalb teilstrukturen von s berücksichtigt werden sollen, ist der, daß ich nicht annehmen will, nur (echte) terminalknoten etikettierende symbole seien einer manifestation zuzuordnen; vielmehr nehme ich an, daß auch nicht-terminalknoten (eben alle finalknoten) etikettierende symbole mitsamt den von ihnen dominierten strukturen manifestationen haben können. Das entspricht nicht nur der empirischen deutung sog. struktursyntaxen (Maurer 1969:185ff), es entspricht auch dem traditionsreich und annahmenmager angelegten theoretischen system von Blanche-Noëlle Grunig (1981:108ff und insbesondere 137ff) (die möglichkeit sog. perforierter inskriptionen – von structures inscrites à trou – lasse ich hier unberücksichtigt).

Die finalknoten von $t_p(s)$ sind die mit $\sigma_1, \ldots, \sigma_n$ etikettierten terminalknoten von s. Um in den teilstrukturen die finalknoten, die nicht mit einem element aus φ etikettiert sind, ebenfalls zu etikettieren, wird das in fig. 10 veranschaulichte verfahren angewendet: Sei σ ein symbol, das einen kopfknoten etikettiert (einen knoten in einer senkrechten kantenfolge), so wird der unmittelbar dominierende knoten mit σ' etikettiert. In fig. 10 wird der mit xa etikettierte kopfknoten unmittelbar durch den mit xa' etikettierten knoten dominiert, der mit xa' durch den mit xa'' etikettierten usw. Die um derart gebildete symbole erweiterte menge der terminalsymbole sei für jede teilstruktur t_j notiert als tupel $T_j = (\tau_{1_j}, \ldots \tau_{m_j})$, wobei für jedes $k_j (1 \leq k_j \leq m_j = j)$ gilt: $k_j = i_j + 1$. Die elemente auf T_j sollen finalsymbole heißen.

In t_5 – beispielsweise – ist $(\zeta_0, ya), (\zeta_1, xa'), (\zeta_2, pi), (\zeta_3, ki), (\zeta_4, mo)$ die folge der finalknoten. t_5 ist durch schraffur hervorgehoben.

Fig. 10 läßt nicht erkennen, daß sich zwei verfahren anbieten, in baumform angegebene strukturbeschreibungen in teilstrukturen zu zerlegen. Das soll an hand von fig. 11 und fig. 12 verdeutlicht werden.



Figur 10
Die teilstrukturen der rechtsnormierten struktur in Fig. 9

Fig. 11 zeigt die teilstrukturbildung von rechts nach links, fig. 12 die von links nach rechts.

Für jeden etikettierten finalknoten jeder teilstruktur wird dann mit der formel (10) ein basiswert errechnet:

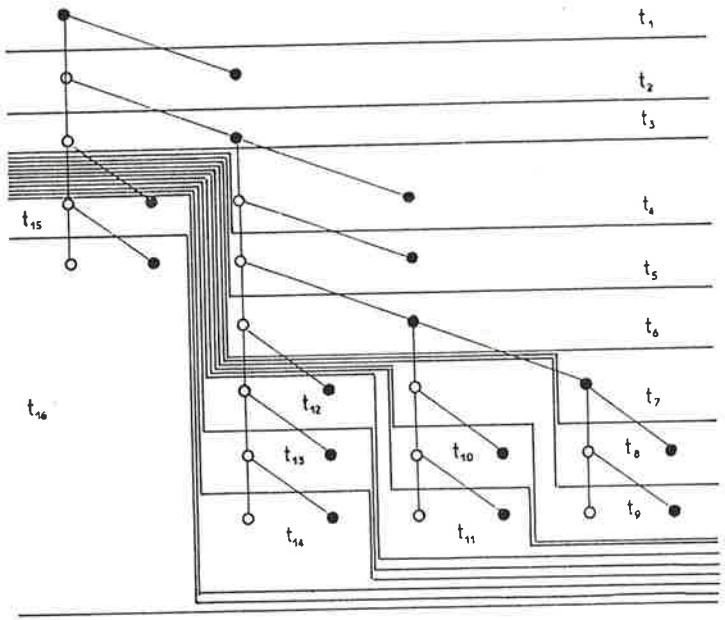
(10) $$\text{bas}(\zeta_{i_j}, \tau_{k_j}) = 2^{q_j - 1} - 2^{q_j - k_j}$$

In t_5 ist xa' das etikett von ζ_1, so daß $i_5 = 1$ und $q_5 = 5$. Setzt man diese werte in (10) ein, so erhält man (11); (12) und (13) zeigen die werte für die mit pi etikettierten finalknoten in t_5 bzw. in t_7:

(11) $$\text{bas}(\zeta_{1_5}, xa') = 2^{5-1} - 2^{5-2} = 8$$

(12) $$\text{bas}(\zeta_{2_5}, pi) = 2^{5-1} - 2^{5-3} = 12$$

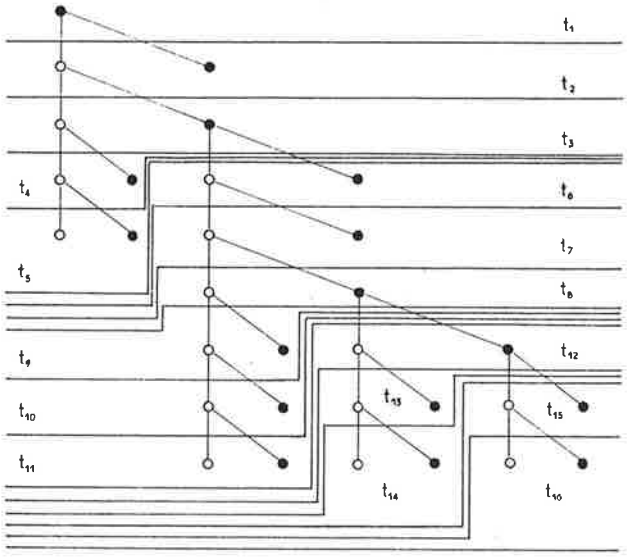
(13) $$\text{bas}(\zeta_{4_7}, pi) = 2^{7-1} - 2^{7-5} = 60$$



Figur 11
Teilstrukturbildung von rechts nach links

In tab. 1 gebe ich die basiswerte für die finalknoten in den einzelnen teilstrukturen der struktur von fig. 10 an.

Es ist leicht zu errechnen, daß eine zerlegung von links nach rechts eine andere gewichtung mit sich bringt, als eine von rechts nach links. Läßt man alle zur selben kopfknotenhierarchie gehörenden knoten außer acht, so erhalten bei der zerlegung von links nach rechts die einem kopfknoten x näheren, bei der von rechts nach links die einem kopfknoten x ferneren knoten einen höheren wert. Zwar wirkt sich der unterschied nicht auf das hier erörterte beispiel aus, aber für den allgemeinen fall muß zwischen den beiden richtungen entschieden werden (eventuell kann man beide berücksichtigen und in geeigneter wiese in die bewer-



Figur 12
Teilstrukturbildung von links nach rechts

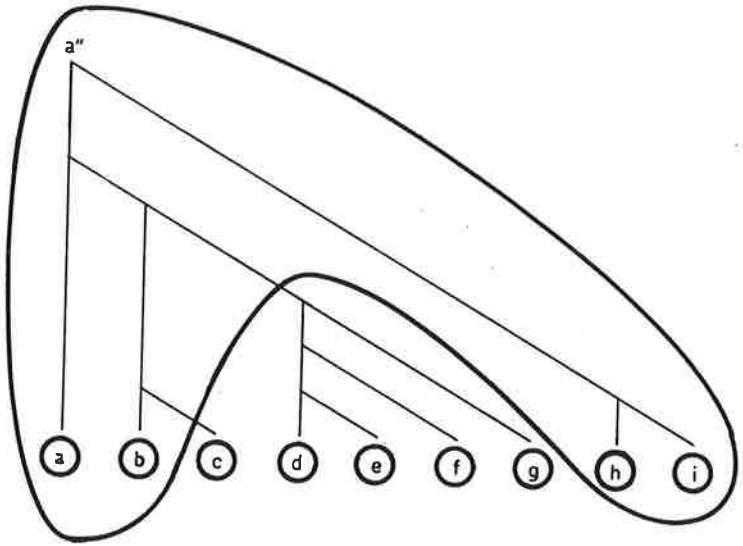
tungskalkulation eingehen lassen). Vorerst scheint es mir plausibel, die mit der zerlegung von rechts nach links verbundene gewichtung als die günstigere anzusehen. Jedenfalls soll hier davon ausgegangen werden.

6. Strukturen wie in fig. 9 – und erst recht in fig. 10 – sind abstrakt, d.h., es sind theoretische gebilde. Vereinbarungsgemäß sind reihenfolgebeziehungen nach dem kopf-/modifikator-reglement effektiv präsent. Man sollte aber nicht glauben, daß genau diese reihenfolgebeziehungen, die über (einfachen oder komplexen) strukturen und somit jedenfalls über abstrakten gegenständen innerhalb der beschreibung erklärt sind, auch in den beschriebenen gegenständen zu finden sind. Vor solchem glauben sollte schon die fabel von jenem wolf warnen, der – bekanntlich mit ganz unerwünschtem ergebnis – nach dem spiegelbild des fleischstücks schnappte, das er im maul hielt. Trotz solch alter weisheit kann man – merkwürdig genug – geradezu von einer

Tabelle 1
Die basiswerte für die teilstrukturen in fig. 10

t_j	y_d^*	x_d^*	y_o^*	γ_a^*	h_a	m_e	p_i	k_i	m_o	q
t_1	τ_1 0 ζ_0									0
t_2	τ_1 0 ζ_0								τ_2 1 ζ_1	1
t_3	τ_1 0 ζ_0	τ_2 2 ζ_1							τ_3 3 ζ_2	2
t_4	τ_1 0 ζ_0	τ_2 4 ζ_1						τ_3 6 ζ_2	τ_4 7 ζ_3	3
t_5	τ_1 0 ζ_0	τ_2 8 ζ_1					τ_3 12 ζ_2	τ_4 14 ζ_3	τ_5 15 ζ_4	4
t_6	τ_1 0 ζ_0	τ_2 16 ζ_1	τ_3 24 ζ_2				τ_4 28 ζ_3	τ_5 30 ζ_4	τ_6 31 ζ_5	5
t_7	τ_1 0 ζ_0	τ_2 32 ζ_1	τ_3 48 ζ_2	τ_4 56 ζ_3			τ_5 60 ζ_4	τ_6 62 ζ_5	τ_7 63 ζ_6	6
t_8	τ_1 0 ζ_0	τ_2 64 ζ_1	τ_3 96 ζ_2	τ_4 112 ζ_3		τ_5 120 ζ_4	τ_6 124 ζ_5	τ_7 126 ζ_6	τ_8 127 ζ_7	7
t_9	$\tau_1=\sigma_1$ 0 ζ_0	$\tau_2=\sigma_2$ 128 ζ_1	$\tau_3=\sigma_3$ 192 ζ_2	$\tau_4=\sigma_4$ 254 ζ_3	$\tau_5=\sigma_5$ 240 ζ_4	$\tau_6=\sigma_6$ 248 ζ_5	$\tau_7=\sigma_7$ 252 ζ_6	$\tau_8=\sigma_8$ 254 ζ_7	$\tau_9=\sigma_9$ 255 ζ_8	8

(Der stern * steht für eine – möglicherweise leere – menge von strichen, die der hierarchie entlang den senkrechten kanten entsprechen, wie dies fig. 10 zeigt)



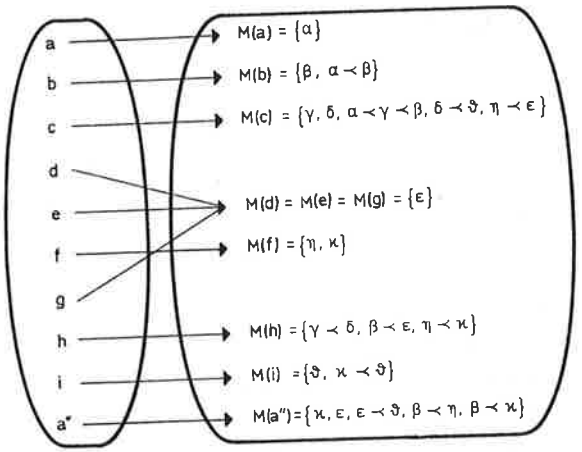
Figur 13
Allgemeines beispiel für substrukturen

"omnipräsenz der ordnungsrelation in der linguistik" sprechen (hierzu ausführlich und kritisch: Grunig 1980-P:1ff), von einer omnipräsenz, unter der man zuweilen sogar die etablierung oder nicht-etablierung von strukturen und einheiten etwa von konstituenten abhängig macht von der begehrten ikonizität. Daß ich darauf verzichte, beschreibende strukturen ikonisch zu gestalten, entspricht einem vielleicht etwas konservativen standpunkt, dem gemäß man möchte sagen können, eine abstrakte linguistische form habe diese oder jene – an der abstraktheit der form gemessen: konkrete – manifestation (hierzu Hjelmslev 1943:94). Dies entspricht einem standpunkt, nach dem man auch sagen kann: "Für die erkenntnis der internen struktur des satzes ist es unwesentlich, welche wortreihenfolge normal ist und welche von der norm abweicht. Deshalb muß die interne struktur des satzes auf der schicht der globalen symbole als invariante für eine beliebige wortrei-

henfolge aufgefaßt werden, sei es eine normale oder eine von der norm abweichende wortfolge." (Šaumjan 1965:93f). – Auch mit dependenzsyntaxen ist grundsätzlich – wenn man die einführung des begriffs der 'majorisierung' (Beleckij & Grigorjan & Zaslavskij 1963:73ff) oder des der 'projektivität' (s. Marcus 1965:181ff) nicht als wesentlichen bestandteil solcher syntaxen auffaßt – die ikonizität im genannten sinne nicht gegeben (s. auch Nebeský 1962:104ff).

Es ist nicht abwegig, die manifestationsrelation als eine funktion aufzufassen, und zwar als surjektion von der menge der formen auf die menge der manifestationen. Dies entspräche jedenfalls der annahme einer funktion $\mathcal{A}$ von der menge der seichtstrukturen auf die menge der oberflächenstrukturen bei Cresswell (1973:210ff) und der etablierung einer abbildung C2 von der menge der PREAL MANIF auf die menge der MANIF bei Grunig (1980-A:229).

Nimmt man als formen elementare oder komplexe substrukturen von zu bestimmender art (hierzu Grunig 1980-P:23ff, ausführlicher Grunig 1981:137ff) und seien die in fig. 13 umrahmten stücke solche substrukturen, so läßt sich für die manifestationsfunktion beispielsweise fig. 14 angeben:



Figur 14
Beispiel für die manifestationsfunktion

a	b	c	d	e	f	g	h	i	a'	
α	β	γ	ε	ε	η	ε		ϑ	κ	A
		δ			κ				ε	A
	α≺β	α≺γ≺β					γ≺δ	κ≺ϑ	ε≺ϑ	B
		δ≺ϑ					β≺ε		β≺η	B
		η≺ε					η≺κ		β≺κ	B

Tabelle 2
Die manifestationen für die struktur in fig. 11
A: morphologische segmente
B: morphotaktische beziehungen

Die manifestationen sind – wie in fig. 14 angedeutet und in tab. 2 verzeichnet – mengen von repräsentationen morphologischer einschließlich morphotaktischer eigenschaften der zu beschreibenden ausdrucksstücke. Die eigenschaften, die in fig. 14 als $\alpha, \beta, \gamma \ldots$ notiert sind, sollen als morphologische segmente verstanden werden, jene mit dem zeichen $\prec$ (für: 'vor') notierten geben reihenfolgebeziehungen (also morphotaktische beziehungen) an. Die manifestationen der substrukturen, wie sie in tab. 2 verzeichnet sind, ergeben zusammen als manifestation der struktur in fig. 13 beispielsweise (14) bis (16), nicht aber (17):

(14) $\alpha \ \gamma \ \beta \ \eta \ \kappa \ \delta \ \varepsilon \ \vartheta$
(15) $\alpha \ \gamma \ \beta \ \delta \ \eta \ \varepsilon \ \kappa \ \vartheta$
(16) $\alpha \ \gamma \ \delta \ \beta \ \eta \ \varepsilon \ \kappa \ \vartheta$
(17) $^{*}\alpha \ \gamma \ \varepsilon \ \beta \ \eta \ \delta \ \kappa \ \vartheta$

Wie fig. 14 und tab. 2 zeigen, sollen die fälle (i) und (ii) nicht ausgeschlossen sein:

(i) Ein und dieselbe manifestation kann aus mehreren elementen bestehen, z. b. $M(f), M(a''), M(b)$.

(ii) Zwei manifestationen können gemeinsame elemente aufweisen, z. b.: sowohl in $M(a'')$ aus auch in $M(f)$ kommt κ vor.

Das motiv dafür, (i) und (ii) zuzulassen, gibt der umstand, daß sowohl mit diskoninuierlichen manifestationen gerechnet werden muß als auch mit solchen, die in der literatur bisweilen als amalgame (oder auch als portemanteau-morpheme) bezeichnet wurden. Als beispiele für diskontinuierliche manifestationen seien angeführt:

(18)	(a)	hebräisch:	g-n-v	*in* ganav ('er, sie hat gestohlen')
	(b)	deutsch:	ge-en, ge-t	*in* gesalzen, geliebt
	(c)	makedonisch:	go - ot	*in* go - budilnikot ('den wecker')

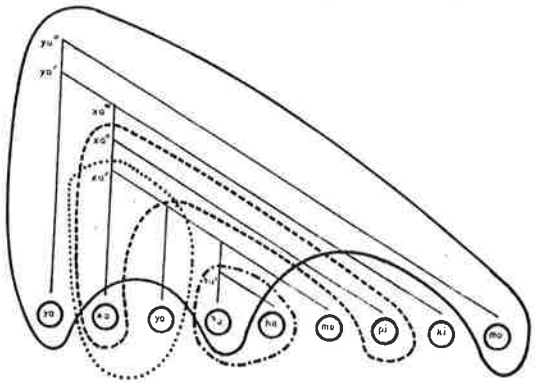
Beispiele für amalgame sind:

(19)	(a)	hebräisch:	tem	('2. ps. & pl. & m.')
	(b)	deutsch:	ging	('geh & 1. ps. & pl. & prät.')
	(c)	französich:	aulx	('knoblauch & pl.')

Manifestationen, in denen amalgamierung und diskontinuität simultan auftreten, liegen vor in (20):

(20)	(a)	hebräisch:	ganavtem	('ihr [m.] habt gestohlen')
	(b)	deutsch:	gegangen	

Die figuren 13 und 14 zusammen mit tab. 2 sind willkürlich erfundene konfigurationen ohne einen bezug auf vorliegende natürlichsprachliche ausdrücke. Konkrete beispiele für die manifestationsfunktion gibt fig. 16 auf der grundlage der struktur in fig. 10 für den hebräischen ausdruck (8), wobei die in fig. 15 vermerkten substrukturen vorausgesetzt werden. Diese beispiele sind stark vereinfacht, zeigen aber wenigstens



Figur 15
Einige substrukturen als urbilder unter der manifestationsfunktion

einige plausible bilder von elementen der menge der substrukturen von fig. 15 unter der manifestationsfunktion.

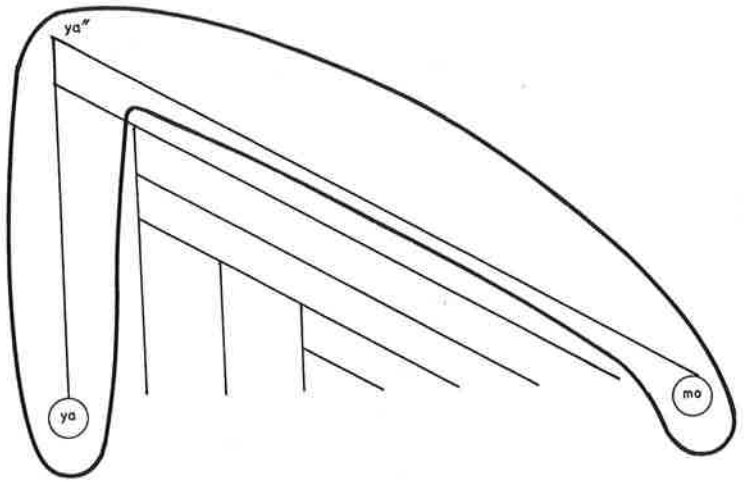
$$
\begin{aligned}
\text{ya} &\longrightarrow /\text{ja'da}/ \\
\text{ya}'' &\longrightarrow /\text{ja'da} \ldots \text{o} \ldots \text{mo'} \int \varepsilon/ \\
\text{xa} &\longrightarrow /\text{xa'xam}/ \\
\text{xa}' &\longrightarrow /\text{xa'xam}/ \prec /\text{jo'ter}/ \\
\text{ya} &\longrightarrow /\text{jo'ter}/ \\
\text{xa}'' &\longrightarrow \begin{cases} /\text{pin'xas xa'xam}/ \\ (*/\text{pin'xas xaxa'ma}/) \\ /\text{pin'xas}/ \prec /\text{xa'xam}/ \end{cases} \\
\text{pi} &\longrightarrow /\text{pin'xas}/ \\
\text{?a} &\longrightarrow \begin{cases} /\text{?axoto haktana}/) \\ (*/\text{?axoto hakatan}/) \end{cases} \\
\text{ha} &\longrightarrow /\text{k-t-n}/
\end{aligned}
$$

Figur 16
Beispiele für bilder unter der manifestationsfunktion

Vereinfacht ist fig. 16 in mindestens zweierlei hinsicht:

(a) Anstatt etwa für yada, ʔaxoto atomare substrukturen in fig. 15 vorzusehen, nämlich ya und ʔa, wäre eine feinere syntaktische analyse angemessener.

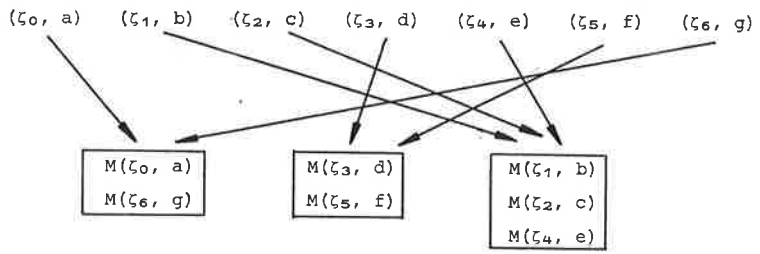
(b) Es sind für eine genauere beschreibung der manifestierenden segmente und beziehungen in manchen fällen mehrere (perforierte) substrukturen notwendig (und formal möglich), die von ein und demselben etikettierten knoten dominiert werden. So müßte etwa für /mo' ∫ ɛ/ ≺ /ja'da/ (oder für die alternative manifestation /ja'da/ ≺ /mo' ∫ ɛ/ in dem ausdruck (9)) die in fig. 17 markierte substruktur vorgesehen werden, die sich von ya″ in fig. 15 unterscheidet:



Figur 17
Eine weitere substruktur, die von dem mit ya″ etikettierten knoten dominiert wird

Für das folgende soll ausgeschlossen werden, daß es mehr als eine substruktur für jeden dominierenden etikettierten knoten gibt. Es soll

außerdem vereinbart werden, daß im falle diskontinuierlicher manifestation (fall (i)) nur das am weitesten links stehende segment in betracht gezogen werden soll. In fällen der amalgamierung (fall (ii)) soll jedes manifestierte paar $(\zeta_{i_j}, \tau_{m_j})$ vor jedem paar $(\zeta_{i'_j}, \tau_{k'_j})$ den vorrang haben gdw $\mathrm{bas}(\zeta_{i_j}, \tau_{k_j}) > \mathrm{bas}(\zeta_{i'_j}, \tau_{k'_j})$.



Figur 18
Amalgamierung (fall ii)

Eine der konsequenzen ist, daß zwischen amalgamierenden und nichtamalgamierenden manifestationen nicht unterschieden werden kann. Dieses manko nehme ich hier in kauf. Um den amalgamierungsfall $M((\zeta_6,g),(\zeta_0,a)), M((\zeta_5,f),(\zeta_3,d)), M((\zeta_4,e),(\zeta_2,c),(\zeta_1,b))$ vom nichtamalgamierungsfall $M((\zeta_6,g),(\zeta_0,a),(\zeta_5,f),(\zeta_3,d),(\zeta_4,e),(\zeta_2,c),(\zeta_1,b))$ zu unterscheiden, könnte man freilich so etwas wie "morphologische kompaktheit" in anlehnung an Lehfeldts (1978:44ff) begriff der "Verbundenheit" einführen. Einen solchen vorschlag unterbreite ich hier nicht. Die übrigen konsequenzen meiner vereinfachung überblicke ich noch nicht, hoffe aber, daß sie unschädlich sind.

Die grundidee des folgenden ist die, eine methode zu entwickeln, mit der man, ausgehend von einer rechtsnormierten struktur sowie den nach (10) ermittelten basiswerten, den eventuellen "linksdrall" der einzelnen manifestierenden segmente gegenüber den jeweils entsprechenden basiswerten berechnen kann.

Es ist unmittelbar einleuchtend, daß es nur die morphologischen segmente sein können, die hierbei interessieren, nicht die morphotaktischen beziehungen zwischen den segmenten (s. fig. 13). Die unter-

scheidung zwischen morphologischen segmenten und morphotaktischen beziehungen – darauf sei geachtet – ist nicht zu verwechseln mit der dichotomie von "segmentalen" und "suprasegmentalen" eigenschaften; "suprasegmentale" manifestationsmittel (akzente, intonation) stehen sowohl untereinander als auch zu "segmentalen" in morphotaktischen beziehungen.

Um dieses ziel zu erreichen, wird nunmehr über der finalkette jeder teilstruktur t_j von (s, φ) ein "manifestationsfächer" aufgespannt, der für jedes paar $(\zeta_{i_j}, \tau_{k_j})$ ein r-tupel potentieller extrem-linker manifestationsplätze $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_r)$ angibt (mit $r = \text{bas}(\zeta_{i_j}, \tau_{k_j})$).

Dies geschieht in zwei Schritten:

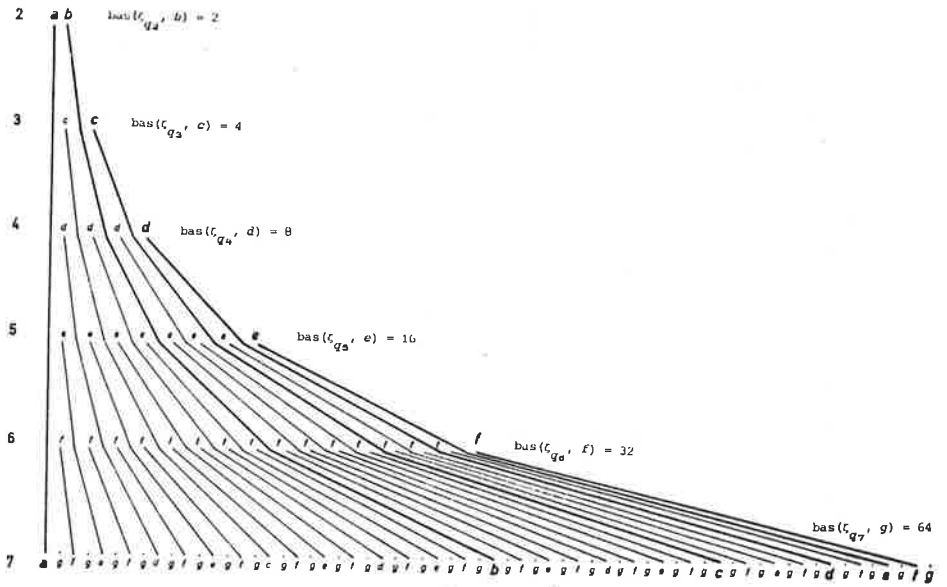
(a) für jedes i, k, j und $\mathbf{p}_w$ ist $\mathbf{p}_w \in \mathbf{P}$ (mit $w = \text{bas}(\zeta_{i_j}, \tau_{k_j})$);

(b) wenn $\alpha \in \mathbf{P}$, dann auch $\alpha - 2^{m_j - i_j} \in \mathbf{P}$, sofern $\text{bas}(\zeta_{q_j}, \tau_{m_j}) \geq \alpha \geq 2^{m_j - i_j} - \text{bas}(\zeta_{q_j}, \tau_{m_j})$.

Als beispiel sei eine kette $\psi_7 = \tau_1 \hat{} \ldots \hat{} \tau_7 = a\hat{}b\hat{}c\hat{}d\hat{}e\hat{}f\hat{}g$ in einer teilstruktur t_j betrachtet. Zeile 7 in fig. 19 gibt die rechte seite des manifestationsfächers für ψ_7 wieder. Wie ersichtlich, findet sich in fig. 19 die rechte hälfte des manifestationsfächers für jede der folgenden ketten: $a\hat{}b$ (in zeile 2 mit $m = 2$) $a\hat{}b\hat{}c$ (in zeile 3 mit $m = 3$). $a\hat{}b\hat{}c\hat{}d$ (in zeile 4 mit $m = 4$), $a\hat{}b\hat{}c\hat{}d\hat{}e$ (in zeile 5 mit $m = 5$) und $a\hat{}b\hat{}c\hat{}d\hat{}e\hat{}f$ (in zeile 6 mit $m = 6$). Jede zeile in fig. 1 zusammen mit ihrer spiegelung um die null-achse in die negativen zahlen – jeder manifestationsfächer also – stellt ein bezugssystem dar, mit dem jede anordnung, in der die manifestation der folge von paaren $(\zeta_{i_j}, \tau_{k_j})$ in jeder teilstruktur von der manifestierten folge – der kette ψ_j also – abweicht, beschrieben werden kann, und zwar so, daß die abweichungen numerisch ausgedrückt werden. Wie dies veranstaltet werden kann, wird im abschnitt 7 vorgeführt.

In fig. 19 sind die potentiellen extrem-linken manifestationsplätze durch fette normalgroße minuskeln oder durch sie vertretende fette kanten angegeben. Die kleinen mageren minuskeln sowie die sie vertretenden mageren kanten sind den jeweiligen potentiellen extrem-linken linkswerten, die sich nach (b) errechnen, zugeordnet.

7. Der nächste schritt besteht darin, die terminale kette φ samt rechtsnormierter strukturbeschreibung s auf die p manifestationsfächer derart zu beziehen, daß für jede teilstruktur t_j und in t_j für jedes



Figur 19
Die rechte seite einiger manifestationsfächer

paar $(\zeta_{i_j}, \tau_{k_j})$ aus der menge **P** eine teilmenge **M** ausgewählt wird, die die realen extrem-linken manifestationsplätze für $(\zeta_{i_j}, \tau_{k_j})$ sind. Grundsätzlich muß **M** nicht eine einermenge sein, denn es sollen diskontinuierliche manifestationen nicht ausgeschlossen werden. Wegen der vereinbarung (i) in abschnitt 6 wird hier aber **M** durchweg als einermenge betrachtet. – Seien $(\zeta_{i_j}, \tau_{k_j})$ und $(\zeta_{i'_j}, \tau_{k'_j})$ zwei voneinander verschiedene paare, so muß es grundsätzlich nicht der fall sein, daß $\mathbf{M}(M(\zeta_{i_j}, \tau_{k_j})) \cap \mathbf{M}(M(\zeta_{i'_j}, \tau_{k'_j})) = \emptyset$. Wegen der vereinbarung (ii) in abschnitt 6 soll dies aber – vor allem auch im zusammenspiel mit vereinbarung (i) – gelten.

Um nun ein verfahren zu entwickeln, mit dem die einzelnen realen manifestationsplätze bestimmt werden können, seien einige vorüberlegungen angestellt. Dabei beziehe ich mich auf zeile 7 in fig. 19. Hier hat g gegenüber allen anderen symbolen (bezogen auf die jeweiligen (final)knoten in ein und derselben teilstruktur t_7) den höchsten basiswert. Steht nun (das weitest-linke stück von) $M(g)$ nach den (weitest-linken stücken der) manifestationen aller anderen finalsymbole von t_7, so trägt $M(g)$ selbst offensichtlich nichts zu einem eventuellen "linksdrall" von ψ_7 bei. Steht $M(g)$ hingegen vor den (weitest-linken stücken der) manifestationen aller anderen finalsymbole von t_7, so hat ψ_7 auf jeden fall einen "linksdrall", und $M(g)$ trägt zu diesem "linksdrall" (a) das maximum dessen bei, was es zum "gesamtlinksdrall" von ψ_7 beitragen kann, und (b) dabei offenbar mehr, als die manifestation irgendeines anderen finalsymbols von t_7 für sich genommen beitragen kann.

Außerdem kann man sagen, daß ψ_7 gar keinen "linksdrall" aufweist, wenn die reihenfolge der manifestierten symbole in ψ_7 und die (der weitest-linken stücke) ihrer manifestationen übereinstimmen. Bei spiegelbildlicher reihenfolge der (weitest-linken stücke der) manifestationen hingegen liegt ein maximum an "linksdrall" vor.

Unter berücksichtigung der vereinbarungen (i) und (ii), die in abschnitt 6 getroffen wurden, kann nunmehr eine bijektion ω der folge $1, \ldots, q$ von ganzen positiven zahlen auf die folge Ψ_j der manifestation der etikettierten finalknoten von t_j angenommen werden, durch die der manifestation jedes paares $(\zeta_{i_j}, \tau_{k_j})$ ein ordnungsindex ω zugeordnet wird, z.B. – bezogen auf t_9 in fig. 10 –:

(21)
$$\begin{aligned}
\omega(M(\zeta_{0_p}, \text{ya})) &= 2 \\
\omega(M(\zeta_{1_p}, \text{xa})) &= 5 \\
\omega(M(\zeta_{2_p}, \text{yo})) &= 6 \\
\omega(M(\zeta_{3_p}, ?\text{a})) &= 8 \\
\omega(M(\zeta_{4_p}, \text{ha})) &= 9 \\
\omega(M(\zeta_{5_p}, \text{me})) &= 7 \\
\omega(M(\zeta_{6_p}, \text{pi})) &= 4 \\
\omega(M(\zeta_{7_p}, \text{ki})) &= 3 \\
\omega(M(\zeta_{8_p}, \text{mo})) &= 1
\end{aligned}$$

An stelle von $\omega(M(\zeta_{0_p}, \text{ya}))$ usw. schreibe ich im folgenden $(M(\zeta_{o_p}, \text{ya}))_\omega$ usw.

Dabei soll der ordnungsindex für jedes paar $(\zeta_{i_j}, \tau_{k_j})$ angeben, an wievielter stelle in der manifestation von Ψ_j die manifestation $M(\zeta_{i_j}, \tau_{k_j})$ vorkommt. Wegen der werte in (21) ergibt sich für Ψ_9 in fig. 10 die folge (22) von manifestationen in Ψ_9 :

(22)
$$\begin{aligned}
\Psi_9 = &(M(\zeta_{8_9}, \text{mo}))_1, (M(\zeta_{0_9}, \text{ya}))_2, (M(\zeta_{7_9}, \text{ki}))_3, \ldots \\
&\ldots (M(\zeta_{6_9}, \text{pi}))_4, (M(\zeta_{1_9}, \text{xa}))_5, (M(\zeta_{2_9}, \text{yo}))_6, \ldots \\
&\ldots (M(\zeta_{5_9}, \text{me}))_7, (M(\zeta_{3_9}, ?a))_8, (M(\zeta_{4_9}, ha))_9
\end{aligned}$$

Die manifestationsfolge $M(\Psi_j)$ soll nunmehr so auf den manifestationsfächer projiziert werden, daß sich jür jedes $K_\omega \in \Psi_j (K_\omega := (M(\zeta_{i_j}, \tau_{k_j}))_\omega$; mit $1 \leq \omega \leq j$)) genau ein realer extrem-linker manifestationsplatz $\mathbf{m}_{K_\omega} \in \mathbf{M}_{K_\omega} \subset \mathbf{P}_{K_\omega}$ ergibt. Mehrere wege können dafür beschritten werden, etwa die beiden folgenden als die wohl nächstliegenden:

(i) $\mathbf{m}_{K_\omega}$ ist das kleinste **p** für K_ω, das größer ist als das seines vorgängers $\text{Vorg}(K_\omega) = K_{\omega-1}$; $(\text{Vorg}(K_1) = \emptyset)$.

(ii) $\mathbf{m}_{K_\omega}$ ist das größte **p** für K_ω, das kleiner ist als das seines nachfolgers $\text{Nachf}(K_\omega) = K_{\omega+1}$; $(\text{Nachf}(K_q) = \emptyset)$.

Wie (23) für (22) zeigt, führen die beiden wege im allgemeinen fall zu unterschiedlichen resultaten:

(23)		nach Verfahren (i)	nach Verfahren (ii)
	$\mathbf{m}((M(\zeta_{8_9}, \text{mo})))_0 =$	-255	-1
	$\mathbf{m}((M(\zeta_{0_9}, \text{ya})))_1 =$	0	0
	$\mathbf{m}((M(\zeta_{7_9}, \text{ki})))_2 =$	2	122
	$\mathbf{m}((M(\zeta_{6_9}, \text{pi})))_3 =$	4	124
	$\mathbf{m}((M(\zeta_{1_9}, \text{xa})))_4 =$	12	128
	$\mathbf{m}((M(\zeta_{2_9}, \text{yo})))_5 =$	64	192
	$\mathbf{m}((M(\zeta_{5_9}, \text{me})))_6 =$	72	216
	$\mathbf{m}((M(\zeta_{3_9}, \text{?a})))_7 =$	96	224
	$\mathbf{m}((M(\zeta_{4_9}, \text{ha})))_8 =$	112	240

Für die spezialfälle, daß die reihenfolgebeziehungen von $M(\Psi_j)$ erhalten bleiben oder daß sie durchweg spiegelverkehrt werden, konvergieren die beiden methoden, wie die hypothetischen beispiele (24) bzw. (25) mit bezug auf (22) zeigen:

(24)		nach Verfahren (i)	nach Verfahren (ii)
	$\mathbf{m}((M(\zeta_{0_9}, \text{ya})))_0 =$	0	0
	$\mathbf{m}((M(\zeta_{1_9}, \text{xa})))_1 =$	128	128
	$\mathbf{m}((M(\zeta_{2_9}, \text{yo})))_2 =$	192	192
	$\mathbf{m}((M(\zeta_{3_9}, \text{?a})))_3 =$	224	224
	$\mathbf{m}((M(\zeta_{4_9}, \text{ha})))_4 =$	240	240
	$\mathbf{m}((M(\zeta_{5_9}, \text{me})))_5 =$	248	248
	$\mathbf{m}((M(\zeta_{6_9}, \text{pi})))_6 =$	252	252
	$\mathbf{m}((M(\zeta_{7_9}, \text{ki})))_7 =$	254	254
	$\mathbf{m}((M(\zeta_{8_9}, \text{mo})))_8 =$	255	255

(25)		nach Verfahren (i)	nach Verfahren (ii)
	$\mathbf{m}((M(\zeta_{8_9}, \text{mo})))_0 =$	-255	-255
	$\mathbf{m}((M(\zeta_{7_9}, \text{ki})))_1 =$	-254	-254
	$\mathbf{m}((M(\zeta_{6_9}, \text{pi})))_2 =$	-252	-252
	$\mathbf{m}((M(\zeta_{5_9}, \text{me})))_3 =$	-248	-248
	$\mathbf{m}((M(\zeta_{4_9}, \text{ha})))_4 =$	-240	-240
	$\mathbf{m}((M(\zeta_{3_9}, \text{?a})))_3 =$	-224	-224
	$\mathbf{m}((M(\zeta_{2_9}, \text{yo})))_2 =$	-192	-192
	$\mathbf{m}((M(\zeta_{1_9}, \text{xa})))_1 =$	-128	-128
	$\mathbf{m}((M(\zeta_{0_9}, \text{ya})))_0 =$	0	0

Sieht man von den beiden spezialfällen ab, so erkennt man leicht, daa das verfahren (i) einen möglichen "linksdrall" deutlicher zu ausdruck bringen kann als das verfahren (ii). Ich wende daher in folgenden das verfahren (i) an und lasse die frage offen, ob das verfahren (ii) möglicherweise vorteile gegenüber (i) mit sich bringt. – Freilich bietet sich auch hier wieder die möglichkeit, die werte, die sich nach (ii) ergeben, durch ein geeignetes verfahren mit denen zu verrechnen, die sich nach (i) ergeben. Ich verfolge diese möglichkeit für den weiteren verlauf meiner darlegungen nicht.

8. Der nächste schritt besteht darin, für jede teilstruktur t_j die werte der realen extrem-linken manifestationsplätze auf die zugehörigen basiswerte zu beziehen, und zwar derart, daß mit der formel (26) eine sog. linksdistanz LDist für jedes paar $(\zeta_{i_j}, \tau_{k_j})$ berechnet werden kann:

$$\text{(26)} \qquad \text{LDist}(\zeta_{i_j}, \tau_{k_j}) = \text{bas}(\zeta_{i_j}, \tau_{k_j}) - \mathbf{m}((M(\zeta_{i_j}, \tau_{k_j}))_\omega)$$

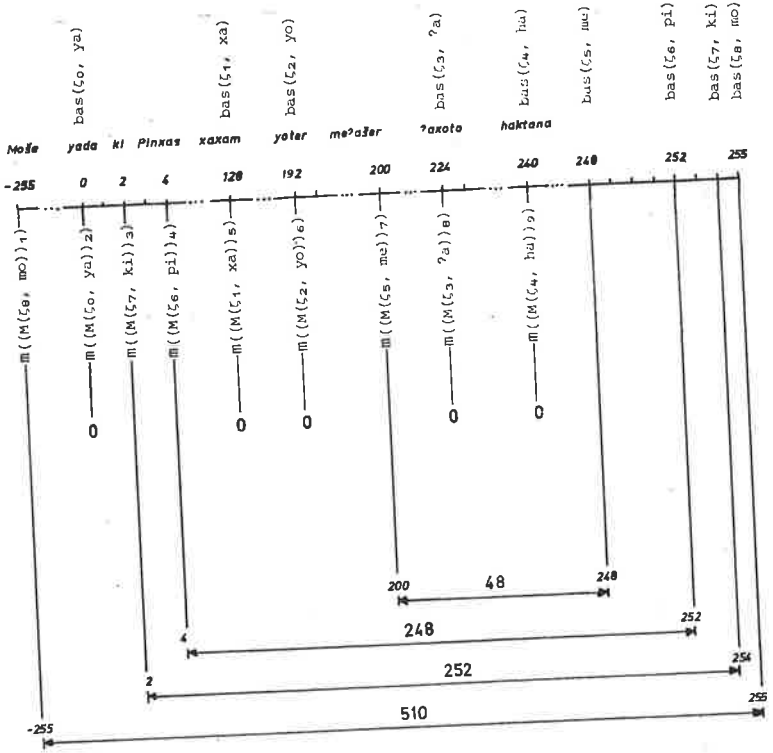
In fig. 20 werden die linksdistanzen für jeden etikettierten terminalknoten in der struktur von fig. 10 gezeigt:

Für die struktur in fig. 10 – möge sie κ heißen – ist $p = 9$. Daher ergibt sich als wert für die linksorientiertheit ihrer manifestation:

$$\text{(27)} \qquad \text{LOr}(M(\kappa)) = \frac{4.0061}{9} = 0.4451$$

Die brauchbarkeit des maßes der linksorientiertheit hängt – wie die jedes vernünftigen sprachtypologischen maßes – von der güte der linguistischen bearbeitung der daten ab, auf die das maß angewendet werden soll. Trotz der aktivitäten, die sich besonders in den letzten dezennien auf syntaktischem gebiete entfaltet haben, ist es meines wissens für keine natürliche sprache gelungen, jedem beliebigen ausdruck in einem (nicht künstlich erzeugten, sondern belegten) korpus eine detaillierte und umfassende strukturbeschreibung zuzuordnen. Beim derzeitigen stand der dinge muß man sich wohl oder übel mit groben strukturzuweisungen begnügen (s. Thümmel 1980:153ff), bei denen stücke eines

gegebenen ausdrucks unanalysiert bleiben, obwohl eine feinere analyse angemessen wäre. Hinzu kommt, daß der feinheitsgrad der strukturzuweisungen bei den verschiedenen sprachen, die typologisch zueinander in bezug gesetzt werden sollen, recht unterschiedlich ist. Kompensiert wird dieses die gleichbeurteilung der einzelnen sprachen störende manko teilweise freilich dadurch, daß die zerlegung in teilstrukturen $t_1, \ldots, t_p$ die berechnung der linksgeneigtheit für gröbere und weniger grobe analysen systematisch nicht nur möglich macht, sondern fordert.



Figur 20
Die linksdistanzen für jedes paar $(\zeta_{i_j}, \tau_{k_j})$
in t_9 von fig. 10

Für t_9 sowie für alle anderen teilstrukturen von fig. 20 werden die linksdistanzen in tab. 3 gegeben.

Tabelle 3
Die linksdistanzen für alle Teilstrukturen von fig. 10

				ω	**m**	LDist
t_1	bas(ζ_0, ya$''$)	=	0	1	0	0
t_2	bas(ζ_0, ya$'$)	=	0	2	0	0
	bas(ζ_1, mo)	=	0	1	-1	2
t_3	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa$'''$)	=	2	3	2	0
	bas(ζ_2, mo)	=	3	1	-3	6
t_4	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa$''$)	=	4	4	4	0
	bas(ζ_2, ki)	=	6	3	2	4
	bas(ζ_3, mo)	=	7	1	-7	14
t_5	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa$'$)	=	8	5	8	0
	bas(ζ_2, pi)	=	12	4	4	8
	bas(ζ_3, ki)	=	14	5	2	12
	bas(ζ_4, mo)	=	15	1	-15	30
t_6	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa)	=	16	5	16	0
	bas(ζ_2, yo$'$)	=	24	6	24	0
	bas(ζ_3, pi)	=	28	4	4	24
	bas(ζ_4, ki)	=	30	3	2	28
	bas(ζ_5, mo)	=	31	1	-31	62
t_7	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa)	=	32	5	32	0
	bas(ζ_2, yo)	=	48	6	48	0
	bas(ζ_3, ²a$''$)	=	56	7	56	0
	bas(ζ_4, pi)	=	60	4	4	56
	bas(ζ_5, ki)	=	62	3	2	60
	bas(ζ_6, mo)	=	63	1	-63	126

t_8	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa)	=	64	5	64	0
	bas(ζ_2, yo)	=	96	6	96	0
	bas(ζ_3, ?a)	=	112	8	112	0
	bas(ζ_4, me)	=	120	7	104	16
	bas(ζ_5, pi)	=	124	4	4	120
	bas(ζ_6, ki)	=	126	3	2	124
	bas(ζ_7, mo)	=	127	1	-127	254
t_9	bas(ζ_0, ya)	=	0	2	0	0
	bas(ζ_1, xa)	=	128	5	128	0
	bas(ζ_2, yo)	=	192	6	192	0
	bas(ζ_3, ?a)	=	224	8	224	0
	bas(ζ_4, ha)	=	240	9	240	0
	bas(ζ_5, me)	=	248	7	200	48
	bas(ζ_6, pi)	=	252	4	4	248
	bas(ζ_7, ki)	=	254	3	2	252
	bas(ζ_8, mo)	=	255	1	255	510

Auf der grundlage der linksdistanzen wird alsdann für jede teilstruktur t_j die linksneigung der manifestation $M(t_j)$ von t_j errechnet, und zwar nach formel (28).

$$\text{LNeig}(M(t_j)) = \frac{\sum_{\nu=1}^{q} \text{LDist}(\zeta_{\nu_j}, \tau_{\nu+1})}{\sum_{\nu=1}^{q} 2\text{bas}(\zeta_{\nu_j}, \tau_{\nu+1})} \tag{28}$$

In tab. 4 werden für t_1 bis t_9 von fig. 10 die werte angegeben, die sich nach (28) für die jeweilige linksneigung errechnen:

Tabelle 4

Die linksneigungen für die teilstrukturen in fig. 10

	$\text{LDist}(\zeta_{\nu_j}, \tau_{\nu+1_j})$	$2\text{bas}(\zeta_{\nu_j}, \tau_{\nu+1_j})$	LNeig
t_1	0	0	0.0000
t_2	2 0	1 2 ×2	
	$\sum$ 2	$\sum$ 2	1.0000

t_3	6 0 0	3 2 0 ×2	
	$\sum$ 6	$\sum$ 10	0.6000
t_4	14 4 0 0	7 6 4 0 ×2	
	$\sum$ 18	$\sum$ 34	0.5294
t_5	30 12 8 0 0	15 14 12 8 0 ×2	
	$\sum$ 42	$\sum$ 98	0.4286
t_6	62 28 24 0 0 0	31 30 28 24 16 0 ×2	
	$\sum$ 114	$\sum$ 258	0.4419
t_7	126 60 56 0 0 0 0	63 62 60 56 48 32 0 ×2	
	$\sum$ 242	$\sum$ 642	0.3769

t_8	254 124 120 16 0 0 0 0	127 126 124 120 112 96 64 0	×2	
	$\sum$ 514	$\sum$ 1.538		0.3342
t_9	510 252 248 48 0 0 0 0 0 0	255 254 252 248 240 224 192 128 0 0	×2	
	$\sum$ 1.058	$\sum$ 3.586		0.2951

Es ist augenfälllig, daß die werte für die linksgeneigtheit zwischen 0 und 1 liegen.

Für die manifestation $M(\varphi, s)$ der kette φ mit der rechtsnormierten strukturbeschreibung s läßt sich dann unter bezug auf die linksgeneigtheit der manifestation der einzelnen teilstrukturen die linksorientiertheit $\mathrm{LOr}(M(\varphi, s))$ berechnen, und zwar nach formel (29):

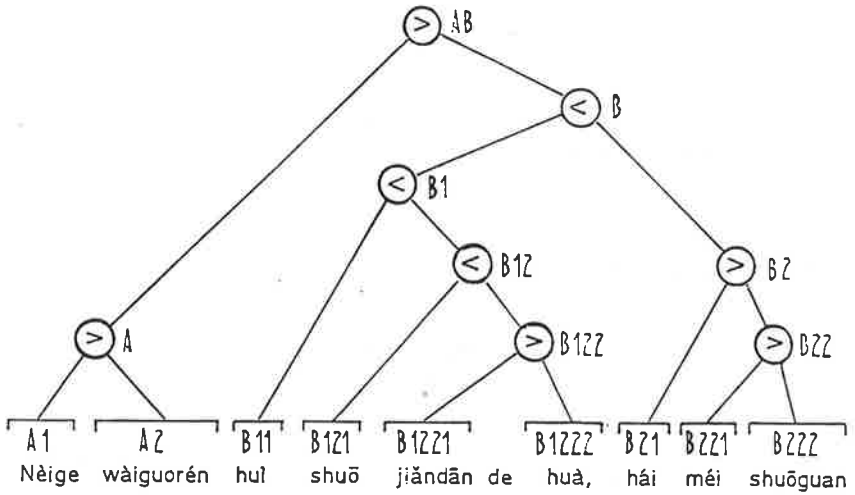
$$\mathrm{LOr}(M(\varphi,s)) = \frac{\sum_{j=1}^{q} \mathrm{LNeig}(M(t_j))}{p} \tag{29}$$

9. Zur typologischen charakterisierung von sprachen kann man auf der grundlage des maßes der linksorientiertheit ein maß der mittleren linksorientiertheit konstruieren, das für jede sprache auf eine geeignete menge K von manifestationen von (φ, s)-paaren (von strukturierten belegen etwa) angewandt wird. Ein solches maß – MLOr – kann mit formel (30) berechnet werden.

$$\mathrm{MLOr} = \frac{1}{|K|} \sum_{\nu=1}^{|K|} \mathrm{LOr}(M(\kappa_\nu)) \tag{30}$$

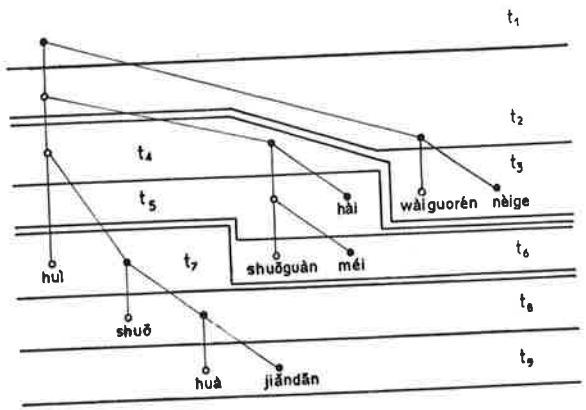
Anstatt dieses maß hier auf entsprechende ausdrucksmengen verschiedener sprachen anzuwenden, beschränke ich mich darauf, das maß der linksorientiertheit zusätzlich an hand eines chinesischen ausdrucks sowie an hand einer strukturbeschreibung, die ihm Henne & Rongen & Hansen (1977:68) zugeordnet haben, zu illustrieren. Es handelt sich um den ausdruck (31) und die strukturbeschreibung in fig. 21.

(31) Nèige$_1$ wàiguórén$_2$ huì$_3$ shuō$_4$ jiǎndān$_5$ de$_6$ huà$_7$, hái$_8$ méi$_9$ shuō$_{10}$ guàn$_{11}$.
'Dieser$_1$ ausländer$_2$ da$_1$ kann$_3$ einfache$_{5,6}$ wörter$_7$ sprechen$_4$, aber$_8$ er$_{1,2}$ spricht$_{10}$ noch$_8$ nicht$_9$ flüssig$_{11}$.'



Figur 21
Konstituentenstruktur für (30)

Mit den symbolen > und < notieren Henne & Rongen & Hansen die kopf-/modifikator-unterscheidung: die spitze weist auf den jeweiligen kopf. Wenn ich statt des symbols ∨ (für "koordination") das symbol < einsetze, läßt sich die strukturbeschreibung in fig. 21 rechs-normieren, wie in fig. 22 gezeigt wird.



Figur 22
Rechtsnormierte Struktur (mit Teilstrukturen) für (30)

In tab. 5 werden für jede teilstruktur die linksdistanzen angegeben. Die linksneigungen werden in tab. 6 aufgeführt.

Tabelle 5
Die linksdistanzen für die Teilstrukturen in fig. 22

				ω	m	LDist
t_1	bas(ζ_0, huì‴)	=	0	1	0	0
t_2	bas(ζ_0, huì″)	=	0	2	0	0
	bas(ζ_1, wàiguóren′)	=	1	1	-1	1

t_3	bas(ζ_0, huì′)	=	0	3	0	0
	bas(ζ_1, wàiguóren)	=	2	2	-2	4
	bas(ζ_2, nèige)	=	3	1	-3	6
t_4	bas(ζ_0, huì″)	=	0	3	0	0
	bas(ζ_1, shuōguàn″)	=	4	4	4	0
	bas(ζ_2, wàiguóren)	=	6	2	-6	12
	bas(ζ_3, n èige)	=	7	1	-7	14
t_5	bas(ζ_0, huì′)	=	0	3	0	0
	bas(ζ_1, shuōguàn′)	=	8	6	8	0
	bas(ζ_2, hái)	=	12	4	4	8
	bas(ζ_3, wàiguóren)	=	14	2	-14	28
	bas(ζ_4, nèige)	=	15	1	-15	30
t_6	bas(ζ_0, huì′)	=	0	3	0	0
	bas(ζ_1, shuōguàn)	=	16	6	16	0
	bas(ζ_2, méi)	=	24	5	8	16
	bas(ζ_3, hái)	=	28	4	4	24
	bas(ζ_4, wàiguóren)	=	30	2	-30	60
	bas(ζ_5, nèige)	=	31	1	-31	62
t_7	bas(ζ_0, huì)	=	0	3	0	0
	bas(ζ_1, shuō)	=	32	4	32	0
	bas(ζ_2, shuōguàn)	=	48	7	48	0
	bas(ζ_3, ,mèi)	=	56	6	40	16
	bas(ζ_4, hái)	=	60	5	36	24
	bas(ζ_5, wàiguóren)	=	62	2	-62	124
	bas(ζ_6, nèige)	=	63	1	-63	126
t_8	bas(ζ_0, huì)	=	0	3	0	0
	bas(ζ_1, shuō)	=	64	4	64	0
	bas(ζ_2, huà′)	=	96	5	96	0
	bas(ζ_3, shuōguàn)	=	112	8	112	0
	bas(ζ_4, méi)	=	120	7	104	16
	bas(ζ_5, hái)	=	124	6	100	24
	bas(ζ_6, wàiguóren)	=	126	1	-126	252
	bas(ζ_7, nèige)	=	127	1	-127	254

t_9	bas(ζ_0, huì)	=	0	3	0	0
	bas(ζ_1, shuō)	=	128	4	128	0
	bas(ζ_2, huà)	=	192	6	192	0
	bas(ζ_3, jiǎndān)	=	224	5	160	64
	bas(ζ_4, shuōguàn)	=	240	9	240	0
	bas(ζ_5, méi)	=	248	8	200	48
	bas(ζ_6, hái)	=	252	7	196	56
	bas(ζ_7, wàiguóren)	=	254	2	-254	508
	bas(ζ_8, nèige)	=	255	1	-255	510

Tabelle 6
Die linksneigungen für die teilstrukturen in fig. 22

	LDist($\zeta_{\nu_j}, \tau_{\nu+1_i}$)	2bas($\zeta_{\nu_j}, \tau_{\nu+1_i}$)	LNeig
t_1	0	0	0.0000
t_2	2 0 — 2	1 0 ×2 — 2	1.0000
t_3	6 4 0 — 10	3 2 0 ×2 — 10	1.0000
t_4	14 12 0 0 — 26	7 6 4 0 ×2 — 34	0.7647

t_5	30 28 8 0 0 — 66	15 14 12 8 0 ×2 — 98	0.6735
t_6	62 60 24 16 0 0 — 114	31 30 28 24 16 0 ×2 — 258	0.6280
t_7	126 124 24 16 0 0 0 — 290	63 62 60 56 48 32 0 ×2 — 642	0.4517
t_8	254 252 24 15 0 0 0 0 — 546	127 126 124 120 112 96 64 0 ×2 — 1.538	0.3550

t_9	510 508 56 48 0 64 0 0 0 0	255 254 252 248 240 224 192 128 0 0	×2	
	1.058	3.586		0.3307

Für die struktur in fig. 22 – möge sie κ' heißen – ist $p = 9$. Daher ergibt sich als wert für die linksorientiertheit ihrer manifestation

(32) $$\mathrm{LOr}(M(\kappa')) = \frac{5.2036}{9} = 0.5782$$

10. Zum schluß soll das maß der mittleren linksorientiertheit durch die folgenden acht punkte allgemein charakterisiert werden:

(a) Anders als bei vielen S-V-O-typologien, anders auch als bei Bartsch & Vennemann (1982) in ihrer handhabung der begriffe 'operator' und 'operand' wird beim maß der mittleren linksgeneigtheit zwischen dem zu manifestierenden und der manifestation des zu manifestierenden unterschieden.

(b) Dem maß der mittleren linksorientiertheit liegt eine idee zugrunde, die der unterscheidung zwischen zentrifugaler und zentripetaler reihenfolge bei Tesnière ähnelt.

(c) Im unterschied zu der Tesnièreschen unterscheidung zwischen 'zentrifugal' und 'zentripetal', die wie vergleichbare unterscheidungen – etwa die zwischen 'konsistent präspezifizierend' und 'konsistent postspezifizierend' bei Bartsch & Vennemann (1982:34f) – grob-klassifikatorisch gehandhabt wird, ist das maß der mittleren linksorientiertheit quantitativ.

(d) Gegenüber dem Andreevschen maß der zentriertheit (in der fassung von Altmann & Lehfeldt) hat das maß der mittleren linksorientiertheit den vorteil, daß es nicht nur über terminale ketten aussagen macht, sondern – wegen der zerlegung in teilstrukturen – über terminale ketten samt der ihnen zugeordneten strukturbeschreibungen.

(e) Während die links- und die rechts-abweichungen von der größten zentralisiertheit bei Andreevs maß nicht unterschieden werden, wird beim maß der linksorientiertheit jede abweichung von der kleinsten mittleren linksorientiertheit von jeder anderen mittleren linksorientiertheit unterschieden.

(f) Durch die generelle rechtsnormierung wird ein besonderer étalon, wie ihn Uspenskij vorsieht, überflüssig.

(g) Das maß der mittleren linksorientiertheit beruht nicht auf einheiten, "die einzelsprachlich gebunden sind und deshalb beim Vergleich von Sprachen nicht ohne weiteres berücksichtigt werden können" (Altmann & Lehfeldt 1973:113). Anders das maß von Andreev, das auf dem begriff 'prädikat' oder auf dem begriff 'verb' beruht.

(h) Das maß der mittleren linksorientiertheit genügt einer von Georg von der Gabelentz (1984:6) vernünftigerweise aufgestellten forderung, der forderung nämlich, daß bei der sprachtypologie "alle grammatischen Möglichkeiten zu berücksichtigen sind".

Literatur

Altmann, G., und Lehfeldt, W. (1973). *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren.* München: Fink.

Andreev, N.D. (1967). *Statistiko-kombinatornye v teoretičeskom i prikadnom jazykovedenii.* Leningrad: Nauka.

Bar-Hillel, Y. (1953). "A quasi-arithmetical notation for syntactic description". *Language* 29, 47-58. Wiederabdruck in Bar-Hillel, Y. (1964). *Language and information. Selected essays on their theory and application.* Reading, Massachusetts, Palo Alto, London: Addison-Wesley; Jerusalem: The Jerusalem Academic Press. 61-74.

Bartsch, R., und Vennemann, T. (1982). *Grundzüge der Sprachtheorie. Eine linguistische Einführung.* Tübingen: Niemeyer.

Beleckij, M.I., und Grigorjan, V.M., und Zaslavskij, I.D. (1963). "Aksiomatičeskoe opisanie porjadka i upravlenija slov v nekotoryx tipax predloženij". In Akademija Nauk Armjanskoj SSR, Erevanskij gosudarstvennyj

universitet, Vyčislitel'nyj centr. (Hrsg). *Matematičeskie voprosy kibernetiki i vyčislitel'noj texniki.* Erevan: Izd. Akademii Nauk Armjanskoy SSR. 71-85.

Bloomfield, L. (1935). *Language.* London: Allen & Unwin.

Clément, D., und Thümmel, W. (1981). "Some remarks on scope phenomena". In Klein, W., and Levelt, W. (ed). *Crossing the boundaries in linguistics. Studies presented to Manfred Bierwisch.* Dordrecht, Boston, London: Reidel. 79-93.

Cresswell, M.J. (1973). *Logics and languages.* London: Methuen.

Dezső, L. (1982). *Studies in syntactic typology and contrastive grammar.* Budapest: Adadémiai Kiadó; The Hague: Mouton.

Gazdar, G., und Klein, E. und Pullum, G., und Sag, I. (1985). *Generalized Phrase Structure Grammar.* Oxford: Blackwell.

Gazdar, G., und Pullum, G.K. (1982). *Generalized phrase structure grammar: a theoretical synopsis.* Bloomington: Indiana University Linguistics Club. (Vervielf.).

Gabelentz, G. von der (1874/1875). "Weiteres zur vergleichenden Syntax. – Wort- und Satzstellung –". *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 8 (1874) 129-165; (1875) 300-338.

Gabelentz, G. von der (1894). "Hypologie [recte: Typologie] der Sprachen, eine neue Aufgabe der Linguistik". *Indogermanische Forschungen* 4, 1-7.

Gabelentz, H.C. von der (1861). "Über das Passivum. Eine sprachvergleichende Abhandlung". *Abhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften, phil.-hist.* Classe 3,4 449-546.

Greenberg, J.H. (1963). "Some universals of grammar with particular reference to the order of meaningful elements". In Greenberg, J.H. (ed). *Universals of language. Report of a conference held at Dobbs Ferry, New York, April 13 - 15, 1961.* Cambridge, Massachusetts: The M. I. T. Press 58-90.

Grunig, B.N. (1978). "Axiome und Theoreme einer Theorie des Artikulators als Illustration der deduktiven Methode". In Clément, D. (ed). *Empirische rechtfertigung von syntaxen. Beiträge zum Wuppertaler kolloquium vom 25. - 29. september 1978.* Bonn: Bouvier. 225-244.

Grunig, B.N. (1980). "Pour et contre l'omniprésence des relations d' ordre en linguistique de la syntaxe à la pragmatique". *DRLAV. Revue de linguistique* 22/23 1-38.

Grunig, B.N. (1981). *Structure sous-jacente: essai sur les fondements théoriques Thèse présentée devant l'Université de Paris X le 28 janvier 1978.* Lille: Atelier Reproduction des thèses, Université de Lille III. Diffusion: Paris: Champion.

Grunig, R., und Thümmel, W. (1980). "Relecture de l'article de Greenberg sur les universaux concernant l'ordre des éléments porteurs de sens". DRLAV. *Revue de linguistique* 22/23. 143-158.

Hawkins, J.A. (1983). *Word order universals.* New York [. . .]: Academic Press.

Henne, H., und Rongen, O.B., und Hansen, L.J. (1977). *A handbook on Chinese language structure.* Oslo & Bergen & Tromsö: Universitetsforlaget.

Hjelmslev, L. (1943). "Omkring sprogteoriens grundlæggelse." In *Festskrift udgivet af Københavns Universitet i Anleddning af Universitetets Aarsfest November 1943.* København: Luno. 3-113.

Keenan, E.L. (1976). "Towards a universal definition of "subject"". In Li, Ch.N. (ed). *Subject and topic.* New York & San Francisco & London: Academic Press. 303-333.

Krupa, V. (1982). "Syntactic typology and linearization". *Language* 58, 639-645.

Lehfeldt, W. (1978). *Formenbildung des russischen Verbs. Versuch einer analytisch-synthetisch-funktionellen Beschreibung der Präsens- und der Präteritumflexion.* München: Sagner.

Marcus, S. (1965). "Sur la notion de projectivité". *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* 11, 181-192.

Maurer, H. (1969). *Theoretische Grundlagen der Progammiersprachen. Theorie der Syntax.* Mannheim & Wien & Zürich: Bibliographischess Institut.

Nebesksý, L. (1962). "O jedné formalizaci větného rozboru". *Slovo a slovesnost* 23, 104-107.

Peterson, T.H. (1979). "Constraining grammars through proper generalization: multiple order grammar". In Meisel, J., und Pam, M.D. (eds). *Linear order and generative theory.* Amsterdam: Benjamins. 99-163.

Pullum, G.K. (1977). "Word order universals and grammatical relations". In Cole, P., und Sadock, J.M. (eds). *Syntax and semantics. Vol. 8: Grammatical relations.* New York & San Francisco & London: Academic Press. 249-277.

Pullum, G.K. (1981). "Languages with object before subject: a comment and a catalogue". *Linguistics* 19, 147-155.

Ramat, P. (1985). *Typologie linguistique.* Paris: Presses Universitaires de France.

Šaumjan, S.K. (1965). *Strukturnaja lingvistika.* Moskva: Nauka.

Staal, J.F. (1967). *Word order in Sanskrit and universal grammar.* Dordrecht: Reidel.

Tesnière, L. (1965). *Éléments de syntaxe structurale. Préface de Jean Fourquet. Deuxième édition revue et corrigée.* Paris: Klincksieck.

Thümmel, W. (1980). "Drei leitende gesichtspunkte beim schreiben einer syntax: grobe strukturzuweisung, beobachtungsvergleich und starke hierarchisierung. In Clément, D. (ed). *Empirische rechtfertigung von syntaxen. Beiträge zum Wuppertaler kolloquium vom 25. - 29. september 1978.* Bonn: Bouvier. 153-168.

Thümmel, W. (1986). "GPSG: Un formalisme pour une syntaxe sans transformation ni coindexation". *DRLAV. Revue de linguistique* 34/35, 287-300.

Uspenskij, B.A. (1965). *Strukturnaja tipologija jazykov.* Moskva: Nauka.

Wundt, W. (1904). *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte.* Erster Band [recte: Zweiter Band]: Die Sprache. Zweite, umgearb. Aufl. Zweiter Teil [recte: einziger teil]. Leipzig: Engelmann.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 105–119

Neue Modelltheoretische Untersuchungen im Zusammenhang mit dem Martingesetz der Abstraktionsebenen*

Rolf Hammerl
Kielce

1. Den Ausgangspunkt unserer Untersuchungen stellt das Martingesetz der Abstraktionsebenen dar (Martin 1974; Altmann, Kind 1983; Hammerl 1987), das in weiterführende Zusammenhänge gestellt werden soll. Dabei wird versucht, alle analysierten Aspekte einer mathematischen Beschreibung zu unterziehen und deren gegenseitigen Beziehungen aufzuzeigen.

Als Datenbasis dienen Untersuchungsergebnisse zum Martingesetz der Abstraktionsebenen für die polnische Sprache (Sambor 1983) und Ergebnisse eigener Analysen des von J. Sambor (1982) gesammelten Materials. Die Ausgangsdaten unserer Untersuchungen werden in Tabelle 1 aufgeführt, wo

x die Nummer der Abstraktionsebenen,
y_x die Zahl unterschiedlicher Begriffe auf der x-ten Abstraktionsebene,
n_x die Zahl der Begriffsketten mit einer Länge von x Begriffen und
L_x die Gesamtzahl der auf der x-ten Ebene ausgesonderten Begriffe darstellt (Begriffswiederholungen auf derselben Begriffsebene sind bei der L_x-Zählung zugelassen, nicht aber bei der yx-Zählung).

2. Zunächst sollte festgestellt werden, welcher Zusammenhang zwischen der Länge der Begriffsketten n_x und der Nummer x der Abstraktionsebenen besteht.

* Diese Studie entstand im Rahmen des Projekts "Sprachliche Synergetik". Der Autor bedankt sich bei der **Stiftung Volkswagenwerk** für die freundliche Unterstützung.

Tabelle 1
Datenmaterial zum Martingesetz der Abstraktionsebenen
für die polnische Sprache (vgl. Sambor 1983)

x	y_x	L_x	n_x
1	1000	1000	0
2	618	1000	238
3	271	762	251
4	110	511	249
5	44	262	159
6	16	103	68
7	9	35	30
8	3	5	4
9	1	1	1

Wie man sich leicht überzeugen kann, gilt:

$$n_x = L_x - L_{x+1} \tag{1}$$

Aus Tabelle 1 kann abgelesen werden, daß n_x seinen Maximalwert $n_{x_{max}}$ bei $x = 3$ annimmt,wobei für alle

$$x \neq 3 \qquad n_x < n_{x_{max}} \tag{2}$$

gilt. Im folgenden soll die Funktion $n_x = f(x)$ mathematisch beschrieben werden als Konsequenz des nun vorzustellenden Modells.

Ähnlich wie Altmann (1985), der bei der Beschreibung semantischer Diversifikationsprozesse von den sogenannten "Zipfschen Kräften" ausgeht, beschreiben wir die Funktion $n_x = f(x)$ über das Wirken zweier in Gegenrichtung wirkender Kräfte F_1 und F_2. Wenn wir annehmen, daß den Begriffsketten im Prozeß der Informationsverarbeitung und -speicherung irgendeine psychologische Realität zukommt, so kann man auch davon ausgehen, daß für den Umgang mit relativ kurzen Begriffsketten Ökonomie-Argumente sprechen, für den Umgang mit relativ langen Begriffsketten Argumente für ein gut ausgebautes, relativ stark differenziertes Begriffssystem. Die Zahl der Begriffsketten n_x mit einer Länge von x in einer untersuchten Stichprobe ist somit eine Zufallsvariable als Ergebnis der Überlagerung der beschriebenen beiden Kräfte.

Ökonomie-Argumente sollen durch F_1 gekennzeichnet und über

$$F_1 \sim \frac{1}{x} \tag{3}$$

und

$$F_1 = \frac{c}{x} + d \tag{4}$$

beschrieben werden.

Der Trend nach einem stark differenzierten Begriffssystem wird durch F_2 erfaßt, wobei gilt:

$$F_2 \sim x \tag{5}$$

und

$$F_2 = ax + b. \tag{6}$$

Wir nehmen an, daß sich diese beiden Kräfte additiv zusammensetzen, so daß gilt:

$$n_x = \frac{c}{x} + ax + \underbrace{b + d}$$

$$n_x = \frac{c}{x} + ax + \quad e. \tag{7}$$

Die Konstanten a, c und e können wie folgt interpretiert werden:

c stellt den Kürzungskoeffizienten und a den Dehnungskoeffizienten der Begriffsketten dar. Die Konstanten c und a beschreiben somit den von x abhängigen Anstieg bzw. Abfall der Funktion $n_x = f(x)$. Die Konstante e dagegen setzt lediglich die durch a und c gekennzeichneten theoretischen Zahlenwerte in die Größenordnung der empirischen Zahlenwerte n_x.

Unter Anwendung der Methode der kleinsten Quadrate bei der Berechnung der Parameter a, c und e gilt:

$$a = \frac{(\sum_x n_x x \sum_x \frac{1}{x} - \sum_x \frac{n_x}{x} \sum x)(n \sum_x \frac{1}{x^2} - \sum_x \frac{1}{x} \sum_x \frac{1}{x}) -}{(\sum_x x^2 \sum_x \frac{1}{x} - n \sum_x x)(n \sum_x \frac{1}{x^2} - \sum_x \frac{1}{x} \sum_x \frac{1}{x}) -}$$

$$\frac{(n \sum_x \frac{n_x}{x} - \sum_x n_x \sum_x \frac{1}{x})(n \sum_x \frac{1}{x} - \sum_x \frac{1}{x^2} \sum x)}{(n^2 - \sum_x x \sum_x \frac{1}{x})(n \sum_x \frac{1}{x} - \sum_x \frac{1}{x^2} \sum x)} \qquad (8)$$

$$c = \frac{(\sum_x n_x x \sum_x \frac{1}{x} - \sum_x \frac{n_x}{x} \sum_x x)(n^2 - \sum_x x \sum_x \frac{1}{x}) -}{(n \sum_x \frac{1}{x} - \sum_x \frac{1}{x^2} \sum_x x)(n^2 - \sum_x x \sum_x \frac{1}{x}) -}$$

$$\frac{(n \sum_x \frac{n_x}{x} - \sum_x n_x \sum_x \frac{1}{x})(\sum_x x^2 \sum_x \frac{1}{x} - n \sum_x x)}{(n \sum_x \frac{1}{x^2} - \sum_x \frac{1}{x} \sum_x \frac{1}{x})(\sum_x x^2 \sum_x \frac{1}{x} - n \sum_x x)} \qquad (9)$$

$$e = \frac{\sum_x n_x x - cn - a \sum_x x^2}{\sum_x x} \qquad (10)$$

Setzt man in die Gleichungen (8), (9) und (10) die Zahlenwerte der Tabelle 1 ein, so erhält man folgende Ergebnisse:

$$\begin{aligned} a &= -68{,}963321 \\ c &= -547{,}50887 \\ e &= 628{,}02503 \end{aligned}$$

Setzt man nun wieder diese Zahlenwerte in Gleichung (7) ein, so kann man die in Tabelle 2 angegebenen theoretischen Zahlenwerte $\hat{n}_x$ berechnen.

Da die Funktion $\hat{n}_x = f(x) = c/x + ax + e$ die Abhängigkeit zwischen n_x und x nur im Intervall von $x = 1$ bis $x = 8$ gut beschreibt (für $x = 9$ erhält man einen negativen Wert für n_x), definieren wir die für unsere Untersuchungen abgeleitete Funktion $n_x = f(x)$ wie folgt:

$$f(x) = \begin{cases} \hat{n}_x & \text{für } x = 1, 2, \ldots, 8 \\ 0 & \text{für } x > 8 \quad (x \in N) \end{cases} \qquad (11)$$

Tabelle 2

Empirische Zahlenwerte n_x (nach Tabelle 1) und theoretische Zahlenwerte $\hat{n}_x$ (nach Gleichung (7))

x	n_x	$\hat{n}_x$	$(n_x - \hat{n}_x)^2$	$(n_x - \bar{n}_x)^2$
1	0	11.55	133.40	12345.679
2	238	216.34	469.16	16100.818
3	251	238.63	153.02	19568.932
4	249	215.29	1136.36	19013.376
5	159	173.71	216.38	2293.356
6	68	122.99	3023.90	1858.558
7	30	67.07	1374.19	6578.944
8	4	7.88	15.05	11472.766
	$\bar{n}_x = 111.111$		$\sum(n_x - \hat{n}_x)^2 =$ $= 6521.46$	$\sum(n_x - \bar{n}_x)^2 =$ $= 89232.429$

Aus den genannten Gründen wurde in Tabelle 2 der Zahlenwert $\hat{n}_x$ für $x = 9$ nicht aufgeführt. Bei der Überprüfung der Abweichungen zwischen n_x und $\hat{n}_x$ auf Signifikanz kann der F-Test nicht angewandt werden, da $\hat{n}_x = f(x)$ keine lineare Funktion darstellt. Zwar kann diese Regressionsfunktion nach der Durchführung bestimmter Transformationen und der von Krause, Metzler (1983:235) besprochenen quasilinearen Regressionsanalyse linearisiert werden, die Ergebnisse der Regressions- und Signifikanzprüfung beziehen sich jedoch nur auf die transformierten Variablen. Aus diesen Gründen soll zur ersten Abschätzung der Anpassung der theoretischen Werte an die empirischen Werte der von Kacprzak, Sobczyk, Zalewa, Zinczuk (1980, 313) vorgestellte Parameter Φ^2 mit

$$\Phi^2 = \frac{\sum_{x=1}^{n} (n_x - \hat{n}_x)^2}{\sum_{x=1}^{n} (n_x - \bar{n}_x)^2} \qquad (12)$$

berechnet werden, wo

$$\bar{n}_x = \frac{\sum_{x=1}^{n} n_x}{n} \qquad (13)$$

gilt. Φ^2 gibt an, welcher Teil der allgemeinen Varianz der Variablen n_x Spiel des Zufalls ist, d.h. nicht durch die Anpassung der Regressions-

funktion an die empirischen Daten erklärt wird. Somit zeugen kleine Φ^2-Werte von einer guten Anpassung.
Für unser Beispiel gilt:

$$\Phi^2 = \frac{\sum_{x=1}^{n}(n_x - \hat{n}_x)^2}{\sum_{x=1}^{n}(n_x - \bar{n}_x)^2} = \frac{6521.46}{89232.479} \approx 0.073.$$

Durch die Regressionsfunktion wird rund 93% der Gesamtvarianz der empirischen Werte n_x erklärt; dieses Ergebnis kann als zufriedenstellend angesehen werden. Unser Modell, welches vom Wirken zweier entgegengesetzter Kräfte spricht, spiegelt wesentliche Aspekte des untersuchten Sachverhaltes wider. Hier sei noch einmal vermerkt, daß auf analogen Grundlagen basierende Modelle auch von anderen Wissenschaftlern in sprachstatistischen Untersuchungen erfolgreich angewandt wurden (Altmann 1985). Dies läßt vermuten, daß die mit diesen Modellen untersuchten Sachverhalte somit in übergeordnete Zusammenhänge gestellt werden können.

3. Neben der Länge n_x der Begriffsketten interessiert uns noch eine weitere Größe, die Subsumptionspotenz g_x, die nach

$$g_x = \frac{L_x}{y_x} \tag{14}$$

berechnet werden kann. Die Subsumptionspotenz g_x gibt an, wieviel Begriffe der Ebene $x-1$, für die ein Oberbegriff auf der Ebene x angegeben werden kann, durchschnittlich unter einem Begriff der Ebene x subsumiert werden. Aus Tabelle 3 kann abgelesen werden, daß die Subsumptionspotenz g_x mit wachsendem x größer wird, bei $x = 6$ ihren Maximalwert erreicht und dann wiederum auf den Anfangswert von $g_x = 1$ abfällt.

Dieses Ergebnis ist so zu deuten, daß nicht unbedingt die Begriffe der höchsten Ebenen (zumindest bei der von Martin (1974), Altmann, Kind (1983) und Sambor (1983) angenommenen Zuordnung der Begriffe zu Begriffsebenen) die größte Zahl von Unterbegriffen auf der jeweils niedrigeren Ebene haben müssen. Bei einer anderen Zuordnung der Begriffe zu Begriffsebenen (z.B. bei einer Zusammenfassung der Ebenen 6-9 zu einer Ebene mit $x = 6$) kann erreicht werden, daß g_x mit wachsendem x stets zunimmt. Inwieweit eine Zusammenfassung

der Ebenen 5-9 angemessen ist, soll nicht an dieser Stelle geklärt werden. Eine erste Analyse des Datenmaterials zur polnischen Sprache (vgl. Sambor 1982) läßt aber die Vermutung zu, daß sich die Abstraktionsebenen mit $x = 5, 6, 7, 8$ und 9 hinsichtlich der auf diesen Ebenen auftretenden Begriffe nicht wesentlich unterscheiden.

Tabelle 3
Berechnung der Subsumptionspotenz g_x nach Gleichung (14)

x	y_x	L_x	g_x
1	1000	1000	1.00
2	618	1000	1.62
3	271	762	2.81
4	110	511	4.65
5	44	262	5.95
6	16	103	6.44
7	9	35	3.89
8	3	5	1.67
9	1	1	1.00

Im folgenden soll untersucht werden, welcher Zusammenhang zwischen der Subsumptionspotenz g_x und der Variablen x besteht. Es wird also danach gefragt, nach welchen Gesetzmäßigkeiten sich die Subsumptionspotenz g_x mit wachsendem x (d.h. auf den jeweils höheren Begriffsebenen) ändert.

Bei der Lösung dieses Problems sind wir zunächst wieder – analog der n_x-Berechnung nach Gleichung (7) – vom Wirken zweier Kräfte ausgegangen. Hier wurde aber nicht angenommen, daß beide Kräfte in entgegengesetzter Richtung wirken, sondern daß die Einwirkung dieser Kräfte unter einem bestimmten Winkel β erfolgt.
g_x sollte dann nach

$$g_x = \left[F_1^2 + F_2^2 + 2F_1F_2\cos\beta\right]^{1/2} \tag{15}$$

mit $F_1 = c/x + d$ und $F_2 = ax + b$ $(a, b, c, d = \text{konstant})$ berechnet werden. Setzt man diese Werte für F_1 und F_2 in Gleichung (15) ein, so erhält man:

$$g_x = \left[o/x^2 + p/x + rx^2 + sx + t\right]^{1/2} \tag{16}$$

Die Konstanten o, p, r, s, t können leicht aus den Konstanten a, b, c, d und β berechnet werden.

Wie unsere Untersuchungen zeigten, beschreibt Gleichung (16) die Verteilung der empirischen Werte für g_x nicht mit einer ausreichenden statistischen Sicherheit. Aus diesem Grunde sind wir von einer anderen Annahme ausgegangen:

Das Verhältnis der Zahlenwerte $(g_{x+1})/(g_x)$ hängt in einer bestimmten Weise von $(x+1)$ ab und kann durch

$$\frac{g_{x+1}}{g_x} = \frac{c}{(x+1)^a} \tag{17}$$

beschrieben werden, wo c und a Konstanten darstellen.

Gleichung (17) stellt einen Spezialfall der von uns bei der Beschreibung des Martingesetzes der Abstraktionsebenen (bei der Beschreibung des Zusammenhangs zwischen y_x und x) angewandten Formel dar (vgl. Hammerl 1987).

Die Konstanten in Gleichung (17) können nach

$$c = e^w \tag{18}$$

mit

$$w = \frac{\sum_x \frac{g_{x+1}}{g_x}(x+1)\sum_x(x+1) - \sum_x \frac{g_{x+1}}{g_x}\sum_x(x+1)^2}{\left[\sum_x(x+1)\right]^2 - n\sum_x(x+1)^2} \tag{19}$$

und

$$a = \frac{n\sum_x \frac{g_{x+1}}{g_x}(x+1) - \sum_x \frac{g_{x+1}}{g_x}\sum_x(x+1)^2}{\left[\sum_x(x+1)\right]^2 - n\sum_x(x+1)^2} \tag{20}$$

berechnet werden.

Unter der Voraussetzung, daß $g_1 = 1{,}00$ und mit $a = 0{,}9072903$ und $c = 4{,}2710961$ wurden die in Tabelle 4 angeführten theoretischen Zahlenwerte $\hat{g}_x$ berechnet.

Sollen die Abweichungen zwischen den theoretischen Zahlenwerten $\hat{g}_x$ und den empirischen Zahlenwerten g_x aus Tabelle 4 mit dem F-Test untersucht werden, so muß die Funktion $g_x(x) = f(x)$ aus Gleichung (17) zunächst linearisiert werden.

$$\underbrace{\frac{g_{x+1}}{g_x}} = \frac{c}{(x+1)^a}$$

$$l_x = \frac{g_{x+1}}{g_x} = \frac{c}{(x+1)^a}$$

$$\underbrace{\ln l_x} = \underbrace{\ln c} - a\,\underbrace{\ln(x+1)}$$

$$G_x = C - a\,(X+1) \tag{21}$$

Die nach Gleichung (21) berechneten Daten für $\hat{G}_x$ (theoretische Daten) und die nach

$$G_x = \ln \frac{g_{x+1}}{g_x} \tag{22}$$

berechneten Daten G_x (empirische Daten) werden ebenfalls in Tabelle 4 aufgeführt.

Die Anwendung des F-Testes erbringt folgende Zahlenwerte:

$$F_{emp} \approx 8{,}27 \qquad F_{2,6}(\alpha = 0{,}05) = 5{,}14.$$

Somit kann die Anpassung der theoretischen Zahlenwerte $\hat{G}_x$ an die empirischen Zahlenwerte G_x bei einer Irrtumswahrscheinlichkeit von kleiner als 0,05 als statistisch signifikant bezeichnet werden. Somit wird auch für die nichttransformierten Daten $\hat{g}_x$ und g_x eine gute Anpassung angenommen.

Die für die Beschreibung des Zusammenhangs zwischen g_x und x angewandte Gleichung (17) ist sehr allgemein und läßt auch die Beschreibung solcher Zahlenwerte für g_x zu, wo z.B. $g_{x+1} \leq g_x$ (für alle x) ist, $g_{x+1} \geq g_x$ (für alle x) oder wo – wie im untersuchten Beispiel – für kleine x $g_{x+1} > g_x$ und für große x $g_{x+1} < g_x$ gilt. Wichtig ist aber immer, daß

Tabelle 4
Berechnung der theoretischen Zahlenwerte für die Subsumptionspotenz $\hat{g}_x$ (nach Gleichung 17), Berechnung der $\hat{G}_x$-Werte (nach Gleichung 21) und der G_x-Werte (nach Gleichung 22)

x	g_x	$\hat{g}_x$	G_x	$\hat{G}_x$	$(G_x - \hat{G}_x)^2$	$(G_x - \bar{G}_x)^2$
1	1.00	1.00	0.48242	0.82298	0.11598	0.23272
2	1.62	2.28	0.55075	0.45511	0.009147	0.30332
3	2.81	3.59	0.50368	0.19409	0.095845	0.25369
4	4.65	4.36	0.24652	-0.0083568	0.064962	0.06077
5	5.95	4.32	0.07914	-0.17377	0.063963	0.00626
6	6.44	3.63	-0.50411	-0.31363	0.036282	0.25412
7	3.89	2.66	-0.84558	-0.4347866	0.16875	0.71500
8	1.67	1.72	-0.51282	-0.54165	0.00083116	0.26298
9	1.00	1.00	-	-	-	-
$\bar{g}_x$ = 3.2256 $\bar{G}_x$ = 0.00000 SSE = 0.5557601 SST = 2.0889						

$$\frac{g_{x+1}}{g_x} \sim \frac{1}{(x+1)^a}$$

ist und daß $(g_{x+1})/g_x$ mit wachsendem x abnimmt (mit $a > 0$), was durch

$$\lim_{x \to \infty} \left[\frac{g_{x+1}}{g_x} \right] = 0 \qquad (23)$$

erfaßt werden kann.

Hiermit wurde bestätigt, daß die von uns bei der Beschreibung des Martingesetzes der Abstraktionsebenen angewandte Formel (vgl. Hammerl 1987) auch auf die Beschreibung verwandter Sachverhalte bezogen werden kann. Somit wurde ein erster Schritt in Richtung der empirischen Validierung des dieser Formel zugrunde liegenden Modells getan.

4. Die folgenden Ausführungen sind dem Problem der Untersuchung des Abstraktionsniveaus der Begriffsklassen gewidmet. Untersuchungen zum Abstraktionsniveau lexikalischer Einheiten wurden schon

von mehreren Forschern durchgeführt (vgl. Flesch 1950; Gillie 1957; Günther, Groeben 1978; Paivio 1963; Kisro-Völker 1984); verschiedene Auffassungen über den Begriff des Abstraktionsniveaus lexikalischer Einheiten führten zu verschiedenen Methoden deren Messung. Kisro-Völker (1984) hat eine kritische Analyse dieser Methoden vorgenommen und selbst eine "objektive" Methode zur Bestimmung des Abstraktionsniveaus von Substantiven vorgeschlagen; bei der Anwendung dieser Methode wird vorausgesetzt, daß z.B. für alle Substantive aus einem bestimmten Wörterbuch Begriffsketten vorliegen. Das Abstraktionsniveau eines bestimmten Substantivs aus der Menge der somit erfaßten Substantive wird dann als Mittelwert aus den Nummern x_i der Abstraktionsstufen bestimmt, auf denen das jeweilige Substantiv innerhalb der untersuchten Begriffsketten auftritt. Auch diese Methode kann nicht als "objektive" Methode bezeichnet werden, da sowohl über die Tatsache der Verwendung konkreter Wörterbücher (die ja selbst von bestimmten Autorenkollektiven erarbeitet wurden) und über das Eingehen von verschiedenen Konventionen bei der Ableitung der jeweiligen Begriffsfolgen eine gewisse Subjektivität bei der Datenerfassung nicht zu vermeiden ist. Vor allem bei der Ableitung von 1000 deutschen Begriffsketten anhand des Handwörterbuchs der deutschen Gegenwartssprache (1984) konnten wir uns davon überzeugen, daß die von Altmann, Kind (1983:2-3) genannten Schwierigkeiten für eine möglichst objektive Ableitung von Begriffsketten aus einem bestimmten Wörterbuch bei weitem nicht alle wichtigen Aspekte dieses Sachverhalts betreffen (vgl. Hammerl 1986, Schierholz 1982). Die Bestimmung des Abstraktionsniveaus lexikalischer Einheiten nach der von Kisro-Völker (1984) vorgeschlagenen Methode ist schon bei der Beschränkung auf nur ein einzelnes Wörterbuch äußerst zeitaufwendig. Die Anwendung der Computertechnik bei der Bewältigung dieser Aufgaben bringt erst dann wesentliche Arbeitserleichterungen und Arbeitszeitverkürzungen, wenn es gelingt, nach der Einspeicherung aller wesentlichen Wörterbuchdaten eines einsprachigen Wörterbuchs auch alle zulässigen semantischen Beziehungen zwischen den Lexembedeutungen dieses Wörterbuchs zu erfassen.

Mit der Methode von Kisro-Völker bestimmt man das Abstraktionsniveau z.B. der Substantive als Mittelwert aus konkreten Zahlenwerten der Variablen x, die ja Nummern von Abstraktionsebenen betrifft (mit

$x = 1, 2, 3, \ldots$). Das Abstraktionsniveau wird somit als Zahlenwert einer Ordinalskala berechnet.

Wir wollen im folgenden versuchen, das mittlere Abstraktionsniveau a_x der Begriffe der jeweiligen Abstraktionsebene x als Zahlenwert einer Intervallskala abzuschätzen, ohne die jeweiligen Intervallgrenzen angeben zu können. Es scheint uns auch nicht sinnvoll, genaue Intervallgrenzen zu berechnen, da wir es erstens mit Mittelwerten des Abstraktionsniveaus von Begriffen zu tun haben und da zweitens die untersuchten Begriffsmengen, d.h. die jeweilige Menge von Begriffen auf der entsprechenden Abstraktionsstufe, hinsichtlich der untersuchten Eigenschaft "Abstraktionsniveau a_x" als unscharfe Mengen im Sinne von Zadeh (1983) (vgl. auch Hoffmann, Piotrowski 1979, Rosch 1978, Piotrowski 1984) anzusehen sind.

Das Abstraktionsniveau a_x sagt somit etwas über die inhaltliche Seite der jeweils untersuchten lexikalischen Einheiten aus, genauer über den begrifflichen Kern deren Bedeutungen. Aus diesem Grunde wird in diesem Kapitel vom mittleren Abstraktionsniveau a_x von Begriffen gesprochen.

Bei der Berechnung von a_x gehen wir von folgenden Annahmen aus:

a) Die Aktivierung eines bestimmten Begriffs unseres Begriffssystems zieht die Aktivierung aller derjenigen Begriffsfolgen nach sich, in denen der untersuchte Begriff auftritt.

b) Die mittlere Aktivierungshäufigkeit der Begriffe der Ebene x soll nach

$$h_x = \frac{L_x}{L_1} \tag{24}$$

berechnet werden.

c) Das mittlere Abstraktionsniveau a_x ist eine Funktion der Auftrittshäufigkeit h_x, so daß gilt:

$$a_x = f(h_x) \tag{25}$$

Eine erste Approximation dieser Abhängigkeit soll über eine indirekte Proportionalität zwischen a_x und h_x erreicht werden. Somit wird angenommen, daß

$$a_x = \frac{c}{h_x} = c\,\frac{L_1}{L_x} \tag{26}$$

gilt, wo c eine Konstante darstellt (vgl. auch Schierholz 1982).

Da uns die Konstante c aus Gleichung (26) nicht bekannt ist, können wir a_x auch nicht direkt berechnen; möglich ist lediglich die Berechnung des Verhältnisses von Abstraktionsniveaus. In Tabelle 5 wird folgende Größe eingeführt:

$$a_{x+1,x} = \frac{a_{x+1}}{a_x}. \tag{27}$$

Diese Größe sagt etwas über das Verhältnis des Abstraktionsniveaus der Begriffe auf zwei benachbarten Abstraktionsebenen $x+1$ und x aus.

Wie aus Tabelle 5 ersichtlich wird, nimmt das Verhältnis des Abstraktionsniveaus der jeweils benachbarten Abstraktionsstufen im Bereich von $x = 1$ bis $x = 6$ stets zu; für den Bereich von $x = 6$ bis $x = 9$ kann diese Tendenz nicht bestätigt werden. Auch diese Untersuchungsergebnisse können unserer Meinung nach darauf hinweisen, daß die höchsten Abstraktionsebenen zu einer einzigen Abstraktionsebene zusammengefaßt werden sollten.

Tabelle 5
Zahlenwerte für h_x (nach Gleichung 24),
a_x (nach Gleichung 26) und $a_{x+1,x}$
(nach Gleichung 27) für die polnische Sprache

x	L_x	h_x	a_x	$a_{x+1,x}$
1	1000	1.000	c · 1.000	1.000
2	1000	1.000	c · 1.000	1.313
3	762	0.762	c · 1.313	1.490
4	511	0.511	c · 1.957	1.950
5	262	0.262	c · 3.817	2.544
6	103	0.103	c · 0.709	1.873
7	55	0.055	c · 18.182	11.000
8	5	0.005	c · 200.000	5.000
9	1	0.001	c · 1000.000	

In diesem Kapitel kam es vor allem darauf an zu zeigen, daß eine Berechnung des Abstraktionsniveaus von Begriffen auf einer Intervallskala

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 121–134

Polylexy and Compounding*

Ursula Rothe
Bochum

The reduction of the length of a word results in a growth of its constituents (a) as well as in an increase of the number of its meanings (b). These hypotheses have been corroborated by means of data from many languages (Geršić, Altmann 1980; Rothe 1983; Sambor 1984; Fickermann, Markner-Jäger, Rothe 1984). Hypothesis (a) is applicable not only to words but to morphemes and sentences as well (Gerlach 1982, Köhler 1982, Heups 1983, Schwibbe 1984, Teupenhayn, Altmann 1984).

Köhler (1984) interprets these law-like relationships as based on the capacity of our memory, which has to store the information about a construct itself as well as the information about its structure (i. e. the relations of the constituents within the construct). Both kinds of information have to be simultaneously resident in the memory during the act of communication. This law has been corroborated not only in linguistics, where it is called Menzerath's law, but in several other sciences, where its analogy is known as the "allometric law" (cf. Naroll, v. Bertalanffy 1956). It is a law of relative growth, in which a given dimension of a part of a system (i. e. a subsystem) is compared to the corresponding dimension of another part of the same system (another subsystem) or of the whole (i. e. the system under consideration itself).

In all cases the proportional relationships concerned exist between the construct and its components, between the whole and its parts, between the system and its subsystems. Earlier linguistic and biological investigations preferred to examine the dependencies between physical variables, but the linguists are also interested in the questions (a) whether physical variables (such as word length) have any effect on

* Diese Studie entstand im Rahmen des Projekts "Sprachliche Synergetik". Die Autorin bedankt sich bei der **Stiftung Volkswagenwerk** für die freundliche Unterstützung

non-physical properties of the constituents (such as the polylexy of a word, as mentioned above), (b) whether non-physical properties have any effect on further non-physical properties (such as the affinity of lexemes to combine with other lexemes) and if so, (c) what form the dependencies take.

Starting from the axiom that language is a system, that its entities are subsystems, that some basic properties of these entities form dimensions, and accepting the principle of the dimensional invariance of the systems (cf. Sahal 1977), it follows that Menzerath's law is valid for many more relationships than was formerly assumed (cf. Altmann, Schwibbe, Kaumans, Köhler, Wilde, forthcoming).

In the present paper the relation between the number of meanings of a word (polylexy) and its (paradigmatic) property to combine with other words, i. e. its tendency to form compounds, will be investigated.

The hypothesis can be formulated as follows:

"the greater the polylexy of a word
the more compounds it forms".

The hypothesis will first be tested on German data selected from the lexicon Wahrig 1985, and we assume that in case of corroboration the hypothesis is valid for other languages, too.

* *
*

It is necessary at this point to make some remarks on the different types of compounding and on the conventions concerning their classification.

In German, morphologic compounding (as in *Eisenbahn*) is the most frequent form, but many languages also form compounds with hyphens (German: *Altbau-Modernisierung* (cf. Kürschner 1974:190), French: *wagon-lits*). A further type is the so-called free compounding (German: *Vereinte Nationen*, English: *word length*), which surely requires different treatments from language to language. In French the most frequent compounds are periphrastic, as in *chemin de fer* or *machine à coudre*. However a differentiation must be made in Chinese, where compounding is, among other things, a means of synthesizing meaning (radical) and pronounciation (phonetic) of a given lexeme (phonetic compounding, for ex.: *fǔ mǔ* which designates *father, mother* and means *parents*).

In all these cases the boundaries between compounding and derivation must be drawn for each language separately. In German, hyphenated words can be considered to be true compounds, due to the fact that there are often alternatives to fixed compounding. On the other hand, free compounding (like *Vereinte Nationen*) fall into another category. They often occur with proper names, which are not meaningful lexemes in the broader semantic sense. But besides the above mentioned 'formal' differences semantic considerations must be taken into account.

We must distinguish between three kinds of semantic categories of compounds:

regular:	*Steinofen*	(paraphrase with components both used in their original meaning: oven out of stone),
partially motivated:	*Hochofen*	(paraphrase: (?) oven which is high/tall, actually an iron-to-steel converter),
	Eselsbrücke	(bridge not for/out of/from mules, direct paraphrase only in the metaphorical sense, actually a mnemonic device),
unmotivated:	*Hochzeit*	(no direct paraphrase).

All components of these four compounds have corresponding autonomous homophones in the lexicon. In the first case the homophones are synonymous; in the second the synonymy lies within the semantic variations of polysemic words, only in the last example the signification of the homophones *Hoch-* and *-zeit* is not identifiable via direct paraphrase (examples cf. Vögeding 1981:95,96).

Analogically the classification used to distinguish real compounds from derivates by means of affixes or affixoids will here be based on the paraphrasing method. The paraphrasing conventions which each compound must meet in our present concerns shall be the following:

1. The compound can be analysed into two components which exist autonomously in the same lexicon,

2. the components must belong to the same field of meaning as described in their entry (i. e. they must be considered within the variation of their polysemic field); if the meaning lies outside this variation, they are homonyms, the status of which in the compound must be tested by paraphrase:
3. the interpretation of the compound by means of paraphrase consists of the two isolated components (in their original unisemic/polysemic sense) and an expression that reflects the relationship of the two components in the compound.

Examples:

	compounds	**no compounds**
nouns:	*der Stiefelschaft*	*die Pflegschaft*
	das Lebewesen	*das Schulwesen*
	die Holzbrücke	
	die Eselsbrücke (metaph.)	
		das Hauptgericht
	das Tage-, Bergwerk	*das Bollwerk*
verbs:	*'übersetzen* (e. g. a boat, particle separable)	*über'setzen* (e. g. a book, particle) not separable)
		mutmaßen

It is clear that this is only a convention; it does not claim to be an all-inclusive definition of a compound. In Augst 1975, for example, elements like *Haupt-*, *-wesen*, *-zeug-*, *-werk*, etc. are classified as parts of compounds with a remark of the kind "oft als 1./2. Teil in Zusammensetzungen, ohne genaue Bedeutung". Erben 1981, Fleischer 1971, Kürschner 1974 don't ignore such homonymous relations, but give further criteria (diachronic view, "Reihenbildung", intonation) and define the category between affixes and components of compounds: affixoids.

* *
*

The data were selected as follows: for each word that was the first component of a compound the number of meanings (as noted in

Wahrig) and the number of compounds it forms were counted. We realize that considering the compounds only with regard to their first component (which is often but not in all cases the determinans) is only a partial treatment of the problem. But the analysis of compounds with regard to the second component (which is in most cases the determinandum) will be the task of a later examination, though we assume that the relationship will be the same.

In selecting the relevant components of compound words, all word-classes were taken into account. In order to distinguish a component of a compound from affixes, prefixes or further morphological elements, we took as criterion its autonomous existence in the German lexicon, that is, its existence as a lexeme that can also be used in isolation. Morphological combining techiniques such as the "Fugenmorphem" (*die Taste – Taste-n-klavier*) or the elimination of syllables (*tasten – Tast(en)sinn*) are special means of and were eliminated or added in the course of the segmentation.

Since the number of lexemes having only one meaning is enormous, we counted them only up to the entry *Dulder*, after verifying that their number (773) was sufficient. The total of words counted as far as *Dulder* was 985. From then on until the end of the lexicon we processed only components having more than one meaning. This section of the data yielded 873 value-pairs, amounting to 1858 encounted lexemes in total. Polylexy classes that contained less than 10 items – which was the case especially in the classes with 10 meanings and more – were pooled into a single class.

* *
*

To test the above hypothesis we assume that the relative rate of change of the number of compounds formed by a lexeme (x) is proportional to the relative rate of change of its number of meanings (y) and set up the equation (cf. Altmann 1980):

$$\frac{dy}{y} = \frac{bdx}{x}, \qquad (1)$$

the solution of which results in the function

$$y = ax^b. \tag{2}$$

As will be shown below, the fitting of the function to the present data yields significant results.

To estimate the coefficients a and b, the function was linearized by means of a logarithmic transformation:

$$y = ax^b \Longrightarrow \ln y = \ln a + b \times \ln x. \tag{3}$$

and the coefficients were obtained using the method of least squares.

For measuring the goodness of fit the usual methods of the theory of linear regression including the F-test were used. The level of significance was fixed at $\alpha = 0.05$. The results for either nouns, verbs, adjectives/adverbs or for all word-classes in total are presented in Tables 1 to 4 and in the corresponding Figures 1 to 4.

* *
*

Tab. 1.
Relation between the polylexy of nouns and the number of compounds

Polylexy (x)	Average number of Compounds (y) observed	computed	Frequency
1	4.44	5.7954	654
2	8.94	10.4284	162
3	12.54	14.7050	41
4	17.79	18.7654	194
5	28.89	22.6723	239
6	27.81	26.4610	31
7	30.83	30.1541	12
8	31.00	33.7674	10
≥ 9	36.52	37.3124	18
a = 5.8116		F(1,7) = 116.22	
b = 0.8460		p = 0.000013	

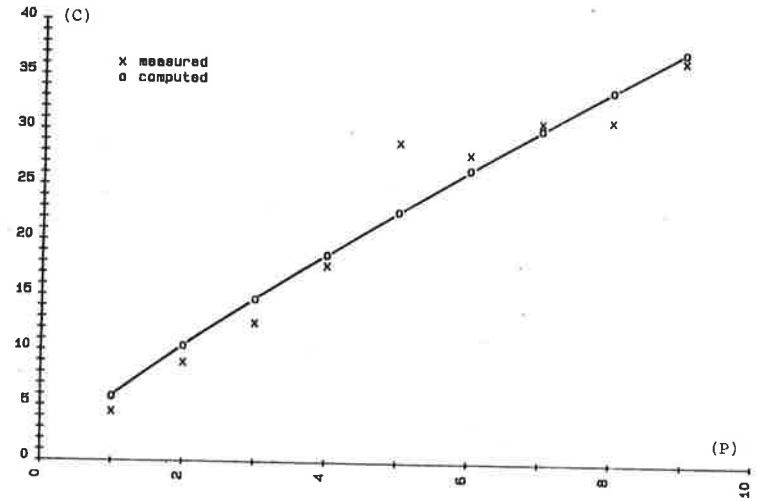


Fig. 1.
Relation between the polylexy of nouns (P) and the number of compounds (C)

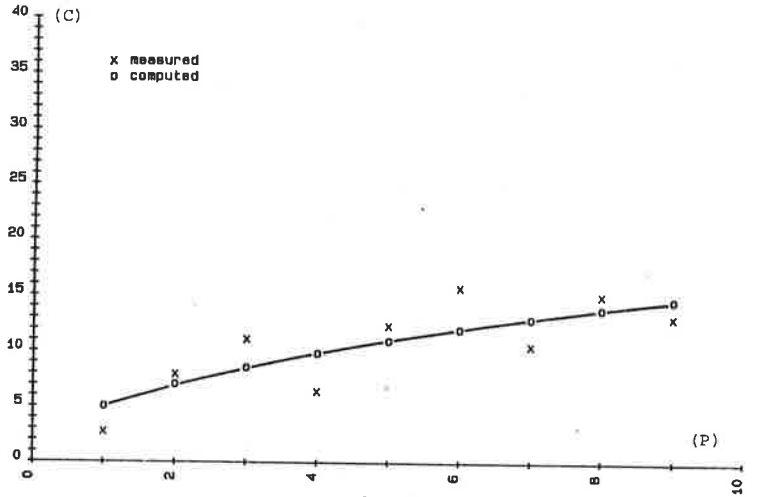


Fig. 2.
Relation between the polylexy of verbs (P) and the number of compounds (C)

Tab. 2.
Relation between the polylexy of verbs
and the number of compounds

Polylexy (x)	Average number of Compounds (y) observed	computed	Frequency
1	2.70	5.0102	69
2	7.92	7.0312	71
3	11.10	8.5727	20
4	6.35	9.8673	20
5	12.29	11.0046	28
6	15.71	12.0306	21
7	10.53	12.9727	15
8	15.00	13.8474	10
≥ 9	13.06	14.6681	16
a = 4.9833		F(1,7) = 12.93	
b = 0.4926		p = 0.0087	

Tab. 3.
Relation between the polylexy of adjectives/adverbs
and the number of compounds

Polylexy (x)	Average number of Compounds (y) observed	computed	Frequency
1	3.87	4.7322	47
2	5.61	8.6801	18
3	9.24	12.3776	21
4	21.63	15.9214	49
5	20.33	19.3551	39
6	27.30	22.7036	10
≥ 7	20.32	25.9828	12
a = 4.6981		F(1,5) = 18.22	
b = 0.8788		p = 0.00000013	

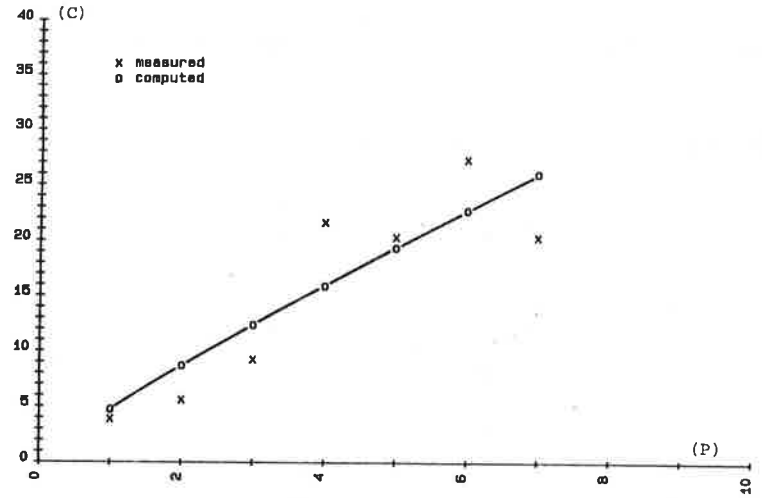


Fig. 3.
Relation between the polylexy of adjectives/adverbs (P)
and the number of compounds (C)

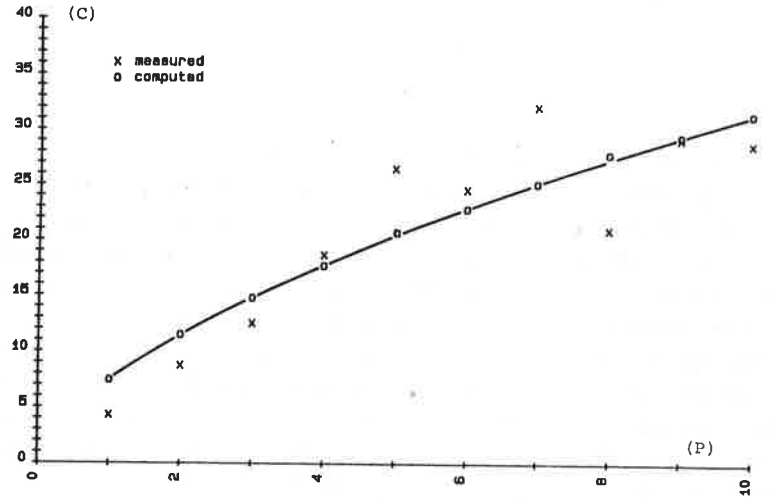


Fig. 4.
Relation between the polylexy (P) and the number of compounds (C)
(all word classes)

Tab. 4.
Relation between the polylexy and the number of compounds (all word classes)

Polylexy (x)	Average number of Compounds (y) observed	computed	Frequency
1	4.27	7.44	771
2	8.73	11.47	257
3	12.53	14.78	88
4	18.67	17.69	267
5	26.46	20.34	310
6	24.52	22.80	66
7	32.03	25.10	32
8	20.92	27.29	24
9	29.14	29.37	21
> 10	28.68	31.37	22
a = 7.4356		F(1,8) = 31.80	
b = 0.6256		p = 0.0049	

* *
*

All fits yield highly significant results. The best fits are obtained with nouns and adjectives/adverbs. The fit for the verbs is the least significant of all, but is still in agreement with the theoretical curve. Prepositions and pronouns have a quite restricted occurrence (though they can form compounds, e.g.: *seinetwegen*, *mitbringen*). So they only could be pooled to the data containing all word classes.

Though the results are satisfactory there are some linguistic and methodological problems that will be mentioned below. The relatively weak agreement of the verb class with the underlying hypothesis, though still significant, gives rise to some further speculations. In the process of compounding, verbs show two morphological peculiarities which probably cause this result.

The first is the shortening of the verb in the compounds up to the pure stem – a property that concerns above all languages with a

morphological infinitive. So the compound with *mischen* for example as first component does not yield *Mischenventil*, but *Mischventil* (cf. also the example *Tastsinn* mentioned above).

Possibly the tendency to combine with further elements decreases because of the infinitive. The infinitive morpheme *-en* must be deleted on the one hand, whereas on the other hand the lexeme is extended by means of compounding; these are two opposite cognitive efforts: shortening (reduction) and lengthening (extension). The second property is the frequent nominalisation of verbs, especially with the derivative morpheme *-ung*, like in *Mischung*, out of *mischen*. Such nominalisations also very often form compounds. We assume that these nominalisations are the more frequent the greater the polylexy of the verb concerned. This hypothesis is set up on empirical grounds and has not been tested as yet. To what extent this circumstance will necessitate a modification of the present hypothesis must be ascertained by further investigations of all kinds of complex words (lexemes with prefixes, infixes, suffixes as well as composed lexemes). Possibly the consideration of the second component will bring some additional aspects of this hypothesis to light, since it is the second component that determines in German the word class of the whole compound.

Such an examination will possibly allow us to give a full interpretation of the two parameters. What we can see for the moment is that b expresses the slope of the curve, and that parameter a represents the mean number of compounds that are formed by words with one meaning.

There is no problem in using a continuous curve for our data since the data are only seemingly discrete. The average number of compounds is a continuous quality and the number of meanings is a quantity that must be rendered discrete in order to be measurable at all. As a matter of fact, in the case of many linguistic phenomena (cf. the speech process) the boundary between discrete and continuous values is not quite sharp, so that it is not necessary to give reasons for a special approach in every case.

* *
*

Now – according to the rule of transitivity – from the dependencies

(i) *"the shorter a word the greater its polylexy"*

and

(ii) *"the greater the polylexy of a word the more compounds it forms"*

it follows that

(iii) *"the shorter a word the more compounds it forms"*.

One can state that, since the longer words probably have more additional morphological elements (affixes, suffixes, derivational morphemes etc.), the shorter words are more probably pure stems. That is, the morphological reduction of a word (or, more generally, the syntagmatic reduction of a unit) to a minimum (here lexeme:stem = 1:1), results in a more productive paradigmatic affinity of the unit (i. e. semantic diversification and combinatorial diversification). The combinatorial diversification can thus further be interpreted as a means of lengthening a unit, and this again concerns its syntagmatic level. Thus the syntagmatic state of the lexicon remains steady. That the same is true for the paradigmatic state can be derived easily by pursuing the above transitive relationships from the opposite direction. To generalize this mechanism within the framework of a systemic approach, we can say that the equilibrium in the lexicon is realized on the one hand by stabilizing the syntagmatic state of the lexicon by means of its paradigmatic productivity and on the other by stabilizing the paradigmatic state by means of syntagmatic extension. Of course this is only one aspect, but it possibly can be brought into correspondence with the oscillation of the lexicon observed in Köhlers self-regulation model (1986), when examining the word length as dependent upon its frequency. For the interpretation of these phenomena in the linguistic regulation-control scheme see Rothe (forthcoming).

What we can state for the moment is that there must exist a feedback mechanism which has a regulating influence on any language unit under consideration. Word compounding is the result of such a feedback mechanism in the way that it counteracts the processes of formal reduction and semantic diversification which are – according to Zipf – permanently current in the language.

* *
*

References

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

Altmann, G., Beöthy, E., and Best, K.-H. (1982). "Die Bedeutungskomplexität der Wörter und das Menzerath'sche Gesetz." *Zs. für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 35, 537-543.

Altmann, G. (forthcoming 1988) "Hypotheses about compounds." In Köhler, R. (ed). *Studies in Language Synergetics.*

Altmann, G., Schwibbe, M., Kaumans, , Köhler, R., and Wilde (forthcoming 1988). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.*

Augst, G. (1975). *Lexikon zur Wortbildung, Forschungsberichte des Instituts für deutsche Sprache.*

Fickermann, I., Markner-Jäger, B., and Rothe, U. (1984). "Wortlänge und Bedeutungskomplexität". *Glottometrika* 6, 115-126.

Fleischer, W. (1971). *Wortbildung der deutschen Gegenwartssprache.* Leipzig: VEB Bibliographisches Institut, Tübingen: Niemeyer.

Gerlach, R. (1982). "Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie". *Glottometrika* 4, 95-102.

Geršić, L.S., and Altmann, G. (1980). "Laut – Silbe – Wort und das Menzerathsche Gesetz". *Forum Phoneticum* 21, 115-123.

Heups, G. (1983). "Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen". *Glottometrika* 5, 113-133.

Köhler, R. (1982). "Das Menzerathsche Gesetz auf Satzebene". *Glottometrika* 4, 103-113.

Köhler, R. (1984). "Zur Interpretation des Menzerathschen Gesetzes". *Glottometrika* 6, 177-183.

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

Kürschner, W. (1974). *Zur syntaktischen Beschreibung deutscher Nominalkomposita. Auf der Grundlage generativer Transformationsgrammatiken.* Tübingen: Niemeyer.

Naroll, R.S., and v. Bertalanffy, L. (1956). "The principle of allometry and the social sciences". *General Systems* 1, 76-89.

Rothe, U. (1983). "Wortlänge und Bedeutungsmenge. Eine Untersuchung zum Menzerath'schen Gesetz an drei romanischen Sprachen". *Glottometrika* 5, 101-112.

Rothe, U. (1988). "The regulation of formal and semantic stability of words". In Köhler, R. (ed). *Studies in Language Synergetics.*

Sahal, D. (1977). "Towards a theory of systems". *General Systems* 12, 41-55.

Sambor, J. (1984). "Menzerath's law and the polysemy of words". *Glottometrika* 6, 94-114.

Schwibbe, M.-H. (1984). "Text- und wortstatistische Untersuchungen zur Validität der Menzerathschen Regel". *Glottometrika* 6, 152-176.
Teupenhayn, R., and Altmann, G. (1984). "Clause length and Menzerath's law". *Glottometrika* 6, 127-131.
Vögeding, J. (1981). *Das Halbsuffix '-frei'. Zur Theorie der Wortbildung.* Tübingen: Narr.
Wahrig (1985). *Deutsches Wörterbuch.* Mosaik.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 135–145

Problems of quantitative text analysis

M. G. Boroda
Tbilisi
V. A. Dolinskij
Moscow

(A survey of seminars held by the research group "Text as an object of interdisciplinary investigations" in Voronovo (USSR), 8 – 12 October 1985 and Tartu (USSR), 27 – 31 January 1986)

Recent years have witnessed a remarkable growth of interest in, and the beginning of a new stage in the development of wide-rangeing studies of text as a specific phenomenon of a complex nature. Along with the extensive use of quantitative-mathematical methods, this new stage is characterized by attempts to connect the results of investigation (and, at times, the very formulation of problems) with hypotheses and arguments about *qualitative* aspects of texts such as their artisitic form, genre, language, etc. In particular, the tendency is manifested by a transition from "purely statistical" models to ones based on *psychological* hypotheses about principles of text perception and text generation.

Other important features of the new stage include an increasing interest shown in quantitative text analysis by specialists from *different* fields, the emergence of new objects of investigation: *nonlinguistic* texts (e. g. musical texts), specific types of texts such as free-association flows, etc. At the same time there is a tendency towards *integrating* all these studies with a view to uncovering general principles of the generation of a text as a coherent whole. Under these circumstances it is only natural that research groups should be established dealing with problems of quantitative text analysis.

One such group – *"Text as an object of interdisciplinary investiga-ns"* – was set up on the initiative of a number of Soviet researchers at e beginning of 1985. The group comprises representatives of various sciplines – linguists, mathematicians, musicologists, psychologists, c. – from higher educational establishments of Moscow, Leningrad, ilisi, Tartu, Yerevan, Kiev, Minsk, Riga and other cities of the USSR. ie aim of the group is to further progress in integrated quantita-e/qualitative text investigations, in studies of general and specific e. language-, genre-, style-, or purpose-specific) principles and regu-ities of text organization, in the construction of mathematical models text generation and, at least potentially, in devising joint research ojects in these fields.

The group has an open structure. Its work is arranged by an or-nizing committee, the members of which are: Ju. A. Tuldava (chair-an, Tartu State University), M. G. Boroda (Tbilisi State Conserva-ry), V. K. Detlovs (Latvian State University), and A. A. Polikarpov Ioscow State University). Seminars are held twice a year, devoted reports delivered by members of the group as well as guest-lecturers d to discussions of focal issues. The following text will give a survey the first two seminars (October 1985, the village of Voronovo, and nuary 1986, Tartu) sponsored by the M. Lomonossov State Univer-y (Moscow) and Tartu State University respectively.

The first seminar was arranged by the Department of General, omparative-Historical and Applied Linguistics of Moscow State Uni-rsity.

At the plenary session papers were read by P. M. Alekseyev and . A. Tuldava.

The paper by P. M. Alekseyev – "On the sign character of texts nd the levels of text description in quantitative linguistics" – pre-nted an analysis of the peculiarities of natural sign systems and a ultilevel framework for generalizing linguistic information. The pa-er by Ju. A. Tuldava – "Some theoretical problems of statistical text rganization" – gave a survey of the various theories and methods of resent-day quantitative linguistics. The author analysed over ten dif-rent analytical formulae designed for the description of rank distribu-ons – variants of Zipf's law – and underscored the urgent need for an ntological *substantiation* of the law. He also dwelt on the necessity of

unifying the terminology used in quantitative-linguistic text analysis.

The seminar then continued in the form of three working-groups. The first – "General problems of the statistical organization of lexis" – began with the report of G. Ja. Martynenko "On Gaussian and non-Gaussian distributions in philology". The author demonstrated that there is no ground for treating Zipf's law as a "paradigm of linguistic distributions": instability of parameters and the increase of dispersion with the increase of sample size also characterize many natural phenomena. An important feature of such distributions consists in the *differentiation of elements* according to their *function in the text.*

Ju. K. Krylov, in his paper "Text generation as a weakly stationary random process", showed that in terms of the theory of random processes text generation can, in the first approximation, be described as *autoregression of the 2nd – 3rd degree*, irrespective of sample size.

Ju. K. Orlov, in his report "On some applications of the generalized Zipf-Mandelbrot law", described certain specific features of the model of statistically homogenous texts, constructed by the author on the basis of V. Kalinin's model, and making it possible to compare the statistical structures of *texts of different lengths* by "reducing" them to a "common length" (referred to the author as "Zipf-size").

V. E. Ostapenko, in his report "Quantitative text models", dealt with methods of selecting *terminological vocabulary* for use in information-retrieval systems analysing the lexical structure of text. An analysis of rank, as well as dispersional and parabolic distributions reflecting the growth of vocabulary size in relation to the increase of text length, enabled the author to classify almost half of the 1000 most frequent lexemes of his corpus.

The papers by M. V. Ovsjannikov and N. S. Manasjan were devoted to problems of modelling the frequency characteristics of various linguistic units on the basis of the *lognormal law.* L. M. Sutyagina, in her paper "On horizontal distributions", presented the results of a taxonomical analysis of frequency characteristics of minimum-length samples.

Considerable interest was evinced by the participants of the seminar in the report by A. N. Lebedev "Some ontological foundations of Zipf's law". Proceeding from the hypothesis that images are encoded in human memory in *packages of coherent waves of neuronic activity,*

nd from the supposition of an *approximate equality between the re-*
ative frequency of a word in speech and the relative frequency of the
ctivation of its image, the author suggested a novel formula for de-
cribing the dependence of the frequency F_i of a word in a text on the
ank i of the word (as usual in linguostatistics, words are ranked in
rder of non-increasing frequency): $F_i = Rq^{-1}ln(1 - q/i)$, where R is
he critical phase difference between the encoding waves of neuronic
ischarges and q – the range of the word's rank fluctuations.

The theme of the second working group was formulated as "Theore-
ical and applied aspects of quantitative vocabulary analysis". A. A. Po-
ikarpov's paper – "Polysemantic spectra" – presented the results of an
nvestigation of rank-polysemantic and spectral-polysemantic distribu-
ions in a large corpus of entries in Russian and English explanatory
dictionaries. As demonstrated by the author,the parameters of these
distributions depend on the *type of the dictionary* as well as the *de-*
gree of analyticity of the language (These ideas were expounded on in
more detail by the same author at the second seminar of the group).
Ye. A. Il'jušina, in her report "An analysis of the systemic organization
of artistic texts with the method of rank distributions", demonstrated
the possibility of using random numbers in the comparative analysis
of the rank structures of linguistic units. M. A. Marusenko, in his re-
port "The optimization approach to the analysis of complex linguistic
objects", dwelt on some unsolved problems of text attribution. The
author suggested a five-part classification of linguostatistical text fea-
tures and a method for measuring the *feature distance between words*
in Euclidian space.

Interesting results in the quantitative analysis of *musical texts* were
described in the paper by V. K. Detlovs on the quantitative regularities
of mode organization in melody. The author demonstrated that there
exist distinct *connections* between mode organization and the direction
of pitch movement in folk melodies.

The meetings of the third working group – "New problems in quan-
titative linguistics" – were introduced by A. A. Polikarpov with the pa-
per "Asymmetrical dualism of the language sign: the systemic-quanti-
tative aspect". Developing the ideas of S. O. Karcevskij about a *sys-*
temic interdependence between synonymy and polysemy, the author ad-
vanced the hypothesis that in languages where lexical units are cha-

racterized by a high degree of polysemy, synonymic relations between
these units are correspondingly *weaker*. Parameters of synonymic and
polysemic distributions prove to be, in a sense, *mutually complemen-*
tary. The paper by A. V. Malov, "A systemic analysis of the quan-
titative characteristics of English lexis", dealt with rank-polysemantic
distributions in English dictionaries. The speaker presented a number
of analytical formulae approximating these distributions.

V. A. Dolinskij in his report "The response-distribution as a func-
tion of the frequency and polysemy of the stimulus: An associative
experiment", described the results of an investigation of the rank dis-
tributions of response words in *free-association experiments* with Rus-
sians and Americans. As shown in the paper, the form of the distribu-
tion depends on the frequency and the polysemy characteristics of the
stimulus-word and is described by Zipf's law.

The report of V. P. Belyanin "A psycholinguistic approach to the
typology of artistic texts" presented a classification of texts of fiction
based on the psychological types of an individual's attitude towards the
world which determine the choice of particular lexico-semantic groups
for the utterance.

The paper by Tan'Aoshuan "The syntactic arrangement of texts in
amorphous languages and the linguistic convention" discussed the epis-
temological difficulties that arise in the process of formalizing certain
relatively uninvestigated peculiarities of Chinese.

Also heard at the seminar were reports by Yu. B. Safronova ("Some
systemic-quantitative characteristics of synonymy, polysemy and anto-
nymy in an English dictionary of synonyms"), O. V. Bushuyeva ("Some
systemic-quantitative correlational regularities of the polysemic charac-
teristics of words in text and in vocabulary (on the basis of English")),
A. M. Karapetiants ("On the rhythmical organization of communica-
tion and text in Chinese") and O. V. Barakin ("Text perception by a
personality and by a social group").

During the work of the seminar a round-table discussion was held
on the problem "Zipf's law at the present stage of the development of
quantitative linguistics". Attention was focussed on a number of live
issues: – What is the *ontological content* of Zipf's law? – What are the
criteria of its fulfilment? – What are the *differences and similarities*
between the *natural and social phenomena* that correspond to the law?

It was emphasized by the speakers that the accumulation and elaboration of mathematical formulae describing the linguistic reality, important though it may seem, should not become an end in itself.

To sum up, the first seminar showed that it is both natural and expedient to combine the efforts of scholars of different fields in the study of qualitative and quantitative aspects of text organization.

The second seminar was arranged by the group of structural linguistics of Tartu State University. The participants included members of the group as well as guests. About 40 papers were heard, dealing with problems of the structural organization of text and vocabulary, linguistic distributions, the dialogue, the classification and automatic analysis of texts, and quantitative musicology. A number of speakers developed the ideas presented in their papers at the first seminar.

The plenary session was devoted to papers by Ju. M. Lotman, R. G. Piotrowski, A. N. Lebedev, A. V. Zubov.

The contribution by Ju. M. Lotman described three major functions of text: conveyance of constant information, generation of new information, cultural memory. R. G. Piotrowski spoke about the antinomies between natural and computer languages in automatic text processing and about possible ways of overcoming or "by-passing" them by means of an analysis, deeper than any carried out so far, of the communicative-semiotic aspects of texts. A. N. Lebedev carried on with the presentation of his psycho-physiological model of quantitative text organization, with the central notions of periodical processes as the basis of memory and the word as a set of consecutive packages of neuronic impulses. A. V. Zubov described the static and dynamic components of text and outlined methods for analysing their form. A number of questions raised at the plenary session became the subject of further discussion in the working groups.

The central report of the working group "Linguistic distributions" was delivered by Ju. A. Tuldava ("On the interpretation of linguistic distributions"). The author outlined a general approach to the analysis of linguistic distribution as a specific object, showed the importance of distinguishing the mono-objective, the poly-objective and the multidimensional (complex) types of distribution, characterized three levels of distribution analysis (form, content and origin – the genetic aspect). Special attention was paid to *quantitative evaluation of a linguistic object.* As the author demonstrated, insufficient regard to the content aspect of such evaluation may entail serious *mistakes* in the quantitative analysis of text organization. Also heard in the section were reports by G. G. Sil'nitskij about the significance of correlation between verbal text characteristics in determining the text's degree of individuality. and by N. S. Manasyan about the analysis of a linguistic distribution on the basis of Russian-language software texts.

The section "Classification and automatic analysis of text" was devoted to problems of automatic text processing with micro-computers (the report by K. J. Lepa), the measurement of the readability and reading difficulty of texts (reports by J. A. Mikk and L. I. Vassil'čenko, H. Liiv, K. Soomere), the logico-semantic structure of English science and technology texts (the report by A. A. Larina), methods of automatic text attribution (the report by G. V. Yermolenko), classificational tasks of quantitative linguistics (the report by G. Ja. Martynenko), the use of factor analysis in text classification (the report by Ju. A. Tuldava) and a quantitative approach to the analysis of free-association processes of normal persons and of the mentally ill (the report by M. G. Boroda and V. E. Paškovskij). All the reports were characterized by endeavours to find new approaches to the problems under investigation. Thus Ju. A. Tuldava proposed a novel method of *factor analysis* of essential quantitative-linguistic text characteristics and showed its efficiency in solving problems of *classifying fictional texts.* The report by G. V. Yermolenko demonstrated the important role of high-frequency and low-frequency words in the *stylistic individuality* of the text.

New ways of measuring the *readability and reading difficulty* of instructional texts were described in the reports by H. P. Liiv, J. A. Miik and L. I. Vassil'čenko. An efficient approach to the study of *cloze procedure tests* was suggested in the report by K. E. Soomere. K. J. Lepa – discussed some significant limitations in the *processing of texts with microcomputers*, A. A. Larina – the possibility of computer-modelling the logico-semantic structure of English scientific texts on the basis of their *paragraph strucure.*

In their joint report, M. G. Boroda and V. E. Paškovskij showed the significance of, and presented a method for, the analysis of the *rhythmical characteristics of free-association flows for diagnostic purposes* as

well as for the study of text generation processes. The characteristics in question are connected with the emergence of *chains of words* equal in length or stressed in the same place, etc.

The fundamental issues of quantitative text organization were taken up in two working groups – "Statistical text organization" and "Structural organization of text and vocabulary". In the former there arosed an animated discussion of the alternative mathematical models of lexical-statistical text organization proposed by M. V. Arapov, Ju. K. Krylov and Ju. K. Orlov. As was pointed out by the participants in the discussion, the first two models have an important asset in that they seek to establish a connection between hypotheses about the mechanisms generating quantitative regularities in texts and certain intuitively appealing general *qualitative* principles: the principle of distribution of words into *classification cells* (M. V. Arapov) and the principle of *symmetry* between the functional properties of elements in language and in coherent text (Ju. K. Krylov). It was emphasized during the discussion that in these two models lexical distributions (e. g. Zipf's distribution) are *deduced* from certain simple principles and not postulated, which makes them valuable both from the ontological point of view and from that of constructing an adequate *dynamic* model of coherent, specifically artistic texts.

Other reports heard in the group included those by V. N. Byčkov about the problem of "optimum sample size" (conceived by the author as the size on which the text follows the Zipf-Mandelbrot law in its canonical form) and an attempt to solve it on the basis of a non-linear Zipf model, that by S. I. Gindin on his multi-purpose system of quantitative text parameters, and that of I. Š. Nadarejšvili on the statistical structure of Rusthaveli's epic "The Knight in the Panther's Skin".

The working group "Structural organization of text and vocabulary" was concerned with problems of quantitative investigations of the *systemic* character of text and vocabulary organization as well as with problems of defining structural units and specifying criteria for their classification. These problems were dealt with by A. A. Polikarpov, who in his paper presented the results of his studies of polysemic distributions in explanatory dictionaries, and showed that a number of parameters of these distributions are sensitive to *typological differences between languages* and, on the other hand, to the *type of the explanatory*

dictionary (see p. 4). The problem of structural units was discussed in M. D. Jakubovskaja's paper on the dependence of the frequency characteristics of a text unit based on the *criteria* used in determining the unit. The joint contribution of V. I. Perebejnos, T. A. Grjaznukhina, M. P. Muravitskaja, N. P. Darčuk and L. I. Komarova described the results of investigations of the structural-quantitative organization of *synoptical* texts.

The papers by L. V. Orlova and I. A. Boldak were concerned with the large textual units (superphrasal unities and paragraphs) in a scientific text, and the problems of their quantitative analysis. Relations between quantitative distribution characteristics of text units and certain qualitative characteristics of language and text were dealt with in contributions by S. V. Raitar ("Quantitative estimation of the contextual determination of the Estonian word") and S. P. Klyavina ("A linguo-statistic comparison of functional styles"). In the subsequent discussion the great importance of further research into the problem of natural units of text structure and natural criteria for their classification was emphasized. The point was made that the solution of the problem requires large-scale systematic purposeful studies of quantitative text organization.

A special working group was devoted to current problems of the *quantitative theory of dialogue*. The topics under discussion comprised the optimal organization of modern linguistic processors and in this connection the typological classification of dialogues (B. Ju. Gorodetskij), the construction of formalized models of the behaviour of communicative interaction participants (V. M. Sergeyev and M. A. Sivertsev), general principles of modelling man-computer natural language dialogues (H. J. Oim, M. E. Koit and T. A. Roosmaa), and dialogue as a specific type of text (G. V. Andrusenko). The discussion revealed a close connection between the problems confronting contemporary studies of dialogue and the general issues of quantitative/qualitative text analysis and, hence, the usefulness of co-operation, exchange of ideas and methods between experts from the two fields.

Lively interest among the seminar participants was aroused by the working-group "Problems of quantitative musicology", where both the papers and discussions testified to a remarkable diversity of research interests in, as well as the determination to tackle the fundamental

problems of, *the field of musical language and text analysis.*

V. K. Detlovs devoted his report to the problem of the quantitative analysis of mode. Working with stylistically varied material, the author traced certain *metastylistic regularities* and showed the "stage-managing" role of *mode functionality* in regard to pitch "behaviour". The material investigated belonged to a broad group of folk and composed melodies of different styles.

The problem of correlation between "language" and the "textual" on the level of n-note melodic sequences (referred to by the authors as "intonations") – a certain analogue of n-grams – was dealt with in the paper by I. V. Bakhmutova, V. D. Gussev, and T. N. Titkova, which described *machine recognition methods* for "common" and "original" intonations and presented a list of composers arranged in order of the "common"/"original" intonation-proportion in their works.

The language aspects of musical thinking were discussed in the paper of M. G. Boroda, who presented an analysis of general and evolutionary principles as well as stylistic characteristics of the *length distribution* of melodic units of different hierarchical levels (defined by the author), such as the "Formal (F-) motife" and the "Metrico-Rhythmical (MR-) segment", etc. The speaker advanced the hypothesis that European music during the last four centuries *has grown increasingly analytical* and that the development and alternation of musical styles within this general process can be described by postulating the existence of "synthetical explosions" (when musical language of a given style, etc., becomes *synthetical*, in a linguistical sense of this term) and "*analytico-synthetic waves*".

Of considerable interest was also the paper by J. Ross and I. Rüütel about *automatic notation* of monophonic tunes. The authors solve this urgent problem of quantitative musicology with methods close to those used for establishing the *fundamental tone frequency in speech.*

The discussions dealt not only with the papers given, but also with such related topics as musical text segmentation, establishment of natural and strictly defined musical units as well as natural criteria for their classification and the use of methods of quantitative linguistics in musical text and musical language studies.

The sessions of the working groups were followed by a general discussion of the major papers of the seminar and a round-table discussion

of a number of unsolved problems connected with interpretations of and investigations based on Zipf's law. The discussion, a continuation of the debate begun at the first seminar of the group, revealed a need for a fundamental elaboration of problems related to the *level* of investigating quantitative regularities in text – in particular, to the *rank of structural units* and the *type of criterion used for their classification* (distinction). The discussion also brought to light the urgency of terminological unification in quantitative linguistics and the significance of systemic-quantitative organization investigations of non-linguistic texts for understanding the principles underlying such regularities as e. g. Zipf's law and, more broadly, for understanding the principles of text generation.

The final plenary session adopted a concluding statement prepared by the organizing committee which summarized the principal results of the seminar and set forth recommendations for the group's further work.

The authors would like to express their gratitude to Dr. K. Soomere who translated the text into English.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 147–169

Verteilungen der Satzlängen*

G. Altmann
Bochum

> It seems likely that the less that goes into one's writing the more stereotyped it becomes, and possibly becomes a Markov process when one is almost asleep.
>
> (Good 1965:227)

Satzlänge ist eine linguistische Größe, die nicht mechanisch zu bestimmen ist. Während in der gesprochenen Sprache die Grenzen des Satzes vage sind, ist die Situation in der geschriebenen Sprache etwas besser, jedoch nicht eindeutig. Die Textautoren deuten die Grenze durch Konventionen an, aber oft haben sie selbst Probleme damit. Die gängigen Definitionen des Satzes, von denen es über 200 gibt, helfen bei der formalen Analyse keineswegs.

Satzlänge ist zugleich eine Größe, auf die die Texterzeuger weniger achten als auf die grammatische Struktur und Bedeutung des Satzes. Im Prinzip ist es möglich, unendlich lange Sätze zu bilden, aber in der Realität wird diese Möglichkeit beträchtlich eingeschränkt. In der gesprochenen Sprache tendiert der Sprecher dazu, die Sätze voneinander gar nicht zu trennen bzw. alles, was er sagen will, in einen einzigen Satz einzubauen. Der Hörer dagegen hat damit seine Dekodierungsprobleme: er muß das Gehörte ordnen, die Sätze und Teilsätze aufeinander beziehen, die Zusammenhänge im Kurzzeitgedächtnis speichern, usw. Damit die Kommunikation reibungslos verläuft, muß sich der Texterzeuger dem Hörer anpassen und eine Gliederung des Textes vornehmen. Diese Tätigkeit ist sicherlich unbewußt, jedoch nicht chaotisch: der Sprecher folgt einem Gesetz, das auf dem Satzniveau

* Diese Studie entstand im Rahmen des Projekts "Sprachliche Synergetik". Der Autor bedankt sich bei der **Stiftung Volkswagenwerk** für die freundliche Unterstützung.

die Kommunikationseffektivität steuert und zu einer Satzlängenverteilung führt, die wir unten herleiten wollen. Wir nehmen an, daß alle Spracheigenschaften im Text gesetzesartig verteilt sind und diese Gesetze aus der Interaktion (Kooperation und Konkurrenz) gewisser Faktoren, die als Ordnungsparameter (vgl. Haken 1978) wirken, ableitbar sind (vgl. Köhler, Altmann 1987). Im folgenden werden wir uns auf die Satzlängen beschränken.

Als Initiator der "Satzlängenforschung" wird Sherman (1888) bezeichnet (vgl. Bailey 1965:228). Seit 100 Jahren der Forschung wurden mehr als 3 Dutzend Arbeiten verfaßt, die uns nicht vollständig zugänglich waren. In diesen Arbeiten ging es vor allem darum,

(i) die Satzlänge als ein Stilcharakteristikum zu benutzen (vgl. Parker 1889; Yule 1939, 1944; Williams 1940, 1970; Wake 1957; Lesskis 1962, 1963, 1964; Markworth, Bell 1967; Morton, Levison 1966; Mistrík 1967; Clayman 1981);

(ii) als Kriterium zur Entscheidung über strittige Autorschaft eines Textes heranzuziehen (vgl. Yule 1939; Morton 1965; Morton, McLeman 1966; Brandwood 1969);

(iii) ein Modell der Satzlängenverteilung aufzustellen (Lesskis 1962; Buch 1969; Williams 1970; Sichel 1974).

Zahlreiche andere Probleme, die in der Diskussion entstanden (vgl. z.B. Spang-Hanssen 1963; Janson 1964; Diskussion zu Morton 1965; Parolková 1970), betrafen nie das Modellieren, sondern vor allem die Definition des Satzes. Die Kritik wendete sich hauptsächlich gegen die Bestimmung des Satzes in der griechischen Literatur, wo sie angeblich nicht so eindeutig ist wie eine statistische Ermittlung es erfordert. Vorwürfe dieser Art erinnern an die recht unfruchtbaren Diskussionen der 30-er Jahre darüber, was ein Phonem ist, ob es Phoneme tatsächlich gibt, was ein Wort ist usw.

Abgesehen davon, daß man in der gesprochenen Sprache nur mit Schwierigkeiten Sätze abgrenzen kann, ist es in der geschriebenen Sprache mit sehr wenigen operationalen Definitionen (Kriterien) nicht nur möglich, sondern auch hinreichend. Jeder der erwähnten "Satzlängenforscher" benutzte sicherlich irgendwelche (zumindest implizite) Kriterien, die man als Bedingungen für die Gültigkeit seiner Schlüsse betrachten kann.

Wir wollen uns daher hier keinen Satzdefinitionsproblemen widmen,

sondern die Resultate anderer Forscher zur Überprüfung des unten vorgeschlagenen Modells verwenden. Die Segmentierung eines Textes in Sätze bereitet kaum Schwierigkeiten in modernen Texten, auch wenn man über die Behandlung des Semikolons, der direkten Rede, der durch Ausrufezeichen abgetrennten Interjektionen usw. Entscheidungen treffen muß. Es geht aber immer um Entscheidungen, nicht um Tatsachen, denn Sprache kümmert sich nicht um Existenz bzw. klare Grenzen irgendwelcher Einheiten. Spracheinheiten sind mehr oder weniger vage Abstraktionen, je nach dem, ob sie vom Sprecher oder vom Linguisten betrachtet werden. In älteren Texten (tradierter Folklore) wird die Tatsache, daß es sich bei Satzsegmentierung um Entscheidungen des Sammlers handelt, viel deutlicher, aber nicht gravierender: der Sammler folgt einer orthographischen Konvention, auf die sich die Sprachgemeinschaft "geeinigt" hat.

Bei der Modellierung der Satzlängenverteilung kann also das Definitionsproblem durch Aufstellung von Kriterien gelöst werden. Dagegen ist nichts einzuwenden und wie es scheint, hat es keinen Einfluß auf das Modellieren, auch wenn unterschiedliche Forscher den Satz unterschiedlich definiert haben.

Was aber eine grundsätzliche Bedeutung haben kann, ist die Einheit, in deren Anzahl die Satzlänge gemessen wurde. Alle Forscher haben stillschweigend angenommen, daß die Satzlänge in Anzahl der Wörter gemessen werden soll. Dies ist nicht so natürlich, wie es betrachtet wurde, denn die Messung in kleineren Einheiten, wie Silben oder gar Lauten/Buchstaben, kann zu drastisch unterschiedlichen Resultaten führen, wie Clayman (1981) gezeigt hat. Die unmittelbare Komponente des Satzes ist eigentlich der Teilsatz (Clause) und daher müßte die Verteilung der Satzlänge in Anzahl von Wörtern anders – und komplizierter – sein als die in Anzahl von Clauses. Die Ebene der Teilsätze ist eine intervenierende Ebene (zwischen Satz und Wort), deren Realität sich auch bei dem Menzerathschen Gesetz eindeutig manifestiert (vgl. Altmann 1983). Daher wäre es praktischer, erst die Satz/Clause-Verteilung zu modellieren und durch ihre Modifikation, d.h. durch Inbetrachtziehung der intervenierenden Clause-Ebene, die Satz/Wort-Verteilung zu erhalten. Diesen Weg werden wir hier gehen.

Bisher gab es zwei Modelle für die Satz/Wort-Verteilung, nämlich die Lognormalverteilung und die *Sichel*-Verteilung. Die Lognormalver-

ung wurde wohl aufgrund ihrer Form angepasst und gilt als das eptierte Modell (vgl. Williams 1940; Lesskis 1962, 63, 64; Cox in rton 1965; Buch 1969, u.a.). Die Begründung liegt wahrscheinlich in, daß man für einen Text (Autor) einen festen Satzlängenmittvert als sein Charakteristikum annimmt und zufällige, "normale", och logarithmisch transformierte Abweichungen von diesem Mittelt vermutet.

Diese Vermutung ist problematisch aus zwei Gründen:

(i) Wie mehrere Autoren festgestellt haben, kann sich der Satzlängenmittelwert im Verlauf des Textes oft sehr beträchtlich ändern. Daher sind große Stichproben aus großen Texten oder mehrere zusammengefaßte Stichproben für die Überprüfung eines Modells nicht gut geeignet.

(ii) Die Annahme der völligen Zufälligkeit der Satzlängenerzeugung läßt keinen Platz für die Intentionen eines Textautors, für seine Rücksicht auf den Hörer (d. h. für die Wirkung der Zipfschen Kräfte), für die Störung der intervenierenden Ebene usw. Ein derartiges Modell ist dann letzten Endes in eine "Theorie der Längen von Spracheinheiten" nicht inkorporierbar, da es auf nichts Rücksicht nimmt und keine Faktoren (nur Zufall) in Betracht zieht.

Es ist eher anzunehmen, daß der Autor entweder eine implizite nntnis dessen besitzt, wie die Längen im Text verteilt sein müssen, mit er sein Kommunikationsziel erreicht, oder daß er durch gewisse ktoren zu einer bestimmten Längenverteilung gezwungen wird, oder des. Die Abweichungen vom Mittelwert sind dann nicht zufällig, ndern folgen einem Gesetz, und erst die Abweichungen von diesem setz können zufällig sein oder einem anderen Gesetz folgen. Nach Aussage von Good im Motto folgt ein Autor dem Grundgesetz ganz reotyp, wenn er keine bewußten Textgestaltungen beabsichtigt.

Einen anderen Weg ging Sichel (1971, 1974), der für andere Zwecke e ziemlich komplizierte Verteilung ableitete und feststellte, daß sie linguistische Zwecke gut geeignet ist. Nach Sichel kann man die tzlänge als eine Poisson-verteilte Variable betrachten, d.h.

$$f(x|r) = \frac{e^{-r} r^x}{x!}, \quad x = 0, 1, \ldots$$

keitsverteilung liefert. Wir werden hier den zweiten Weg wählen, weil er einfacher und übersichtlicher ist.

Wir nehmen an, daß jegliche Verteilung der Längen im Text gesetzesartig gestaltet wird und der Texterzeuger, sofern er den Text spontan erzeugt, diesen Gesetzen folgt. Um sie abzuleiten, reicht es, Annahmen über die Differenz zweier benachbarter Wahrscheinlichkeiten zu machen, d.h. über

$$P_x - P_{x-1} = \Delta P_{x-1}. \tag{1}$$

Diese Differenz ist nicht konstant, sondern hängt von dem jeweiligen P_{x-1} ab, sodaß wir besser den Quotienten

$$D = \frac{P_x - P_{x-1}}{P_{x-1}} = \frac{\Delta P_{x-1}}{P_{x-1}} \tag{2}$$

in Betracht ziehen. Überlegungen über D führten sowohl zum System von Pearson (s. Ord 1972), als auch zum System von Katz (1946, 1965) und Ord (1967) und unser Modell hat Überschneidungen mit dem Ordschen System. Die D gestaltenden Faktoren werden wir mit kleinen Buchstaben bezeichnen:

a – Wirkung des Sprechers (Stil, u.a.)
c – Wirkung des Hörers, der Sprachgemeinschaft, Rücksicht auf den Hörer u.a
] "Zipfsche Kräfte"
b – Textfaktoren
d – Faktor der Ebene

Diese Größen stellen Ordnungsparameter dar, in denen möglicherweise noch andere Faktoren enthalten sind, die wir aber nicht explizit kennen müssen. Wir können nun D folgendermaßen zusammensetzen:

Falls wir die Satzlänge in Clauseanzahl messen, dann ist

$$D = \frac{b - ax}{cx}, \tag{3}$$

wo a und b "gestaltend", während c "bremsend" wirkt; b wirkt "global", a und c "lokal". Der Sprecher versucht in die Bindung der Textsorte seinen eigenen Stil einzubringen, daher besitzt a negatives Vorzeichen.

ei r selbst eine Zufallsvariable mit eigener Wahrscheinlichkeits-
eilung ist. Die Findung dieser sekundären Verteilung ist schwer
ründbar und meistens – auch in anderen Wissenschaften – Sache
Glücks. Es ist aber eine fruchtbare Modellierungshilfe, die in der
guistik des öfteren benutzt wurde (vgl. z. B. Grotjahn 1982; Alt-
ın, Burdinski 1982). Sichel benutzt eine recht komplizierte Funk-
, nämlich

$$g(r) = \frac{1}{2} \cdot \frac{(2\sqrt{1-b}/ab)^c r^{c-1}}{K_c(a\sqrt{1-b})} \exp\left[-(1/b-1)r - \frac{a^2 b}{4r}\right]$$

$K_c(.)$ die modifizierte Bessel-Funktion zweiter Art der Ordnung c
so daß er schließlich die Verteilung von X als

$$f(x) = \frac{(1-b)^{c/2}}{K_c(a\sqrt{1-b})} \cdot \frac{(ab/2)^x}{x!} K_{x+c}(a)$$

ält. Als Satzlängenverteilung ist sie besonders geeignet, wie Sichel
74) zeigte, wenn $c = -1/2$ ist.

Sichels Verteilung bringt viele Vorteile mit sich, weil sie zahlreiche
ezial- oder Grenzfälle hat, die in der Linguistik bereits angewendet
rden, sodaß man sie nicht außer Acht lassen sollte, sondern ihre
gweite in der Linguistik weiter prüfen und ihre Interpretation ver-
hen sollte.

Im Folgenden werden wir versuchen, von linguistischen Annahmen
gehend, ein anderes Model für die Satzlängenverteilung abzuleiten,
sowohl die Satz/Clause als auch die Satz/Wort-Variante abdecken
nte.

Die Erzeugung von Sätzen ist ein (stochastischer) Prozess, in dem
Variable "Satzlänge", wie auch immer gemessen, einem Rhythmus
gt, der von einigen Faktoren – die wir in der Sprache der Synerge-
als Ordnungsparameter bezeichnen können – gesteuert wird. Als
sultat dieser Steuerung entsteht eine Wahrscheinlichkeitsverteilung
Satzlänge, die man auf zwei unterschiedliche Weisen ableiten kann.
tweder modelliert man den zugrundeliegenden stochastischen Pro-
s oder man stellt aufgrund linguistischer Annahmen die Differenzen-
ichung auf, deren Lösung die entsprechende diskrete Wahrscheinlich-

Falls wir die Satzlänge in Wortanzahl messen, kommt noch die Wirkung der intervenierenden (Clause-) Ebene dazu und wir erhalten

$$D = \frac{b - ax}{cx + d}. \tag{4}$$

Die Lösung von (3) ergibt:

$$\begin{aligned}
\frac{P_x - P_{x-1}}{P_{x-1}} &= \frac{b - ax}{cx} \\
P_x &= (1 + \frac{b - ax}{cx}) P_{x-1} \\
&= \frac{(c-a)x + b}{cx} P_{x-1} \\
&= \frac{c-a}{c} \cdot \frac{x + b/(c-a)}{x} P_{x-1}. \qquad (5)
\end{aligned}$$

Schreibt man $q = (c-a)/c$ und $k = b/(c-b)$ wobei $c > a > 0, b > 0$, sodaß $0 < q < 1, k > 0$, dann wird (5) zu

$$P_x = \frac{x+k}{x} q P_{x-1} \tag{6}$$

mit der Lösung

$$P_x = \frac{(k+1)(k+2)\ldots(k+x)q^x}{x!} P_0. \tag{7}$$

Um auf eine übliche Form zu kommen, setzen wir $r = k+1$ und erhalten die negative Binomialverteilung

$$P_x = \binom{r+x-1}{x} p^r q^x, \tag{8}$$

denn aus der Bedingung $\sum P_x = 1$ ergibt sich $P_0 = p^r (p = 1-q)$. Wie man sieht, besteht q nur aus der Interaktion des Sprechers und des Hörers, während k (bzw. r) auch den Faktor der Textsorte enthält.[1]

[1] Korrekter wäre der Ansatz $(D = b - c + a - ax)/cx$, wo im Zähler noch eine Interaktion des Sprechers und Hörers $(a - c)$ stattfindet, damit $r > 0$ sein kann; oder man läßt in (3) $b > -1$ zu.

Im Fall (4) haben wir

$$P_x = \left(1 + \frac{b - ax}{cx + d}\right) P_{x-1}$$

$$= \frac{(c-a)\,x + b + d}{cx + d} P_{x-1}$$

$$= \frac{c-a}{c} \cdot \frac{x + \frac{b+d}{c-a}}{x + \frac{d}{c}} P_{x-1}$$

Mit $q = (c-a)/c$, wo $c > a > 0, k = (b+d)/(c-a) > m > d/c > 0$ erhalten wir

$$P_x = \frac{k+x}{m+x} q P_{x-1} \tag{9}$$

mit der Lösung

$$P_x = \frac{(k+1)(k+2)\ldots(k+x)}{(m+1)(m+2)\ldots(m+x)} q^x P_0, \quad x = 0, 1, \ldots \tag{10}$$

Setzt man wieder (vgl. Anmerkung 1) $K = k+1$, $M = m+1$, dann erhält man

$$P_x = \frac{K(K+1)\ldots(K+x-1)}{M(M+1)\ldots(M+x-1)} q^x P_0, \tag{11}$$

was man auch als

$$P_x = \frac{K^{(x)} q^x}{M^{(x)}} P_0, \quad \text{mit } A^{(0)} = 1 \tag{12}$$

oder

$$P_x = \frac{\binom{K+x-1}{x}}{\binom{M+x-1}{x}} q^x P_0 \tag{13}$$

schreiben kann, wobei man P_0 als

$$P_0 = [{}_2F_1(K, 1; M; q)]^{-1} \tag{14}$$

definieren kann, wo ${}_2F_1$ die hypergeometrische Funktion darstellt

$${}_2F_1(K, 1; M; q) = 1 + \frac{K}{M} q + \frac{K(K+1)}{M(M+1)} q^2 + \ldots . \tag{15}$$

Formel (11) bis (14) stellt die *Hyperpascal-Verteilung* dar. Man sieht, daß mit $M = 1 (d = 0)$ die Hyperpascal-Verteilung zur negativen Binomialverteilung wird.

Sollte die empirische Überprüfung (s.unten) diese zwei Modelle unterstützen, so möchten wir sie als *Sherman-Gesetze* bezeichnen, da Sherman als erster die Satzlängen ins Spiel brachte.

Alle Textgesetze haben bedauerlicherweise die Eigenschaft, fast immun gegen Falsifizierung zu sein. Dies rührt von der Tatsache her, daß sie nur für spontan, "in einem Zug" erzeugte oder vollständige Texte gelten. Die Vollständigkeit ist z.B. für das Zipf-Mandelbrotsche Gesetz von Bedeutung (vgl. Orlov, Boroda, Nadarejšvili 1982), die Spontanität und kontinuierliche Erzeugung ist z.B. für den Satzlängenrhythmus von Belang. Wird das obige Modell an einem Text bekräftigt, dann nimmt man automatisch an, daß die Erzeugungsbedingungen erfüllt wurden; ist das aber nicht der Fall, dann kann man immer behaupten, daß an dem ursprünglichen, spontanen Text der Autor oder die Redakteure einiges geändert haben, wodurch eine signifikante Fluktuation der Variablenwerte entstand. Dies ist aber durchaus berechtigt, wenn man bedenkt, wie ein Text entsteht. Bedingungen solcher Art (z.B. Einwirkung anderer Kräfte) werden ja auch in der Mechanik durch Newtons Axiome nicht ausgeschlossen.

Um aber nicht zu solchen Immunisierungsmaßnahmen greifen zu müssen, empfiehlt es sich bei der Stichprobenerhebung folgendes zu beachten:

(1) Da sich der Satzlängenrhythmus im Text ändern kann, empfiehlt es sich, keine Zufallsstichproben aus einem Text zu nehmen, sondern in einem geschlossenen Texteil die Sätze nacheinander zu messen.

(2) Stichproben möglichst nicht zusammenfassen. Es ist anzunehmen, daß eine größere Stichprobe, die aus mehreren Stellen eines

Texts genommen wurde, dem Modell eher folgt als eine Stichprobe, die aus mehreren Texten genommen wurde. Mittelung ist hier also fehl am Platz. Leider findet man in der Literatur vor allem Stichproben dieser Art. Es ist nur der Homogenität des Autors oder des Ganzen zu danken, wenn das Modell trotzdem paßt.

Paßt das Modell in solchen Fällen nicht, so muß es so erweitert erden, daß es der Stichprobenmischung Rechnung trägt. Dies kann ntweder durch Mischung von Verteilungen oder, wie Sichel verfuhr, urch Zusammensetzung von Verteilungen, indem man einen Parameer in (3) oder (4) als eine Zufallsvariable betrachtet, oder durch beide Verfahren, erreicht werden. Dies wird sich im Laufe der Forschung eigen.

Für die Plausibilität des obigen Modells spricht – per analogiam – uch der Umstand, daß Grotjahn (1982) für die Länge der Wörter gemessen in Silben auch die negative Binomialverteilung auf eine andere Weise erhielt. Es bietet sich der verallgemeinernde Schluß an, daß die Länge einer beliebigen sprachlichen Einheit, gemessen in Einheiten der unmittelbar niedrigeren Ebenen, der negativen Binomialverteilung folgt, während die Einschaltung einer intervenierenden Ebene zu der Hyperpascal-Verteilung führt.

Diese Hypothese müßte natürlich auf breiter Basis geprüft werden. Fraglich ist auch, was geschieht, wenn zwei intervenierende Ebenen zwischengeschaltet werden, z.B. wenn man die Satzlängen in Silbenzahl mißt (vgl. Clayman 1981). Muß man dann in (3) bzw. (4) Produkte von linearen Funktionen ansetzen oder die Ebenen als neue Variablen betrachten? Am problematischsten ist natürlich die Frage, ob es sinnvoll ist, über zwei oder mehr intervenierende Ebenen hinweg nach Abhängigkeiten zu suchen. Auf diese Frage wird man nur in der Zukunft eine Antwort geben können.

Zur Überprüfung der beiden Modelle (8) und (13) braucht man Schätzer der Parameter und empirische Daten. Die Schätzungsmethoden für die negative Binomialverteilung findet man in Johnson, Kotz (1968:131-135), wir benutzten sie nach Bedarf nur als Anfangswerte der Optimierung zum Erreichen des kleinsten Chiquadrats bei der Anpassung. Zur Optimierung haben wir den Simplex von Nelder und Mead oder den einfachen Algorithmus von Hook und Jeeves (s. Himmelblau

1972) verwendet.

Bei der Hyperpascal-Verteilung, die in der statistischen Literatur selten erwähnt wird (s. Ord 1972:88; Katz 1946) kann man Anfangsschätzer aus den Quotienten der ersten vier Wahrscheinlichkeiten (P_0 bis P_3) gewinnen oder aus den Momenten, die man aus der Rekurssionsformel

$$P_x = \frac{K+x-1}{M+x-1} q P_{x-1} \tag{16}$$

erhalten kann, nämlich

$$\mu' = \frac{Kq + (1-P_0)(1-M)}{p}$$

$$\mu'_{r+1} = \left[\mu'_r (1-M) + qK \sum_{i=0}^{r} i^{(r)} \mu'_i + q \sum_{i=0}^{r-1} i^{(r)} \mu'_{i+1} \right] / p \tag{17}$$

Zahlreiche Rechnungen haben gezeigt, daß die iterativen Optimierungsmethoden immer zu besseren Resultaten führen als die klassischen Punktschätzungen. Weiter hat sich herausgestellt, daß es für die drei Parameter der Hyperpascal-Verteilung immer mehrere Lokalminima gibt, wobei es meistens von den Anfangswerten abhängt, gegen welches Minimum die Prozedur strebt. Die Anpassung durch Optimierung ist also eine "trial-and-error" Prozedur, während die klassischen Punktschätzungen meistens unbefriedigende Resultate lieferten.

Für die Überprüfung der Satz/Wort-Verteilung (Hyperpascal) haben wir die gesamte, uns zugängliche Literatur ausgewertet. Die mit * gekennzeichneten Daten zeigen eine signifikante Abweichung von dem Modell. Der Chiquadrat-Anpassungstest wurde unter der Bedingung durchgeführt, daß die theoretische Häufigkeit einer Klasse größer als 1.00 war. In einigen Fällen wurden jedoch starke Zusammenfassungen von Klassen vorgenommen. Die Originaldaten sind in der Literatur, die in der Spalte "Quellen" angegeben ist, zu finden.

Drei "gute" Anpassungen findet man als Illustration in der Tabelle 1.

Zu den Resultaten ist folgendes zu bemerken: In den Quellen wurden die Häufigkeitsklassen in Intervalle 1-5, 6-10 usw. zusammengefaßt.

Tabelle 1
Einige Anpassungen der Hyperpascal-Verteilung
an die Satzlängen (Satz/Wort- Variante)

	Paulus: Brief an die Philiper		Herodot: Buch 1		Philo Judaeus: Werk 1	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	9	8.68	16	16.08	7	7.77
1	26	28.48	35	36.19	27	25.58
2	31	22.16	47	43.05	42	34.80
3	15	15.83	39	37.82	33	34.86
4	5	10.23	24	27.62	21	29.66
5	4	6.65	16	17.77	21	22.77
6	4	4.26	11	10.42	12	16.28
7	3	2.71	5	5.69	14	11.05
8	2	1.71	4	2.93	13	7.20
9	2	1.07	1	1.44	2	4.55
10	1	0.67	1	0.68	4	2.80
11	-	-	1	0.31	1	1.69
12	-	-	-	-	3	1.00
K		0.3770		5.9987		3.1217
M		0.0694		0.8277		0.4553
q		0.6043		0.3106		0.4802
χ^2		8.50		1.72		13.57
FG		6		6		8
P		0.20		0.43		0.09

Zwecks Anpassung haben wir sie durch Transformation $(x-5)/5$, wo x die obere Intervallgrenze war, auf die Skala $0, 1, 2, \ldots$ überführt.

Die meisten Daten stammen aus dem Griechischen, da dieses Problem die klassischen Philologen besonders interessiert. Yule (1939) hat einige englische Daten gesammelt, eine Verteilung ist aus dem Slowakischen (Mistrík 1967). Wie man in der Tabelle 2 sieht, gibt es in 245 Fällen insgesamt 22 signifikante Abweichungen von dem "Sherman-Gesetz" auf der $\alpha = 0.05$ Ebene. Von diesen 22 kann man 5 Daten des Brown-Corpus (Nr. 182, 185, 186, 191, 192) als für den Test irrelevant betrachten, da wir von vornherein angenommen haben, das zusammengesetzte Stichproben nur unter sehr günstigen Umständen dem Gesetz folgen (Homogenität der Texte). Das Gleiche gilt für die Essays von Bacon (Nr. 200), Essays von Macaulay (Nr. 210), Gesamtzählung von Ilias (Nr. 86), Gesamtzählung von Lamb (Nr. 207), und die drei Datensätze von De Imitatione Christi (Nr. 211-213), wo man nicht weiß, wer der Autor war, ob es mehrere Autoren gab oder ob starke Änderungen eines Textes durchgeführt wurden. Es gibt also nur 9 Datensätze, die als Falsifikation bezeichnet werden können (Nr. 96, 106, 128, 142, 156, 205, 206, 228, 232). Da uns der Ursprung und die Entstehungsgeschichte dieser Daten nicht bekannt ist, werden wir keine "Rechtfertigungen" suchen (vgl. dafür z.B. Morton, McLeman 1966). Wir schließen jedoch auch Druckfehler in den Daten nicht aus. Ein offensichtlicher Druckfehler ist in Wakes (1957) Zählung von Aristoteles "De Generatione" zu finden, wo für $x = 36 - 40$ statt 28 nur 8 stehen soll, wie die Gesamtsumme zeigt. Einen Druckfehler vermuten wir auch in den Daten von Aristoteles "Historia Animalium VI". Wie auch immer, gibt es in den 245 Fällen nur 9 Fälle, die mit dem Modell nicht übereinstimmen, was durchaus zulässig ist (3,7%).

Zur Überprüfung der Clause-Variante des Sherman-Gesetzes haben wir einige Stichproben aus Texten in mehreren Sprachen ausgezählt, wobei die Sätze nicht zufällig genommen, sondern nacheinander untersucht wurden. Die Daten und Resultate der Anpassung der negativen Binomialverteilung sind in der Tabelle 3 enthalten. Als Clause wurde praktisch der Teil des Satzes betrachtet, der ein finites Verbum enthielt – auch wenn es durch eine Ellipse vertreten wurde. In mehreren Fällen mußten wir entscheiden, ob ein Semikolon Satzende bedeutet, wie Ausrufezeichen u.a. zu deuten sind. Andere Forscher würden wahrscheinlich andere Werte bekommen, aber der Unterschied wäre wohl nur an kleinen Differenzen der Parameter zu bemerken.

Wie man sieht, folgen alle Daten der negativen Binomialverteilung, sodaß wir das Modell vorläufig akzeptieren können. Es bedarf natürlich noch zahlreicher Überprüfungen, bis das Gesetz fest etabliert ist, und umfangreicher Analysen, bis man herausbekommt, wie sich die Werte der Parameter bei bestimmten Textsorten, Autoren u.a. verhalten. Die obige Untersuchung ist nur der erste Impuls für die weitere Forschung.

Literatur

Altmann, G., Burdinski, V. (1982). "Towards a law of word repetitions in text-blocks". *Glottometrika* 4: 147-167.

Altmann, G. (1983). "H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz". In Faust, M., Harweg, R., Lehfeldt, W., und Wienold, G. (Hrsg). 1983. *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.* Tübingen: Narr. 31-39.

Bailey, R.W. (1969). "Statistics and style: A historical survey". In Doležel, L., Bailey, R.W. 1969. 217-236.

Brandwood, L. (1969). "Plato's seventh letter". *Revue de l'organisation international pour l'étude de langues anciennes par ordinateur* 4: 1-25.

Buch, K.R. (1969). "A note on sentence-length as random variable". In Doležel, L., Bailey R.W. 1969. 76-79.

Clayman, D.L. (1981). "Sentence length in Greek hexameter poetry". In Grotjahn, R. (Hrsg). *Hexameter Studies.* Bochum: Brockmeyer. 107-136.

Doležel, L., Bailey, R.W. (1969). (Hrsg). *Statistics and style.* New York: Elsevier.

Good, I.J. (1965). *Diskussion zu Morton (1965).* 225-227.

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

Himmelblau, D.M. (1972). *Applied nonlinear programming.* New York: McGraw Hill.

Janson, T. (1964). "The problem of measuring sentence-length in classical texts". *Studia Linguistica* 18: 26-36.

Johnson, N.L., Kotz, S. (1969). *Discrete distributions.* New York: Houghton Mifflin.

Katz, L. (1946). "On the class of distributions defined by the difference equation $(x + 1)f(x + 1) = (a + bx)f(x)$ (abstract)". *Annals of Mathematical Statistics* 17: 501.

Katz, L. (1965). "Unified treatment of a broad class of discrete probability distributions". In Patil, G.P. (Hrsg). 1965. *Classical and contagious distributions.* Calcutta: Statistical Publishing Society. 175-182.

Köhler, R., Altmann, G. (1987). *Synergetic modelling of language phenomena.* (In press).

Kučera, H., Francis, W.N. (1967). *Computational analysis of present-day American English.* Providence, Rhode Island: Brown University Press.

Leed, J. (Hrsg). (1966). *The computer and literary style.* Kent, Ohio: Kent State University Press.

Lesskis, G.A. (1962). "O razmerach predloženij v russkoj naučnoj i chudožestvennoj proze 60-ch godov XIX v". *Voprosy jazykoznanija* 11,2: 78-95.

Lesskis, G.A. (1963). "O zavisimosti meždu razmerom predloženija i charakterom teksta". *Voprosy jazykoznanija* 12,3: 93-112.

Lesskis, G.A. (1964). "O zavisimosti meždu razmerom predloženija i ego strukturoj v raznych vidach teksta". *Voprosy jazykoznanija* 13,3: 99-123.

Markworth, M.L., Bell, L.M. (1967). "Sentence-length distribution in the corpus". In Kučera, H., Francis, W.N. (Hrsg). 1967. 368-405.

Mistrík, J. (1967). "Dĺžka vety pri štylistickej charakteristike". *Slovenská reč* 32: 19-25.

Morton, A.Q. (1965). "The authorship of Greek prose". *Journal of the Royal Statistical Society* A 128: 169-233.

Morton, A.Q., Levison, M. (1966). "Some indicators of authorship in Greek prose". In Leed, J. (Hrsg). 141-179.

Morton, A.Q., McLeman, J. (1966). *Paul, the man and the myth.* London: Hodder and Stoughton.

Ord, J.K. (1967). "On a system of disrete distributions". *Biometrika* 54: 649-656.

Ord, J.K. (1972). *Families of frequency distributions.* London: Griffin.

Orlov, Ju.K., Boroda, M., Nadarejšvili, I.Š. (1982). "Sprache, Text, Kunst". *Quantitative Analysen.* Bochum: Brockmeyer.

Parker, H.A. (1889). "Curves of literary style". *Science* 13 (Nr. 321).

Parolková, O. (1970). "Determinace substantiva a délka vety". *Filologické Studie* 2: 29-39.

Sherman, L.A. (1988). "Some observations upon the sentence-length in English prose". *University of Nebraska Studies* 1: 119-130.

Sichel, H.S. (1974). "On a distribution representing sentence-length in prose". *Journal of the Royal Statistical Society* A 137: 25-34.

Sichel, H.S. (1971). "On a family of discrete distributions particularly suited to represent long-tailed data". In Laubscher, N.F. (Hrsg). *Proceedings of the Third Symposium on Mathematical Statistics.* Pretoria: CSIR: 51-97.

Spang-Hanssen, H. (1963). "Sentence length and statistical linguistics". In *Structures and Quanta.* Copenhagen: Munksgaard. 58-72.

Wake, W.C. (1957). "Sentence-length distribution of Greek authors". *Journal of the Royal Statistical Society* A 120: 331-346.

Williams, C.B. (1969). "A note on the statistical analysis of sentence-length as a criterion of literary style". In Doležel, L., Bailey, R.W. (Hrsg). 69-75. (= *Biometrika* 31: 356-361).

Williams, C.B. (1970). *Style and vocabulary: Numerical studies.* London: Griffin.

Yule, G.U. (1939). "On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship". *Biometrika* 30: 363-390.

Yule, G.U. (1944). *The statistical study of literary vocabulary.* Cambridge: University Press.

Tabelle 2
Anpassung der Hyperpascal-Verteilung
an Satzlängenverteilungen (Wort-Variante)

	Text	Quelle	K	M	q	X^2	FG	P
1	Herodot: Buch 1	Morton/Levison 1966	5.9987	.8277	.3106	1.72	6	.94
2	2	-"-	1.7383	.1414	.5275	5.33	6	.50
3	3	-"-	.3997	.0401	.6630	7.93	8	.44
4	4	-"-	2.5904	.2360	.4185	6.18	7	.52
5	5	-"-	.6777	.1196	.6067	4.95	9	.84
6	6	-"-	3.9842	.4468	.4067	10.84	8	.21
7	7	-"-	1.0770	.1535	.5784	5.83	8	.67
8	8	-"-	1.3763	.1595	.5221	1.57	8	.99
9	9	-"-	2.6499	.3928	.4289	9.59	6	.14
10	Gesamt 1-9	-"-	1.3607	.1800	.5321	6.11	10	.81
11	Thukydides: Buch 1	-"-	1.4555	.1708	.6158	13.97	10	.17
12	2	-"-	.6461	.6081	.6889	10.46	12	.58
13	3	-"-	.9488	.0947	.6380	8.75	10	.56
14	4	-"-	7.4964	.9219	.3534	9.78	7	.20
15	5	-"-	.7094	.0751	.6844	3.64	10	.96
16	6	-"-	1.4469	.2319	.6541	14.90	9	.09
17	7	-"-	1.0563	.1566	.6687	5.57	12	.94
18	8	-"-	1.2638	.3784	.6895	17.04	13	.20
19	Gesamt 1-8	-"-, Sichel 1974	.9779	.1430	.6659	38.85	18	.003*
20	Philo Judaeus: Werk 1	-"-	3.1217	.4553	.4802	13.57	8	.09
21	2	-"-	1.0144	.1602	.6191	15.43	10	.12
22	3	-"-	.5887	.1743	.7419	9.76	15	.83
23	4	-"-	.9686	.1637	.6300	14.20	10	.16
24	5	-"-	2.5719	.5915	.5067	10.75	9	.29
25	6	-"-	.9217	.2709	.6664	6.93	6	.33
26	Gesamt 1-6	-"-	1.3235	.3198	.5146	9.58	5	.09
27	Lysias: Werk 1	-"-	1.3717	.3349	.5839	5.62	8	.69
28	2	-"-	.5474	.0002	.7477	7.83	7	.35
29	3	-"-	1.5802	.0002	.5461	2.49	7	.93
30	7	-"-	.6693	.0501	.6650	6.88	7	.44
31	6	-"-	2.8875	.5992	.4693	7.14	3	.07
32	12	-"-	.6168	.1506	.6633	7.35	10	.69
33	13	-"-	.6520	.0966	.6114	2.28	10	.99
34	14	-"-	1.2864	.1711	.6245	7.29	8	.51
35	19	-"-	1.0365	.1175	.5636	12.09	8	.15
36	20	-"-	.2218	.0003	.6365	9.25	5	.10
37	Demosthenes: Werk 3	-"-	2.1859	.9429	.5723	8.19	7	.32
38	4	-"-	.7660	.5755	.7572	19.80	12	.07
39	6	-"-	1.1589	1.0282	.7924	8.48	8	.39
40	7	-"-	1.3590	.3191	.6213	4.79	7	.69

Tabelle 2
Fortsetzung

	Text	Quelle	K	M	q	X^2	FG	P
	Demosthenes: Werk	Morton/Levison 1966						
41	8	-"-	.7509	.3909	.7158	11.59	12	.48
42	9	-"-	.3918	.1874	.7203	7.67	11	.74
43	10	-"-	.3062	.0827	.7360	16.49	11	.12
44	13	-"-	.7877	.2854	.7124	11.89	10	.29
45	14	-"-	2.0982	.8859	.6160	7.60	9	.57
46	15	-"-	2.6041	.0085	.4403	3.58	7	.83
47	18	-"	.8036	.5361	.7409	6.55	12	.89
48	19	-"-	1.1679	.8983	.7437	11.31	13	.58
49	20	-"-	.9490	.3809	.6927	13.73	12	.32
50	21	-"-	.8844	.3780	.7153	3.89	9	.92
51	22	-"-	4.5624	1.8594	.4959	11.64	9	.23
52	23	-"-	1.0097	.7256	.6868	12.31	9	.20
53	24	-"-	1.2076	.7012	.6637	10.37	10	.41
54	25	-"-	1.2542	.6579	.6791	10.68	11	.47
55	27	-"-	1.2918	.2896	.6138	13.38	10	.20
56	29	-"-	.9512	.4498	.6912	12.48	11	.33
57	30	-"-	1.3989	.4453	.6717	10.43	10	.40
58	32	-"-	1.1668	.4812	.6366	10.69	8	.22
59	34	-"-	1.1890	.1640	.5713	9.03	9	.43
60	35	-"-	.2198	.0737	.7275	9.91	9	.36
61	37	-"-	2.7954	1.3595	.5524	13.26	8	.10
62	38	-"-	1.0759	.3646	.6180	3.43	7	.84
63	39	-"-	2.0538	.8266	.5517	5.98	8	.65
64	40	-"-	1.8835	.5135	.6816	9.29	13	.75
65	41	-"-	1.7508	.5391	.5948	10.91	8	.21
66	43	-"-	.5920	.3937	.7820	11.12	13	.60
67	44	-"-	1.8878	.0706	.4861	5.69	4	.22
68	45	-"-	2.5419	.8964	.5603	8.39	6	.21
69	47	-"-	.3598	.0463	.7425	21.15	13	.07
70	48	-"-	1.2899	.2980	.6377	8.98	10	.53
71	49	-"-	1.0665	.3009	.7355	7.86	9	.55
72	50	-"-	.2479	.0366	.7829	12.49	13	.49
73	55	-"-	.9012	.1257	.6104	9.28	8	.32
74	56	-"-	.0765	.0164	.7786	16.95	12	.15
75	57	-"-	2.3737	.8680	.5607	3.43	8	.90
76	58	-"-	.3978	.2509	.8030	16.82	13	.21
77	59	-"-	.5792	.1395	.7795	14.03	17	.66
78	60	-"-	1.3912	.0167	.5864	8.14	6	.23
79	61	-"-	1.3050	.2240	.6566	12.79	9	.17
80	Theokrit: Idyllen	Clayman 1981	.5901	.1857	.4694	13.95	7	.05
81	Kallimachos: Hymnen	-"-	.1627	.0461	.5523	5.37	7	.61
82	Hesiod: Werke und Tage	-"-	.7146	.1302	.4439	3.22	5	.67
83	Schild des Herakles	-"-	.3455	.0688	.5255	5.85	6	.44
84	Theogonie	-"-	.6415	.0966	.5374	12.53	5	.03*

Tabelle 2
Fortsetzung

	Text	Quelle	K	M	q	X^2	FG	P
85	Homer: Odyssee	-"-	.7645	.0958	.4247	14.30	9	.11
86	Ilias	-"-	.5717	.0665	.4635	25.20	9	.003*
87	Hermeshymnus	-"-	3.6062	.2457	.2503	4.72	4	.32
88	Demeterhymnus	-"-	.6536	.1703	.5562	6.61	6	.36
89	Apollohymnus	-"-	.8761	.0981	.4852	4.90	6	.56
90	Aphroditehymnus	-"-	1.5376	.1794	.4306	2.73	5	.74
91	Arat: Phainomena 1	-"-	1.1227	.1099	.3316	.62	3	.89
92	2	-"-	.1900	.0229	.5031	5.99	4	.20
93	3	-"-	1.7780	.1289	.4105	2.36	5	.80
94	Gesamt	-"-	2.2505	.1824	.3180	7.48	5	.19
95	Nikander: Theriaka	-"-	1.4148	.1733	.4459	7.22	6	.30
96	Alexipharmaka	-"-	1.3425	.0014	.4898	13.92	5	.02 *
97	Apollonios Rhodios: Argonautika 1	-"-	1.8527	.1655	.3883	13.04	7	.07
98	2	-"-	3.1029	.3359	.3139	2.67	6	.85
99	3	-"-	1.8568	.3117	.4019	10.28	6	.11
100	Gesamt	-"-	2.1667	.2878	.3814	9.61	5	.09
101	Isokrates: Rede 1	Morton/Levison 1966	2.7702	.0902	.2903	10.45	6	.11
102	2	-"-	.3480	.0863	.6194	8.66	7	.28
103	3	-"-	1.1338	.0021	.6100	9.39	8	.31
104	4	-"-	.5739	.0065	.7721	21.04	16	.18
105	5	-"-	1.0155	.2593	.6998	13.01	7	.07
106	6	-"-	1.6913	.0013	.5928	17.88	7	.01*
107	7	-"-	1.0307	.1384	.6704	17.38	11	.10
108	8	-"-	2.3096	.5129	.6253	10.18	13	.68
109	9	-"-	1.5206	.2584	.6435	23.46	14	.05
110	10	-"-	1.5025	.2118	.6568	13.16	14	.51
111	11	-"-	2.3241	.0038	.5544	14.41	11	.21
112	12	-"-	.4266	.1528	.8185	22.39	19	.27
113	13	-"-	1.9805	1.0153	.7355	7.10	7	.42
114	14	-"-	.8671	.1134	.7044	9.84	11	.54
115	15	-"-	1.7402	.2584	.6341	15.13	13	.30
116	16	-"-	.5385	.0309	.7794	11.32	9	.25
117	17	-"-	2.3028	.7097	.5924	9.83	9	.36
118	18	-"-	1.2235	.0918	.5857	4.92	8	.77
119	19	-"-	1.1100	.1612	.6275	10.96	10	.36
120	20	-"-	3.6709	.0055	.4393	2.99	6	.81
121	21	-"-	1.5596	.0438	.4539	2.72	2	.26
122	Xenophon: Hieron	Wake 1957	.9179	.1783	.6104	9.49	7	.22
123	Agesilaos	-"-	1.4450	.0644	.5084	15.11	9	.09
124	Staatseinkünfte	-"-	1.3992	.1378	.5110	6.64	8	.58
125	Hipparchikos	-"-	1.4530	.0073	.4915	9.60	7	.21
126	Staat der Spartaner	-"-	1.6237	.0925	.4458	12.99	7	.07
127	Pferdezucht	-"-	1.8374	.0875	.4448	7.55	8	.48

Tabelle 2
Fortsetzung

	Text	Quelle	K	M	q	X^2	FG	P
128	Xenophon: Kynegetikos I	-"-	1.0272	.1174	.4714	11.77	4	.02*
129	II	-"-	1.5162	.3275	.4603	5.85	6	.44
	Aristoteles:							
130	De Generatione	-"-	.7348	.0982	.5924	7.39	9	.60
131	De Partibus I	-"-	.8163	.1473	.6168	4.61	8	.80
132	II	-"-	1.1093	.1294	.5484	18.18	11	.08
133	III	-"-	.7575	.0716	.5583	8.46	8	.39
134	IV	-"-	2.4630	.1059	.3725	6.02	6	.42
135	De Motu Animalium	-"-	1.0170	.1756	.5580	5.81	8	.67
136	De Incessu Animalium	-"-	.9297	.0930	.5638	18.38	11	.07
137	Historia Animalium I	-"-	.1272	.0146	.6067	8.20	6	.22
138	II	-"-	1.1370	.0881	.4448	2.96	5	.71
139	III	-"-	1.4598	.0188	.4761	8.50	5	.13
140	IV	-"-	1.8663	.1279	.3734	2.62	5	.76
141	V	-"-	21.4983	1.0672	.1128	2.76	4	.60
142	VI	-"-	1.0844	.0576	.4218	15.01	4	.005*
143	VII	-"-	1.3508	.1396	.5079	2.29	5	.81
144	VIII	-"-	.3802	.1030	.5491	8.08	5	.15
145	IX	-"-	.0613	.0100	.5718	6.20	5	.29
146	X	-"-	1.1150	.0664	.5174	2.55	6	.86
147	Ethik A	-"-	.4465	.0935	.6264	5.18	7	.64
148	B	-"-	.3493	.0293	.6146	9.73	6	.14
149	C	-"-	1.2079	.2077	.5229	2.54	6	.86
150	Nikomachische Ethik I	-"-	.5993	.1503	.5950	6.11	7	.53
151	II	-"-	1.3678	.0714	.4834	5.53	6	.50
152	III	-"-	1.3759	.1308	.4044	0.92	4	.92
153	IV	-"-	.0113	.0028	.5887	6.24	3	.10
154	VIII	-"-	.1442	.0184	.4747	2.84	3	.42
155	IX	-"-	1.4794	.2633	.4560	1.83	5	.87
156	X	-"-	4.2720	.374	.2713	11.82	4	.02*
157	Eudemische Ethik I	-"-	1.3523	.1860	.5663	10.31	6	.11
158	II	-"-	.0882	.0264	.6371	2.62	6	.85
159	III	-"-	.2514	.0819	.6269	6.31	6	.39
160	VII	-"-	.1972	.0384	.5655	3.86	5	.57
161	VIII	-"-	1.3549	.3992	.5028	1.30	5	.93
162	Plato: Gorgias	Wake 1957	.2871	.1399	.7358	12.92	11	.30
163	Phaidon	-"-	1.5190	.4341	.6714	15.95	13	.29
164	Phaidros	-"-	1.6619	.0319	.5685	11.81	10	.30
165	Timaios A	-"-	.7178	.0347	.7088	4.86	15	.99
166	B	-"-	1.2107	.1968	.6290	17.50	11	.09
167	C	-"-	1.0471	.1755	.7113	12.44	14	.57
168	A+B+C	-"-	.7906	.0981	.7031	15.81	19	.67
169	Gesetze V	-"-	1.7767	.2863	.5963	20.12	12	.06
170	VI	-"-	1.1889	.0036	.6384	8.91	11	.63
171	VII	-"-	.5847	.1596	.7706	21.36	16	.17

Tabelle 2
Fortsetzung

	Text		Quelle	K	M	q	X^2	FG	P
172	Plato	VIII	-"-	.3507	.0528	.7951	8.71	12	.73
173		IX	-"-	.8730	.0389	.7187	4.34	11	.96
174		XI	-"-	1.2102	.4047	.7165	5.39	13	.97
175		XII	-"-	.9569	.1328	.7271	12.81	15	.62
176		Brief VII	-"-	.5021	.1555	.7894	20.07	16	.22
177	Matuska: Od vcerajska k dnesku (Slovakisch)		Mistrik 1967	.8836	.3287	.7542	16.00	14	.31
	Brown-Corpus:		Markworth/Bell 1967						
178	A (Press, Reportage)		-"-	5.7678	.5804	.3455	14.45	10	.15
179	B (Press, Editorial)		-"-	4.7133	.7149	.3924	17.65	11	.09
180	C (Press, Reviews)		-"-	5.5187	.7188	.3825	17.70	10	.06
181	D (Religion)		-"-	3.1921	1.0210	.5645	15.20	13	.30
182	E (Skills and Hobbies)		-"-	3.7146	.3691	.3908	26.12	9	.002*
183	F (Popular, Love)		-"-	4.4060	.6027	.4091	11.14	5	.05
184	G (Belles Letters)		-"-	1.9534	.1971	.5424	15.05	12	.24
185	H (Miscellaneons)		-"-	2.3547	.3856	.5694	81.71	14	10^{-14}*
186	I (Learned and Scientific Writing)		-"-	3.4490	.2678	.4543	53.45	14	$2x10^{-4}$*
187	K (Fiction and General)		-"-	.8468	.3900	.5720	13.05	11	.29
188	L (Fiction, Mystery and Detective)		-"-	1.1697	.4369	.4845	10.83	9	.29
189	M (Fiction, Science)		-"-	1.4604	.6328	.5071	6.13	8	.63
190	N (Adventure and Western)		-"-	1.5402	.5202	.4607	9.36	9	.40
191	P (Fiction, Romance, Love-Story)		-"-	.3938	.1809	.5912	8.50	12	.74
192	R (Humor)		-"-	2.0233	.6942	.5545	24.46	9	.004*
193	Chesterton, G.K., A Short story of England		Williams 1940/1970	23.8505	.4467	.1618	8.20	9	.51
194	Wells, H.G., The Work, Wealth and Happines of Mankind		-"-	2.1630	.1793	.5342	12.37	12	.42
195	Shaw, Intelligent Women's Guide to Socialism		-"-	.8241	.1862	.7603	30.61	20	.06
	Dunsany, Chronicles of Rodriguez (1922)								
196		Descriptive	Williams 1970	.8884	.2057	.7421	9.23	16	.90
197		Dialogue	-"-	.1331	.0266	.5406	7.02	5	.22
	Bacon, Essays								
198	First half to end of XXVI		Yule 1939	1.3808	.0908	.7542	13.34	19	.82
199	Second half to end of LI		-"-	1.1817	.0254	.7445	14.54	16	.56
200	Zusammen			.9912	.0001	.7921	49.83	23	.001*
	Coleridge, Biographia Literaria. London, 1817								
201	A: Vol. I to p.134		-"-	1.4317	.0150	.7386	14.52	19	.75

Tabelle 2
Fortsetzung

	Text	Quelle	K	M	q	X^2	FG	P
202	Coleridge							
	B: Vol. II to pp. 1-66 and 104- to end	-"-	1.2952	.0872	.7422	17.38	18	.50
203	Zusammen	-"-	1.6303	.2178	.7358	21.49	22	.49
	Lamb,C.,Lamb,M., The Works of Charles and Mary Lamb, ed. E.V. Lucas. London 1905							
204	A: Elia: from the beginning to middle of Mrs. Battle's opinions on whist	Yule 1939	.0153	.0030	.7814	30.14	14	.007*
205	B: Last Essays: Detached Thoughts on books to Barbara S inclusive	-"-	.4978	.1271	.7415	31.95	19	.03*
206	Total	-"-	.0696	.0148	.7936	26.65	16	.04*
	Macaulay, Critical and historical essays. London 1888							
207	A: From first portion of essays on Lord Bacon (1837)	-"-	.4648	.0739	.6946	19.01	13	.12
208	B: From first portion of essay on The Earl of Chathan	-"-	1.6079	.1677	.5425	10.21	9	.33
209	Total	-"-	.7308	.0959	.6535	25.37	11	.01*
	De Imitatione Christi							
210	A: from Lib. I,II and IV	-"-	2.8830	.0517	.3146	30.82	6	.00003*
211	B: From Lib. III	-"-	2.0770	.1397	.3619	26.34	5	.00008*
212	Total	-"-	2.7943	.1336	.3335	77.25	6	10^{-14}*
	Petty, W. [Hull, C.H., The economic writing of Sir Williiam Petty together with the observations upon the bills of mortality more probably by Captain John Grant. Cambridge, UP 1899]							
213	A: Political Arithmetic, 300 Sentences, with 34 added from the Treatise of Taxes	-"-	1.3869	.0003	.8384	39.83	27	.05
214	B: Treatise of Taxes	-"-	2.2821	.4040	.7923	26.30	23	.29
215	C: Random passages	-"-	.7401	.2037	.8643	27.29	24	.29
216	Plato: Apologie	Brandwood 1969	.2931	.1262	.6926	8.38	12	.75*
217	Paulus Briefe:							
	An die Römer	Morton/McLeman 1966	.0100	.0032	.5760	11.89	8	.16
218	An die Korinther I	-"-	.2296	.0501	.4461	10.63	5	.06
219	An die Korinther II	-"-	.5233	.1534	.5365	8.76	6	.19
220	An die Galater	-"-	.4239	.0854	.5001	7.02	4	.13
221	An die Epheser	-"-	1.2171	.4735	.6844	12.91	10	.23
222	An die Philiper	-"-	.3770	.0694	.6043	8.50	6	.20
223	An die Kolosser	-"-	1.3301	.2817	.5622	2.33	9	.99

Tabelle 2
Fortsetzung

	Text	Quelle	K	M	q	X^2	FG	P
	Paulus Briefe:							
224	An die Thessaloniker I	-*-	.1676	.0206	.6805	7.50	7	.38
225	An die Thessaloniker II	-*-	.4529	.1251	.7244	4.88	6	.56
226	An Timotheus I	-*-	1.9180	.4500	.4464	6.36	5	.27
227	An Timotheus II	-*-	-1.8623	.7945	.5222	12.37	5	.03
228	An Titus	-*-	.9323	.5665	.6825	5.15	5	.40
229	An Philemon	-*-	1.3727	.0981	.3286	3.07	1	.07
230	An die Hebräer	-*-	.7050	.0888	.5234	10.57	7	.16
	Diodorus Siculus:							
231	Buch 1, Stichprobe 1	-*-	2.8110	.0003	.4982	22.77	7	.002*
232	Buch 1, Stichprobe 2	-*-	2.3303	.3487	.5248	5.79	8	.67
233	Buch 2	-*-	1.8818	.1584	.4381	5.25	6	.51
234	Buch 21	-*-	-2.0389	.0002	.4788	18.94	7	.01
235	Buch 22	-*-	2.4680	.0006	.3991	8.64	5	.12
236	Buch 32	-*-	2.7037	.0854	.4289	8.40	8	.40
237	Klemens: Rich man's salvation	-*-	.9834	.7312	.7474	15.28	12	.23
238	Exhortation to the Greek	-*-	1.2517	.4112	.6099	4.88	9	.84
239	Brief 1, Stichprobe 1	-*-	.7373	.3162	.5641	8.27	6	.22
240	Stichprobe 2	-*-	.9434	.2528	.4772	8.40	4	.08
241	Stichprobe 3	-*-	2.0107	.4966	.3643	4.19	4	.38
242	Stichprobe 4	-*-	.5524	.1851	.5467	2.99	6	.81
243	Brief 2	-*-	1.8383	.3938	.4264	3.65	4	.46
244	Johannes Brief	-*-	1.5139	.3196	.3439	3.29	3	.35

Tabelle 3
Anpassung der negativen Binomialverteilung
an Satzlängen

	Texte	k	p	χ^2	FG	P
1.	Heisenberg, W., Der Teil und das Ganze. München 1969.	3.7409	.6937	3.74	5	0.59
2.	Němcová, B., Babička.Praha 1961	2.6630	.5504	6.13	7	0.52
3.	Móricz, Zs., Légy jó mindhalálig. Budapest s.d.	1.5961	.4215	8.71	8	0.37
4.	Widjaja, P., Bila Malam Bertambah Malam. Jakarta 1971.	36.9454	.9704	1.05	3	0.79
5.	Peyrefitte, A., Le Mal Français. Plon 1976	20.7028	.9610	2.64	2	0.27
6.	Hronský, J.C., Na Bukovom Dvore. Bratislava 1975.	3.4411	.5466	6.83	9	0.65
7.	West, M.L., The Devil's Advocate. New York 1959.	8.2152	.8489	1.59	4	0.81
8.	Bunge, M., Scientific Research I. New York 1967.	10.5835	.8608	5.73	5	0.33
9.	Twitchell, P., The Tiger's Fang. Menlo Park 1967.	9.3138	.8696	1.82	4	0.77
10.	Fromkin, V., Rodman,R., An Introduction To Language. New York 1978.	6.5237	.7905	7.24	3	0.65

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 171–200

Arborescences Linguistiques

Henri Guiter
Montpellier

Le modèle des arbres généalogiques familiaux a depuis longtemps inspiré les représentations synoptiques d'éléments susceptibles de dériver d'une souche commune, tels que les races humaines ou les langages.

La seule anomalie de ce symbolisme est que, contrairement à l'ordonnance naturelle, ces arbres se présentent le plus souvent avec leurs ramures dirigées vers le bas et leurs souches dirigées vers le haut: cette fantaisie ne les empêche pas de rendre les services que l'on attend d'eux.

De telles constructions abondent dans le volume 30 de la collection "Quantitative Linguistics", intitulé "Statistics in Historical Linguistics" et dû à Sheila M. Embleton (1986).

Notons d'emblée que toutes les données des recherches de S. Embleton sont traitées par ordinateur, et ceci nous suggère déjà quelques réflexions.

Il nous revient d'abord en mémoire certain passage d'un article du regretté Jean Séguy (1973:12), après le long exposé d'une méthode dialectométrique plutôt laborieuse: "Dès lors, beaucoup de nos lecteurs doivent se dire: "Mais qu'attend ce pauvre homme pour appeler l'informatique à son secours?" On y a bien pensé: un seul exemple montrera que ce propos ne pouvait être que chimérique. Je prends dans la carte 1181 *rillons*, au point 679, le mot *héritus*, et j'en fais l'analyse – analyse dans ce cas relativement facile – destinée au programmateur. Je ne m'occupe pas de la transcription phonétique: transcodage que la perforeuse effectuera elle-même en consultant une table, table d'ailleurs richement garnie. Mais, en plus des numéros de la carte et de la localité, en plus de l'étymologie, je dois prévoir les informations suivantes: FR initial > *her*; phonème /h/ réalisé pleinement (en vue d'une carte phonologique où se distingueront les localités à réalisation positive de celles qui présentent une réalisation zéro; le moment venu,

l'ordinateur devra me renseigner sur ce chef pour tous les mots de l'atlas commençant par /h/); e̲ prétonique fermé (si je ne dégage pas cette information, l'ordinateur ne saura pas la découvrir en lisant l'étymon); de même pour u̲ tonique final et de syllabe fermée; (etc...). C'est là le minimum: je néglige certaines informations que je pense n'avoir jamais à utiliser, telle la règle phonétique ō) u̲, commune à tout l'occitan, ou la position des phonèmes les uns par rapport aux autres (distribution). Et ainsi de suite pour les centaines de milliers de vocables contenus dans l'ALG (quant à l'analyse des formes du verbe, nous ne nous sentons ni la force ni le courage d'imaginer un seul exemple). En admettant que je vienne à bout de l'analyse, combien de temps faudra-t-il pour réaliser la programmation monstrueuse? Et combien d'argent?"

J. Séguy ajoute dans une note infrapaginale: "Quelques dialectologues utilisent l'informatique. Mais ils ne mettent en mémoire qu'une quantité dérisoire d'information eu égard aux besoins de la dialectométrie. Contrairement au lexique, les autres composantes du langage, spécialement la phonétique diachronique, ne peuvent se mesurer sur un échantillon au hasard. Et on ne perdra pas de vue que le corpus d'un atlas linguistique n'est lui-même qu'un échantillon assez mince de chaque parler".

Un autre écueil dans l'emploi de l'informatique en linguistique tient au fait que les linguistes ont parfois une culture physico-mathématique assez réduite. Ils n'ont pas le sens des ordres de grandeur, et comprennent mal que la précision d'une mesure ne puisse excéder la définition de l'objet à mesurer. En voici un exemple récent.

Nous avons publié un article intitulé "Datation de divergences romanes" (3), en y appliquant la méthode glottochronologique que nous avions élaborée (Henri Guiter 1984a). Nous avertissions: "Il est commode de construire la courbe représentative de T en fonction de K, et cette abaque nous donnera rapidement, par simple lecture, des résultats qui ne sont pas tenus d'être extrêmement précis: un millimètre y représente vingt-cinq ans". Dans le courant de notre texte, nous répétions qu'il ne s'agissait là que de "dates-repères" ou de "jalons". Peu de temps après, nous recevions d'un collègue qui avait lu notre article, une lettre inquiète: il avait soumis nos données de base à l'ordinateur de son Université, et les résultats fournis par celui-ci différaient des nôtres (de moins de vingt-cinq ans, bien entendu). L'ordinateur donnait les reculs temporels des divergences avec quatre décimales après le nombre d'années, en gros à l'heure près, puisqu'il y a 8760 heures dans une année! Dans notre réponse nous demandions à cet estimable collègue s'il pourrait nous donner, au millimètre près, la distance entre sa ville universitaire et la capitale...

Nous rencontrons dans l'ouvrage de S. Embleton des constructions d'arbres qui appellent des remarques du même ordre.

A la page 36 l'arbre 1-10 présente des "distances négatives", la plus grande en valeur absolue étant −0.15. L'auteur décide de considérer ces distances négatives comme nulles, et, dans un souci d'équité, il annule aussi des distances positives du même ordre, soit 0.10 et 0.16. Or, nous voyons sur le même arbre des distances de 0.39, 0.41, 0.48, etc..., dont la valeur 0.16 représente respectivement 41%, 39% 33%, etc... Ces simplifications ne sont-elles donc pas abusives et peu respectueuses des ordres de grandeur? Quelle valeur peut-on accorder à l'arbre schématique 1-11, qui résulte de leur application?

A la page 38, l'arbre 1-13 voit éliminer la "distance négative" -0.24 et deux distances positives 0.19, alors que 0.24 représente respectivement 44%, 37%, etc... d'autres distances comme 0.54, 0.65, etc... qui entrent aussi en jeu. D'ailleurs, l'obtention de "distances négatives" montre que l'emploi de l'informatique n'est pas particulièrement adéquat.

Cependant, le but principal de l'auteur est d'appliquer la construction d'arbres à des données fournies par la méthode glottochronologique de Swadesh.

Sa bibliographie mentionne deux de nos publications (Henri Guiter 1983 et 1985); mais la seule allusion qui y est faite (p. 41) est celle-ci: "In this connection, the work of Henri Guiter correlating blood-types with glottochronology (e.g. Guiter 1983, 1985) is important". C'est aimable dans sa brièveté, mais c'est inexact. La première de ces deux publications propose essentiellement une nouvelle méthode glottochronologique, et la mise en relation avec la distribution des groupes sanguins n'y est qu'un à-côté. Quant à la seconde, il n'est absolument pas question de groupes sanguins, mais bien de l'établissement d'une relation mathématique entre les résultats de notre propre méthode et ceux d'une autre méthode proposée par A. Ross.

* * *

Avant d'aller plus avant, nous voudrions attirer l'attention sur les fluctuations que peut introduire, pour les distances lexicales entre les langues, le choix des signifiés servant à établir les listes de vocables.

Dans une recherche que nous voulions consacrer essentiellement à l'Europe, nous avons éliminé des termes non différenciés par les langues romanes (lat. *homo* et *uir*) ou représentés par des mots savants (*animal*); à leur place nous avons introduit des noms d'animaux fortement associés aux populations anciennes (Henri Guiter 1977a, 1977b, 1979, 1983) (*vache, cheval, porc, brebis*). Si, au lieu d'adopter la *brebis*, nous avions eu la fantaisie de mettre la *chèvre*, la Romania n'aurait plus présenté que les héritiers d'un étymon unique *capra*, au lieu de la variété résultant des étymons *ouicula*, *ueruex*, *foeta*, *pecora*; le groupe roman aurait été moins fragmenté.

Les conséquences du choix des signifiés retenus peuvent donc agir capricieusement sur la valeur des distances lexicales entre les langues des couples comparés.

Nous devrions avoir toujours les mêmes accords, ou les mêmes divergences, entre une langue romane et une langue germanique; or, si l'on compare neuf langues romanes et huit langues germaniques, on constate qu'il n'en est pas rigoureusement ainsi. Dans nos recherches précédentes (Henri Guiter 1977a, 1977b, 1979, 1983), nous avions considéré les pourcentages d'accords; nous allons maintenant considérer les pourcentages de divergences, c'est à dire le complément de ceux des accords à 100.

	Islandais	Norvégien	Danois	Suédois	Anglais	Frison	Néerlandais	Allemand
Portugais	63	65	65	64	64	67	65	64
Espagnol	64	66	66	66	65	69	67	66
Catalan	65	67	67	66	66	70	68	67
Gascon	61	63	63	62	62	65	63	62
Provençal	64	65	65	64	63	65	64	64
Français	64	65	65	64	63	65	63	64
Italien	63	64	64	63	62	65	63	62
Sarde	67	68	68	68	69	67	67	67
Roumain	64	66	66	65	62	67	65	64

Cette confrontation roman-germanique est celle qui nous offre le plus de couples différents pour deux familles de langues, c'est à dire 72. La valeur moyenne de ces 72 mesures est à peu près exactement 65%; la valeur la plus basse est 61%, la plus élevée 70%.

Portons sur un graphique, en abscisse les valeurs des écarts, de 0.61 à 0.70, et en ordonnée le nombre de fois que nous rencontrons chacune de ces valeurs: 1 pour 61, 6 pour 62, 10 pour 63, 15 pour 64, 15 pour 65, 8 pour 66, 10 pour 67, 4 pour 68, 2 pour 69 et 1 pour 70 (fig. 1). Les points obtenus ébauchent la forme de la traditionnelle courbe en cloche. Sur 72 mesures, 58, soit 80%, sont comprises entre 63 et 67 (et 68, soit 95%, entre 62 et 68). Nous pouvons donc espérer qu'une mesure isolée a une probabilité de 80% pour différer tout au plus de 2 unités de la valeur moyenne 65, considérée comme exacte. Quelle que soit la méthode par laquelle sera évaluée la distance lexicale de deux langues, il faudra toujours penser qu'elle est frappée d'une incertitude génératrice de fluctuations aléatoires.

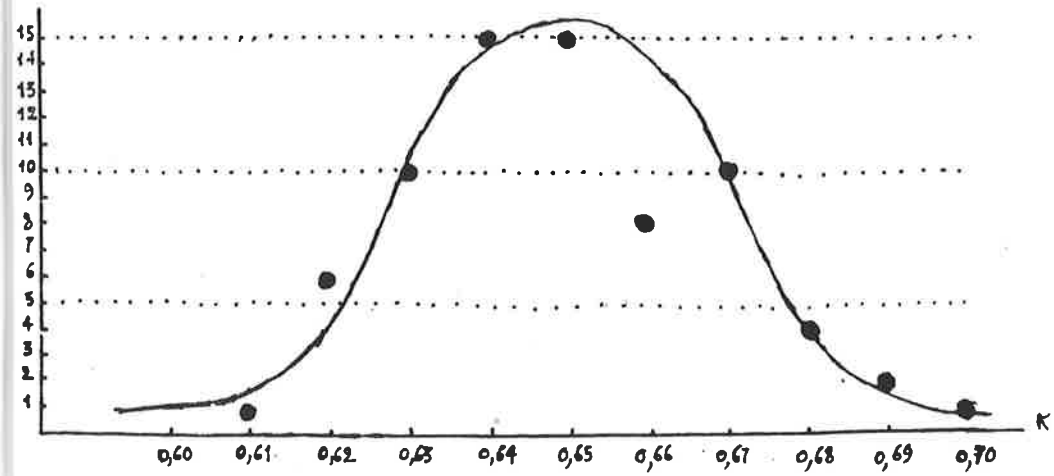


Fig. 1
Germanique — Italique

* * *

"A fructibus eorum cognoscetis eos. Numquid colligunt de spinis uuas, aut de tribulis ficus? Sic omnis arbor bona fructus bonos facit: mala autem arbor malos fructus facit"
(Matth. VII).

Cette citation évangélique semble particulièrement de circonstance lorsqu'il est question d'arbres, comme c'est le cas présentement.

Dans son ouvrage, S. Embleton donne ses sources, ses méthodes et leurs modalités d'emploi; il ne nous appartient pas de les exposer à nouveau ici. Nous voulons seulement examiner les résultats qui apparaissent sur les "arbres", les "fruits", si nous pouvons dire.

Plusieurs arbres, 1-3 (p. 31), 1-6 et 1-7 (p. 33), 1-8 (p. 34), répartissent de la même manière les six langues romanes qu'ils prennent en compte. L'espagnol et le portugais sont groupés d'un côté; le roumain est isolé de l'autre; au milieu, un même bouquet associe une singulière équipe constituée par le catalan, le français et l'italien.

A cette union contre nature nous opposerons des adversaires qui font autorité.

Amado Alonso écrit (1943): "El francés ha desarrollado en su historia tantos y tan extraños caracteres específicos que se opone por sí solo a las otras lenguas occidentales (y aun a todas las románicas)... El francés actual, con su carácter tan apartadizo de los otros idiomas románicos, es, pues, el resultado de una doble hibridación eficaz; la una, la acción del sustrato céltico, más triunfante que en ninguna otra región románica; la otra, la acción del superstrato franco, incomparablemente más persistente que la de ningún otro superstrato germánico. Ambas han convergido en apartar el francés del tipo latino, si lo comparamos con el italiano, el catalán, el castellano y el portugués".

Donnons maintenant la parole à Walther von Wartburg (1955): "Quand, un jour, la linguistique synchronique, cette science si jeune encore et si pleine de promesses, procèdera à une classification des langues romanes, elle mettra certainement à part le roumain et le français, et elle les opposera aux autres langues romanes ... Mais en principe je pense, comme je l'ai déjà dit, qu'au point de vue descriptif, les langues romanes se groupent en trois familles, dont deux ne se composent chacune que d'un membre, le roumain et le français, tandis que la troisième comprendra toutes les autres, le groupe méditerranéen ou méridional ..."

Les "fruits" de ces premiers arbres semblent donc bien peu appréciés par des experts dont il serait difficile de contester le jugement.

Les arbres 1-11 (p. 36) et 1-14 (p. 37) ajoutent le sarde aux six langues précédentes. Ils manifestent une unité portugais-espagnol-catalan-français, alors que le sarde, l'italien et le roumain sont tous isolés. Si l'italien s'est dégagé du malencontreux ensemble où il était inclus auparavant, le français est toujours associé aux langues ibériques, et les critiques précédentes restent valables.

Les études de S. Embleton qui suivent celles-ci, se réclament de données obtenues par la méthode de Swadesh. La première application est faite au germanique. Aux huit langues germaniques modernes que nous avons retenues dans nos propres travaux (islandais, norvégien, danois, suédois, anglais, frison, néerlandais, allemand), l'auteur ajoute huit idiomes supplémentaires: le bas-saxon de Hambourg, le saxon de Transylvanie, l'africaans, le flamand, le yiddish, l'allemand de Pensylvanie, le féroé et le tokpisin (pidgin à base d'anglais usité en Nouvelle Guinée).

On peut se demander si l'africaans et le flamand sont plus différenciés du néerlandais que les trois grandes variétés d'américain ne le sont de l'anglais, autrement dit, s'il valait la peine de les comptabiliser à part. Ou encore, si des parlers atomisés comme le saxon de Transylvanie peuvent être représentés sous une étiquette unique: est-ce l'idiome de Kronstadt ou celui de Hermannstadt? Des réponses à ces questions figurent peut-être dans la thèse doctorale non publiée de l'auteur; on ne peut cependant éviter de se les poser. Dans quelles limites est définie l'identité des langues ainsi invoquées?

Mais venons-en à ce qui nous intéresse particulièrement, c'est à dire aux "fruits" des arbres élaborés avec diverses modalités, 5-1 (p. 108), 5-2 (p. 109), 5-3 (p. 110), 5-4 (p. 112), 5-5 (p. 117). Tous ces arbres présentent des datations aux enfourchures de branches.

Notons d'abord deux points qui confirment nos propres résultats:

1) La date de séparation de l'anglais et du germanique continental est très proche de celle de la séparation du scandinave (si bien que deux arbres sur cinq indiquent l'anglais comme issu de la branche scandinave).
2) Le frison ne fait pas groupe avec l'anglais, mais bien avec le néerlandais et l'allemand.

Que penser des datations? Le premier fait qui nous frappe est que la fragmentation du germanique commun est située dans les premiers siècles de notre ère: 194, 122, 247, 295 ou 189. Or, telle n'est pas l'opinion d'Antoine Meillet (1949) lorsque'il parle de "cet idiome qu'on conviendra d'appeler 'germanique commun'... On s'abstiendra, dit-il, de rechercher ici où et quand s'est parlé l'idiome ainsi désigné; pour le lieu, il faut penser à la partie centrale de l'Europe, peut-être aux plaines du Nord de l'Allemagne; pour le temps, on ne se trompera guère en pensant aux deux ou trois siècles qui ont précédé le commencement de l'ère chrétienne". Les Goths émigrèrent de Scandinavie au milieu du II^{e} siècle, et leur nombre suppose que leur installation y était déjà assez ancienne.

Bien entendu, il est logique que les dates proposées pour l'isolement de l'anglais soient postérieures aux précédentes, respectivement 246, 170, 504, 440, 264, environ un demi-siècle dans trois cas sur cinq, et tombent donc sous le coup de notre précédente remarque.

La branche anglaise n'admet qu'une seule bifurcation, celle qui mène au tokpisin de Nouvelle-Guinée. Les Anglais ne se sont installés sur les côtes méridionales de cette île qu'en 1885. Or, les datations proposées par les cinq arbres sont 1842, 1041, 1853, 1153 et 1834. Trois de ces datations se situent au XIX^{e} siècle, bien avant 1885; elles ne sont cependant pas absurdes, et l'on peut attribuer les écarts aux fluctuations aléatoires. Mais que dire de la datation 1153, ou mieux encore 1041, celle-ci antérieure à la conquête de l'Angleterre par les Normands!

Passant à la branche scandinave, les dates de 893, 874 ou 873 s'accordent assez bien avec l'installation des colons normands en Islande; en revanche, celles de 1176 et surtout 1224 sont abusivement tardives.

La séparation du suédois et du dano-norvégien est indiquée par les dates de 1540, 1493, 1745, 1735 et 1531. En fait, elle est beaucoup plus ancienne; si les équipages des drakkars avaient mêlé jusqu'au IX^{e} siècle des Scandinaves de toutes origines, la formation des monarchies isola les deux groupes, les Suédois tournant vers l'est et le sud-est l'essentiel de leurs activités.

Du côté du germanique continental, trois arbres sur cinq nous donnent pour la séparation de l'allemand d'Allemagne et de l'allemand de Pensylvanie des dates de 1476, 1467 et 1459, antérieures, par conséquent,

à la découverte de l'Amérique! Les deux autres résultats, 1682 et 1754 sont évidemment plus vraisemblables.

Tous les arbres s'accordent pour offrir des couples de dates qui isolent le frison du néerlandais après que celui-ci se soit séparé de l'allemand: 1025 et 1239, 1022 et 1303, 1085 et 1253, 1124 et 1327, 1037 et 1236.

Or, nous lisons chez Antoine Meillet (1949): "Le groupe occidental est moins un que le groupe nordique. Il comprend le groupe anglais, dont les parlers frisons sont assez rapprochés, et le groupe allemand, où l'on distingue le haut allemand, le bas-allemand ou saxon, et le néerlandais".

Plus radical à ce sujet est Carlo Tagliavini (1961): "Da questa bipartizione nacquero i due grandi gruppi in cui si suddivide il germanico occidentale, e cioè lo *anglo-frisone* (che alcuni filologi chiamarono anche *ingwaonico* ted. ingwäonisch) deve aver avuto un certo periodo di unità, che però non possiamo determinare esatamente"

Même si l'on n'adopte pas entièrement ces points de vue, le fait qu'ils aient été émis implique que le frison diffère de l'allemand et du néerlandais plus que ces deux dernières langues ne diffèrent entre elles. Et donc, l'isolement du frison, devrait être antérieur à la séparation de l'allemand et du néerlandais.

Nous limiterons là nos observations sur les arbres du germanique, et nous passons au groupe roman. Les langues modernes prises en compte sont le français, l'espagnol, l'italien, le roumain, le catalan, le portugais, le papiamento de Curaçao, le frioulan, le rhéto-roman, le créole de Sainte-Lucie, le créole de Guadeloupe. Remarquons l'absence du sarde et de l'occitan, qui ne figurent par aucun de leurs dialectes. Les résultats sont exposés au moyen de quatre arbres (pp. 137, 142, 144 et 145).

Trois de ces arbres donnent la date 206 pour la séparation du roumain, date qui se situe en pleine période de colonisation de la Dacie; mais le quatrième arbre rabaisse la date à 544, ce qui n'a plus aucun rapport avec les faits historiques.

L'italien se serait séparé du français en 952, 898, 838 ou 1189. Or il semble bien improbable que, dans le nord de la Gaule, l'unité romaine ait pu survivre à l'invasion franque et à la formation de royaumes francs. Une preuve directe résulte du texte de la "Cantilène de Sainte-Eulalie",

écrite vers la fin du IX^{e} siècle; la langue s'écarte du latin et de l'italien: c'est déjà du français.

Pour des raisons du même ordre que ci-dessus les dates de séparation du français et de l'ibéro-roman 1066, 984, 909, ou 1192 sont, elles aussi beaucoup trop basses.

Les dates de séparation données pour le portugais et l'espagnol sont 1573, 1551, 1525 et 1664. Ici l'absurdité des résultats est vraiment patente. Au XVI^{e} siècle l'espagnol de Cervantes et le portugais de Camões en étaient tous deux à l'Age d'Or de leur littérature. Et les deux langues avaient déjà derrière elles plusieurs siècles de production écrite depuis le Poema del Cid (1140) ou les textes portugais du XII^{e} siècle (J. J. Nunes 1959). Nous connaissons d'ailleurs très bien, par l'Histoire, la genèse de l'entité castillane au X^{e} siècle.

Les résultats sont peut-être moins choquants pour la séparation du catalan et des autres idiomes péninsulaires: 1180, 1144, 1116, 1372; ils n'en sont pas moins très inexacts. Nous citions plus haut des textes espagnols et portugais du XII^{e} siècle; il existe des textes juridiques catalans de même époque, ainsi que les fameuses "Homilies d'Organyà" (P. Russel-Gebbett 1965). On peut penser que c'est la poussée musulmane et la Reconquête qui ont séparé l'est et l'ouest péninsulaires.

Le quatrième arbre mérite une mention particulière parce qu'il fait apparaître en 957 le papiamento antillais! Qu'en dirait Christophe Colomb?

E. Coseriu avait dressé, il y a un quart de siècle, un très sévère réquisitoire contre la méthode Swadesh, après avoir tenté de l'appliquer aux langues romanes. S. Embleton y fait une discrète allusion: "Coseriu (1965), on the base of Romance retention rates which vary between 91% and 78% per millenium, questions the universality of the retention rate and therefore the accuracy of glottochronology as a tool for precise dating" (p. 54).

De fait, les expressions de Coseriu sont beaucoup plus radicales (Coseriu 1965): "En conclusion, la glottochronologie nous paraît être une technique mal fondée, fausse et même théoriquement absurde. Et de par sa facilité même – et superficialité – elle nous paraît représenter un danger pour la linguistique contemporaine ... Par sa fausseté même, la glottochronologie peut servir au moins à nous démontrer exactement le contraire de ce qu'elle se propose: c'est à dire que l'histoire, il faut la

faire pour chaque cas individuellement, et qu'on ne peut pas la déduire mathématiquement d'une façon générale". Ceci semblerait une condamnation sans appel.

Une troisième étude est consacrée à la famille Wakash, langues amérindiennes parlées sur la côte pacifique de la Colombie Britannique et de l'Etat de Washington. Nous sommes fort loin d'être spécialiste en ce domaine; d'ailleurs, la mise en relation avec des événements historiques est totalement impossible. Toutefois, lorsqu'on voit le point de départ de l'arbre reculé jusqu'en 3.506 ou même 4.100 avant notre ère, on est tenté de s'exclamer, comme le juge des "Plaideurs" de Racine: "Avocat, passons au déluge !".

Il faut reconnaitre que notre promenade dans ce bosquet linguistique ne nous a pas montré beaucoup d'arbres dont les fruits ne soient au moins douteux, pour ne pas dire plus.

* * *

"La critique est aisée, et l'art est difficile".

Cet alexandrin, passé en proverbe, pourrait nous être dédié, si, après avoir critiqué le travail de S. Embleton, nous ne proposions pas une méthode donnant des résultats plus corrects. Nous essayons de le faire en prenant comme exemple les langues dont le passé historique est sans doute le mieux connu, les langues indo-européennes de la famille *kentum*: italique, germanique, celtique et grec.

Nous avons déjà publié dans nos travaux rappelés sous la référence (Guiter 1977a, 1977b, 1979, 1983) les pourcentages d'accords entre sept langues romanes (portugais, espagnol, catalan, provençal, français, italien, roumain), sept langues germaniques (norvégien, danois, suédois, anglais, frison, néerlandais, allemand), deux langues celtiques (gallois, irlandais) et le grec moderne. Par la suite, nous avons pris en compte le gascon et le sarde (Guiter 1984a), l'islandais, le breton et le cornique (Guiter 1984b), l'erse d'Ecosse (Guiter 1985).

S. Embleton a adopté, comme seuil des langues modernes à comparer, l'année 1980. C'est sans doute notre âge qui nous avait amené à choisir comme seuil l'année 1900. Ceci est pratiquement sans influence sur les résultats.

Nous avons adapté, pour construire l'arbre du roman, les différences que nous avions comptées entre les diverses langues romanes (fig. 2).

Comme sur une carte routière, nous avons indiqué la distance entre chaque couple d'enfourchures voisines, et aussi entre chaque extrémité de rameau et l'enfourchure précédente. Pour obtenir la distance donnée par l'arbre entre deux langues romanes, il suffit d'ajouter les distances successives rencontrées sur le trajet qui va de l'une à l'autre. Par exemple, entre le français et le portugais: $9+2+3+1+6+1+4=26$.

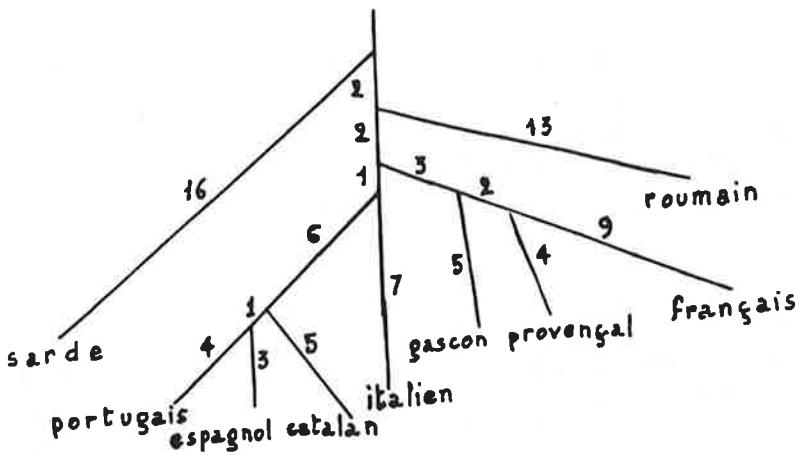


Fig. 2
Roman

Sur un tableau à double entrée nous allons reporter les pourcentages d'écarts entre deux langues, tels qu'ils découlent de nos travaux antérieurement publiés. Après chacun de ces nombres, nous indiquons entre parenthèses la distance comptée sur l'arbre de la figure 2. –

	Portugais	Espagnol	Catalan	Gascon	Provençal	Français	Italien	Sarde
Espagnol	7 (7)							
Catalan	11 (10)	9 (9)						
Gascon	18 (20)	19 (19)	17 (20)					
Provençal	22 (21)	23 (20)	19 (21)	10 (11)				
Français	26 (26)	28 (25)	27 (26)	17 (16)	13 (13)			
Italien	16 (18)	18 (17)	15 (18)	17 (16)	17 (17)	19 (22)		
Sarde	31 (32)	32 (31)	31 (32)	24 (28)	26 (29)	31 (34)	27 (28)	
Roumain	23 (27)	25 (26)	24 (27)	19 (23)	28 (24)	33 (29)	23 (23)	31 (31)

Aucun écart n'est supérieur à 4, alors que chacun des deux nombres comparés peut raisonnablement être approché à ± 2. Il y a 5 écarts égaux à ± 4, c'est à dire 13% des termes; en revanche 23 écarts, soit 63% sont compris entre $+2$ et -2.

Nous construisons de la même façon l'arbre relatif aux langues germaniques (fig. 3). Ici encore, nous mettons côte à côte, pour chaque couple de langues, le nombre résultant de nos mesures directes de pourcentages d'écarts, et celui mesuré le long du trajet qui, sur l'arbre, conduit d'une langue à l'autre.

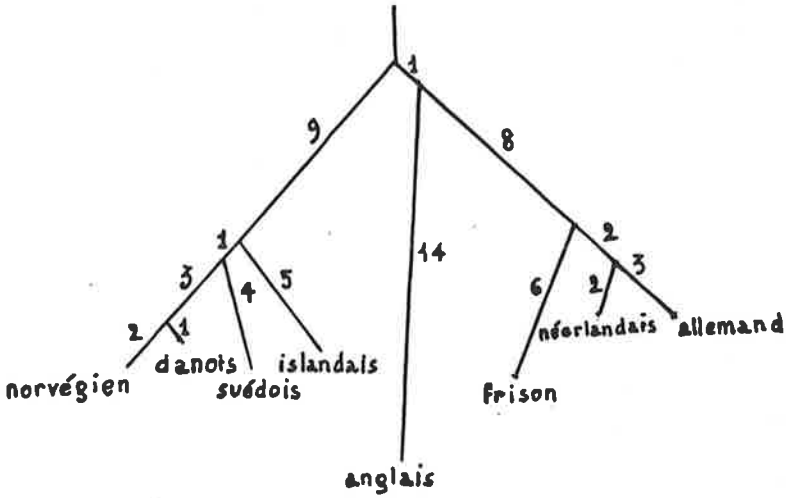


Fig. 3
Germanique

	Islandais	Norvégien	Danois	Suédois	Anglais	Frison	Néerlandais
Norvégien	11 (11)						
Danois	10 (10)	3 (3)					
Suédois	14 (10)	10 (9)	7 (8)				
Anglais	28 (29)	30 (30)	30 (29)	30 (29)			
Frison	31 (29)	30 (30)	31 (29)	29 (29)	27 (28)		
Néerlandais	27 (27)	30 (28)	29 (27)	27 (27)	27 (26)	10 (10)	
Allemand	26 (28)	28 (29)	27 (28)	24 (28)	26 (27)	12 (11)	5 (5)

Ne seraient-ce deux écarts exceptionnels de 4, aucune des autres différences n'excède 2. Sur les 28 couples comparés, 10 différences sont nulles (36%), et 21 (75%) sont comprises entre -1 et $+1$.

Opérons de même pour les langues celtiques (fig. 4).

	Gallois	Cornique	Breton	Irlandais
Cornique	14 (14)			
Breton	20 (19)	9 (9)		
Irlandais	50 (50)	47 (46)	48 (51)	
Erse	52 (51)	48 (47)	49 (52)	19 (19)

Avec les langues celtiques, 80% des différences ne dépassent pas 1.

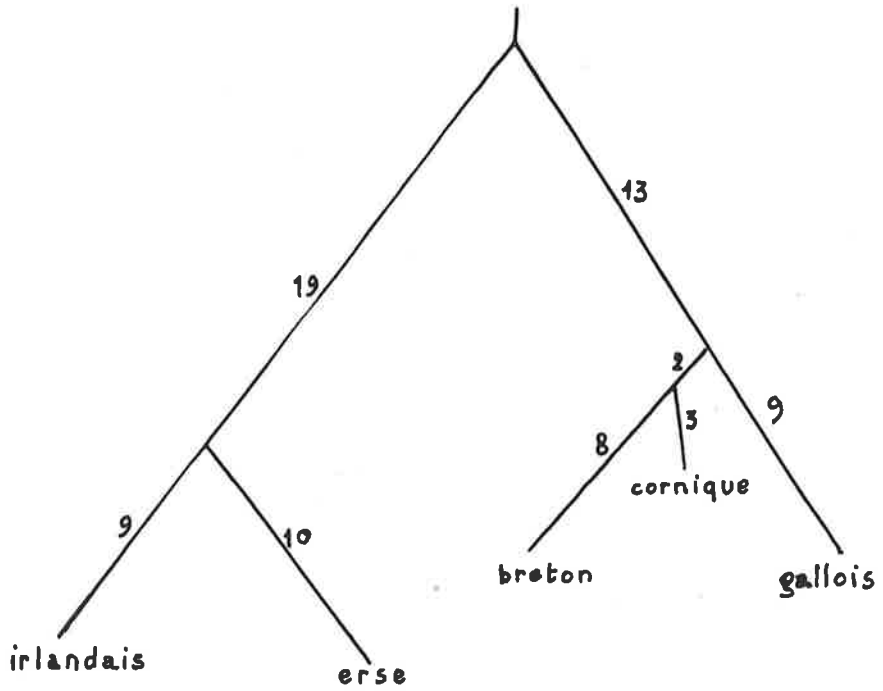


Fig. 4
Celtique

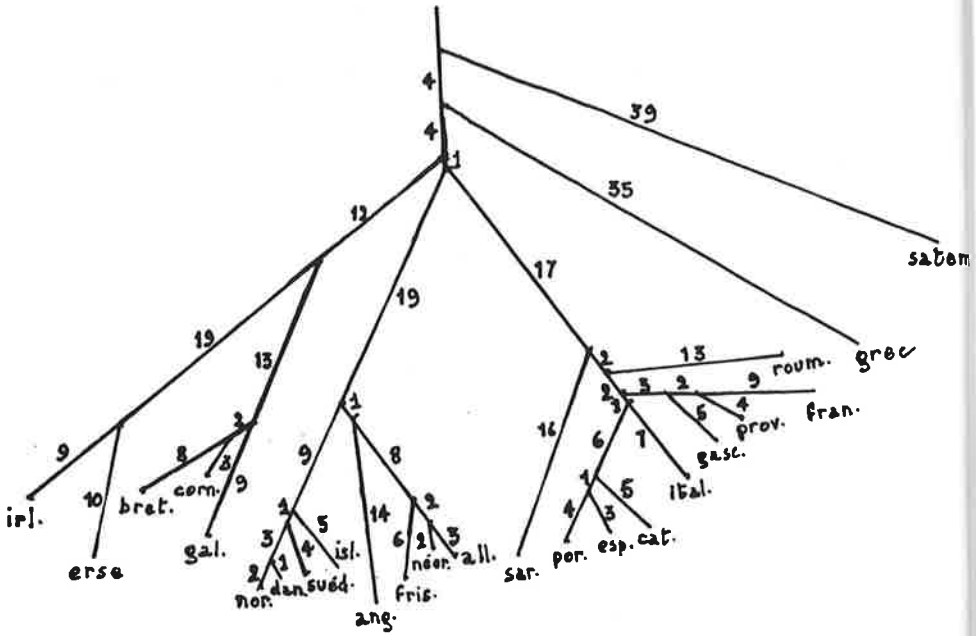


Fig. 5

Bien entendu, pour le grec il n'y a pas d'arbre à dresser, puisque nous ne disposons que d'une seule langue moderne.

Nous pouvons maintenant assembler les arbres que nous venons de dresser, avec le grec pour obtenir l'arbre correspondant à la famille *kentum* (Nous y adjoignons même un départ en direction de la famille *satem* basé sur la moyenne des distances entre toutes les langues *kentum* et les deux langues *satem* que nous avions étudiées) (fig. 5).

Le tableau permettant l'appréciation des écarts pour les 23 langues *kentum* est présenté de la même façon que ceux relatifs aux familles de langues; chaque pourcentage de différences résultant de la comparaison

directe des listes de signifiants pour chaque couple de langue, est suivi, entre parenthèses, de la mesure effectuée sur l'arbre de la figure 5.

Les écarts entre ces deux séries de valeurs, au nombre de 253, varient de -7 à 6. Les fréquences de chaque valeur d'écart sont les suivantes: 1 pour -7, 2 pour -6, 3 pour -5, 14 pour -4, 16 pour -3, 15 pour -2, 35 pour -1, 55 pour 0, 53 pour 1, 27 pour 2, 16 pour 3, 12 pour 4, 3 pour 5 et 1 pour 6.

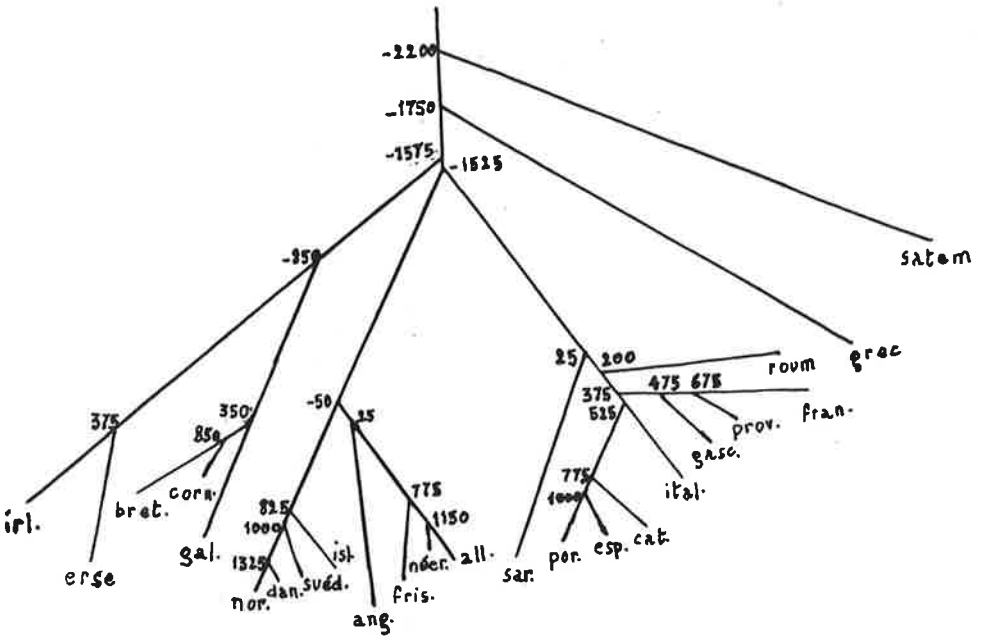


Fig. 6

Sur le graphique de la figure 7, nous portons en abscisse les valeurs des écarts, et en ordonnée leurs fréquences. Les points obtenus ébauchent l'allure d'une courbe en cloche. Dans 57% des cas, l'écart se situe de -1 à 1; dans 73% des cas, de -2 à 2; dans 96% des cas, de -4 à 4.

	Portug.	Esp.	Cat.	Gasc.	Prov.	Fran.	Ital.	Sard.	Roum.
Espagnol	7 (7)								
Catalan	11 (10)	9 (9)							
Gascon	18 (20)	19 (19)	17 (20)						
Provençal	22 (21)	23 (20)	19 (21)	10 (11)					
Français	26 (26)	28 (25)	27 (26)	17 (16)	13 (13)				
Italien	16 (18)	18 (17)	15 (18)	17 (16)	17 (17)	19 (22)			
Sarde	31 (32)	32 (31)	31 (32)	24 (28)	26 (29)	31 (34)	27 (28)		
Roumain	23 (27)	25 (26)	24 (27)	19 (23)	28 (24)	33 (29)	23 (23)	31 (31)	
Islandais	63 (66)	64 (65)	65 (66)	61 (62)	64 (63)	64 (68)	63 (62)	67 (66)	64 (65)
Norvégien	65 (67)	66 (66)	67 (67)	63 (63)	65 (64)	65 (69)	64 (63)	68 (67)	66 (66)
Danois	65 (66)	66 (65)	67 (66)	63 (62)	65 (63)	65 (68)	64 (62)	68 (66)	66 (65)
Suédois	64 (66)	66 (65)	66 (66)	62 (62)	66 (63)	64 (68)	63 (62)	68 (66)	65 (65)
Anglais	64 (67)	65 (66)	66 (67)	62 (63)	63 (64)	63 (69)	62 (63)	69 (67)	62 (66)
Frison	67 (67)	69 (66)	70 (67)	65 (63)	65 (64)	65 (69)	65 (63)	67 (67)	67 (66)
Néerlandais	65 (65)	67 (64)	68 (65)	63 (61)	64 (62)	63 (67)	63 (61)	67 (65)	65 (64)
Allemand	64 (66)	66 (65)	67 (66)	62 (62)	64 (63)	64 (68)	62 (62)	67 (66)	66 (65)
Gallois	69 (68)	70 (67)	71 (68)	66 (64)	69 (65)	67 (70)	66 (64)	68 (68)	65 (67)
Cornique	68 (64)	67 (63)	68 (64)	63 (60)	66 (61)	64 (66)	63 (60)	67 (64)	63 (63)
Breton	68 (69)	68 (68)	68 (69)	64 (65)	67 (66)	64 (71)	63 (65)	67 (69)	63 (68)
Irlandais	75 (74)	76 (73)	76 (74)	76 (70)	76 (71)	75 (76)	74 (70)	75 (74)	72 (73)
Erse	75 (75)	76 (74)	77 (75)	76 (71)	76 (72)	77 (77)	75 (71)	76 (75)	75 (74)
Grec	70 (73)	70 (72)	70 (73)	70 (69)	71 (70)	72 (75)	69 (69)	70 (73)	71 (72)

	Isl.	Norv.	Dan.	Suéd.	Angl.	Fris.	Néer.	All.	Gall.	Corn.	Bret.	Irl.	Erse
Norvégien	11 (11)												
Danois	10 (10)	3 (3)											
Suédois	14 (10)	10 (9)	7 (8)										
Anglais	28 (29)	30 (30)	30 (29)	30 (29)									
Frison	31 (29)	30 (30)	31 (29)	29 (29)	27 (28								
Néerlandais	27 (27)	30 (28)	29 (27)	27 (27)	27 (26)	10 (10)							
Allemand	26 (28)	28 (29)	27 (28)	24 (28)	26 (27)	12 (11)	5 (5)						
Gallois	67 (68)	69 (69)	69 (68)	68 (68)	68 (69)	67 (69)	69 (67)	68 (68)					
Cornique	65 (64)	67 (65)	67 (64)	66 (64)	66 (65)	65 (65)	67 (63)	66 (64)	14 (14)				
Breton	64 (69)	66 (70)	66 (69)	65 (69)	65 (70)	64 (70)	66 (68)	65 (69)	20 (19)	9 (9)			
Irlandais	73 (74)	74 (75)	74 (74)	73 (74)	76 (75)	75 (75)	76 (73)	75 (74)	50 (50)	47 (46)	48 (51)		
Erse	75 (75)	76 (76)	77 (75)	77 (75)	76 (76)	75 (76)	77 (74)	76 (75)	52 (51)	48 (47)	49 (52)	19 (19)	
Grec	72 (73)	73 (74)	73 (73)	73 (73)	74 (74)	75 (74)	74 (72)	73 (73)	73 (73)	73 (69)	72 (74)	79 (79)	81 (80)

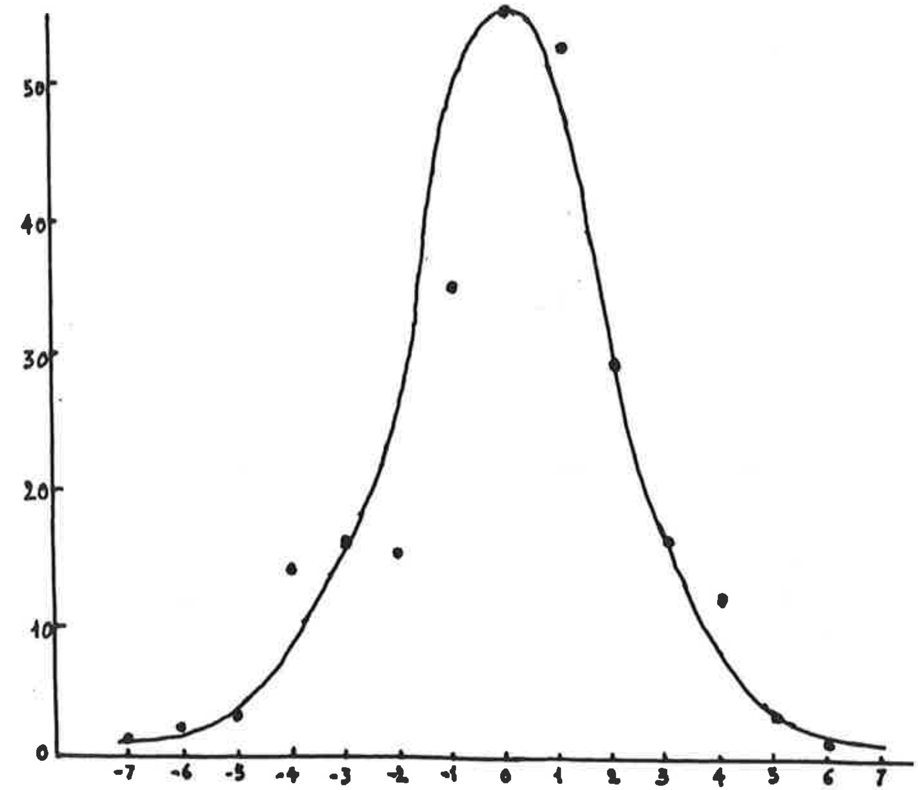


Fig. 7

Ce résultat correspond bien à ce que l'on peut espérer, étant donné les incertitudes frappant les mesures, telles que nous les avions appréciées par la confrontation des langues romanes et germaniques.

Pour l'établissement de ces systèmes, le cerveau humain a remplacé tout ordinateur. Ceci explique, sans doute, qu'il n'apparaisse aucune de ces "distances négatives", qui sont autant d'outrages au "bon sens", ce bon sens que Descartes estimait "la chose du monde la mieux partagée". Conscient de l'approximation des résultats, nous avons arrondi toutes

les distances à l'unité de pourcentage qui en était la plus voisine: aussi, les deux "moitiés" d'un nombre impair diffèrent-elles d'une unité.

Il nous appartient toutefois, pour ne pas être en reste avec S. Embleton, de mettre une date sur chaque enfourchure. Ces dates, nous les avons déjà publiées ça et là; nous les tirions de l'abaque qui les donne en fonction du pourcentage de concordances. Lorsqu'il s'agit de comparer des groupes de langues, nous prenons la moyenne des concordances entre les langues des couples qui les constituent. Par exemple, dans ce même travail, nous avons déjà calculé la moyenne des valeurs correspondant aux 72 couples italo-germaniques: 35% de concordances, et donc 65% de différences.

Pour ne pas surcharger l'arbre de la figure 5 avec deux séries de nombres qui pourraient prêter à confusion, et en rendre la lecture plus difficile, nous allons reproduire cet arbre sans l'indication des distances et y inscrire les dates (fig. 6). Ces dates sont tirées d'une abaque sur laquelle un millimètre représente 25 ans; pour cette raison, ces dates-repères sont arrondies à un multiple de 25. Pouvons-nous tenter de les justifier?

−2200. A. Meillet écrit (Antoine Meillet 1943): "Les populations qui parlaient l'indo-européen commun, occupaient un domaine situé dans une région plutôt septentrionale, et, tout au plus, centrale, soit en Europe, soit à la limite de l'Europe, et de l'Asie On entrevoit donc, à partir du second millénaire avant l'ère chrétienne, une marche conquérante des populations de langue indo-européenne provenant sans doute de la région des Carpathes et de la Thrace". Une première rupture de la nation indo-européenne dans les derniers siècles du troisième millénaire s'accorde assez avec cette dispersion à partir du second millénaire.

−1750. Dans son mouvement vers l'ouest, le groupe *kentum* approchait sans doute des Carpathes lorsque les Grecs s'en séparèrent pour marcher vers le sud. Il leur fallait, dès lors, le temps d'arriver jusqu'en Crète vers 1400, pour y manifester la présence de leur langue à travers les inscriptions du linéaire B. "La stratigraphie cnossienne, en effet, écrit Jacques Raison (Jacques Raison 1963), ainsi que l'a définie Evans, trouve confirmation sur les autres sites de Crète, qui nous offrent en 1400, avec un cataclysme identique, une même séparation des couches". Un délai de trois siècles et demi n'était pas inutile pour atteindre la grande île méditeranéenne à partir d'une région qui pouvait être l'actuelle Volynie.

−1575, −1525. Le XVI^{e} siècle voit se disloquer la communauté italo-germano-celtique. Le lien avec les Celtes semble avoir été rompu avant celui des deux autres ethnies. D'une région qui pourrait être l'actuelle Pologne, les Celtes s'avancent vers l'ouest, les Italiques partent vers le sud par le trajet assez facile de la Moravie, l'Autriche et les Alpes Juliennes. Il est admis que les Latins avaient gagné le Latium dès le XII^{e} siècle, et qu'ils y furent étroitement resserrés par l'invasion des Etrusques, puis par celle des Osco-ombriens, dont ils s'étaient séparés avant de franchir les Alpes. Comme dans le cas précédent, un délai de trois à quatre siècles semble raisonnable pour permettre le trajet jusqu'à l'Italie centrale et le développement des caractères propres de la langue.

−850. Les Gaéliques, qui ont toujours été en pointe à l'avant-garde occidentale des migrants, se séparent des Gallo-brittoniques et déferlent sur l'Europe occidentale. Les préhistoriens (Maurice Louis 1948) datent la première vague celtique, la vague gaélique, véhicule de la culture de Hallstatt, de 900 à 500; la seconde vague, gallo-brittonique, véhicule de la culture de la Tène, de 500 à 150. La séparation des deux groupes celtiques est donc bien intervenue au IX^{e} siècle, lors du départ des Gaéliques.

−50. Nous avons déjà rappelé les indications de Meillet situant aux veilles de notre ère l'existence du germanique commun, et donc la séparation du scandinave et du germanique occidental.

25. Isolement du sarde. Conquise par les Romains en −238, entre les deux premières guerres puniques, parce qu'elle constituait une plateforme avancée vers l'Espagne carthaginoise, la Sardaigne perdit toute importance militaire après la défaite d'Hannibal. Peu attrayante à plusieurs titres, elle fut de plus en plus coupée de Rome, et devint sous l'Empire un lieu de rélégation pour personnages en disgrâce.

Séparation de l'anglais et du germanique continental. C'est un des points sur lesquels nous nous accordons avec S. Embleton; elle doit donc vraiment correspondre à une réalité. On peut penser que, dès cette époque, les tribus qui gagneraient plus tard l'Angleterre, étaient rassemblées sur la rive droite de l'Elbe, à l'écart du reste de la Germanie.

200. Séparation du roumain. La colonisation de la Dacie, entreprise par Trajan dans les premières années du II^{e} siècle, devait prendre fin en 270 avec le repli des légions au sud du Danube. La date de 200 se situe au fort de cette colonisation.

350, 375. Vers cette époque deux événements liés entre eux, intéressent les deux branches des populations celtiques. En 381 les légions de Bretagne proclament Maxime empereur et franchissent la Manche avec lui. Maxime conféra des bénéfices militaires à un grand nombre de volontaires bretons, qui s'installèrent à l'ouest d'une Armorique déjà romanisée. Ils y rétablirent l'emploi d'une langue celtique, le breton, qui se sépara alors de ce qui sera le gallois.

Profitant du départ des légions, deux assaillants se ruent sur la Grande Bretagne, d'une part les Calédoniens descendus d'Ecosse, d'autre part la tribu irlandaise des Scots. Mais entre 396 et 406 Stilicon, envoyé par Honorius, refoule tous les envahisseurs vers le nord, et les Scots s'établissent dans la partie occidentale de l'Ecosse. L'aventure de Maxime aura ainsi provoqué leur séparation de l'irlandais.

A cette même époque un autre bouleversement se produit sur un autre théâtre. Dès la fin du III^{e} siècle les Francs s'étaient implantés au nord de la Gaule; en 355 ils font une poussée, et l'empereur Julien doit venir leur faire lever le siège d'Autun; cependant en 358 il leur laisse le territoire où ils se sont installés. En 406, les Francs progressent encore avec la grande invasion, et deviennent les maîtres de la Gaule septentrionale, dont les relations avec l'Italie sont définitivement compromises.

475. Un événement particulièrement important est la déposition,par le chef hérule Odoacre, du dernier empereur d'Occident, Romulus Augustule. Les Wisigoths dominent le sud-ouest de la Gaule et les relations avec la Gaule du nord, occupée par les Francs ne peuvent guère subsister.

525. Eloignement de l'Espagne et de l'Italie. Les Wisigoths, battus par les Francs en 507, se sont retirés vers l'Espagne et la Septimanie. En 536, les Francs attaquent les Ostrogoths de Provence, et interrompent le contact territorial à travers l'ancienne *Prouincia.* Et en 540, l'Empire d'Orient s'empare de l'Italie (que les Lombards envahiront en 568).

675. Les Francs installés en Provence depuis 536 faisaient peser une dure sujétion sur ce pays. Mais, après les règnes autoritaires de Clotaire II et Dagobert I (613–638), s'installe l'ère des "rois fainéants"

et les rivalités des maires du palais de Neustrie et d'Austrasie. Cette époque anarchique durera près d'un siècle, jusqu'aux succès de Charles Martel (737). Une telle période de détente permet à la Provence de se différencier, et le jalon de 675 en marque sensiblement le milieu.

750, 775. Il est aisé de voir quels sont les faits historiques qui expliquent une scission entre l'est et l'ouest péninsulaires au cours du VIII^{e} siècle. La domination wisigothique avait prolongé l'unité romaine en Ibérie. Mais l'invasion musulmane franchit en 711 le détroit de Gibraltar pour ne s'arrêter qu'en 732 à Poitiers. Cependant, des noyaux chrétiens résistent à l'abri des chaînons de la cordillère cantabro-pyrénéenne, à l'ouest le noyau asturo-galicien, à l'est le noyau catalan. Isolés l'un de l'autre, ils entreprennent une lente reconquête: le Duero est atteint en 791, Girona libérée en 785, Barcelona en 801. C'est donc l'invasion du VIII^{e} siècle qui est responsable de cette séparation entre l'est et l'ouest péninsulaires; avec la Reconquête, chaque langue étendra son domaine vers le sud.

Vers la fin du VIII^{e} siècle, le frison s'isole du germanique continental. Après les assauts que leur ont livrés Pépin d'Héristal (689), puis Charles Martel (716, 719,734), les missions de conversion au christianisme (690, 716,719,755), nous voyons les Frisons refuser à Charlemagne en 792 le contingent militaire qu'il leur demande. Trop occupés en Germanie, en Italie ou en Espagne, Pépin le Bref et Charlemagne avaient laissé s'isoler le petit peuple maritime. Ce n'étaient pas Louis le Pieux, ou ses fils, qui pouvaient redresser la situation.

825. L'Islande fut découverte et peuplée par les Norvégiens au cours de la deuxième moitié du IX^{e} siècle. Notre repère est donc un peu précoce, mais l'ordre de grandeur n'est pas bousculé.

850. "La grandeur de Wessex, ébauchée par Ini (688–725), commença réellement avec Ecbert ... Il envahit le pays gallois, prit Chester et pénétra jusqu'à l'île de Mona (Anglesey). Les Celtes de Cornouailles se soulevèrent; il les repoussa et les soumit (835)". Cette citation empruntée à l'"Histoire Générale" (E. Lavisse et A. Rambaud, t. I 1893) nous explique que les relations maritimes étroites entre Armorique et Cornouailles (cf. Tristan et Iseult!) se soient poursuivies aussi longtemps que ces dernières ont conservé leur indépendance; tout cesse en 835, lorsqu'elles sont occupées par les Saxons.

1000. C'est la date-repère pour la séparation de l'espagnol et du

portugais. Au début du X^{e} siècle, le roi des Asturies et de Galice avança sa capitale jusqu'à Léon, au sud de la cordillère cantabrique. Pour se prémunir contre les algarades de l'ennemi, la zone déserte qui séparait Chrétiens et Musulmans fut hérissée de châteaux (*Castella* au neutre pluriel) et placée sous l'autorité d'un comte. L'un de ceux-ci, Fernán González (935–970) se rendit indépendant. Devenu héréditaire, le comté de Castille finit par être annexé par le roi de Navarre, Sancho el Mayor (1000–1035), et, à la mort de celui-ci, érigé en royaume au bénéfice de son fils cadet Fernando. C'est donc vers la fin du X^{e} siècle que le castillan prend les caractères (Ramón Menéndez Pidal 1964) qui l'éloignent du groupe plus occidental, dont le portugais assure la survivance littéraire.

C'est à la même époque que se produit la rupture de l'unité scandinave, le suédois se séparant du dano-norvégien. Les équipages des drakkars mêlaient des Scandinaves de toute la Péninsule; la formation des monarchies isola les deux groupes.

1150. "L'ancienne unité de la Lotharingie disparait définitivement. Avec ses trois principautés épiscopales, ses deux duchés (Brabant et Limbourg), ses comtés de Hainaut, de Namur, de Luxembourg, de Hollande, de Gueldre, dont plusieurs provinces de la Hollande et de la Belgique conservent encore les noms, elle prend pour des siècles une physionomie nouvelle. Désormais, ce pays tout féodal échappe à l'Empire. En 1185, sollicité par l'empereur Henri VI de prendre part à une expédition contre Philippe-Auguste, le comte de Hainaut, Baudouin V (1171–1206), répond négativement. Les raisons de son refus sont caractéristiques: il invoque la neutralité de sa terre entre la France et l'Allemagne" (E. Lavisse et A. Rambaud, t. III 1894). Cet éloignement politique peut avoir facilité une différenciation lexicale.

1325. Le norvégien et le danois, différents par la phonétique, sont très voisins par le lexique. Il est possible que l'Union de Kalmar (1397) entre les trois royaumes scandinaves ait rendu moins exclusive la suprématie danoise. En tout cas, les langues étaient déjà bien formées au début du XIX^{e} siècle, et il semble peu vraisemblable que les événements de cette époque tardive aient pu avoir une sensible influence.

Cette revue historique manifeste que les "fruits" de notre arbre sont d'une qualité fort honorable, et ne présentent avec les faits de l'Histoire

ou de la protohistoire aucune contradiction choquante qui puisse les faire taxer d'absurdité.

* * *

Comment se fait-il, demanderont certains,que des arbres puissent produire des "fruits" de qualités aussi dissemblables? Ceci dépend de la façon dont ils ont été plantés et cultivés, et, dans le cas présent, de la relation qui a été établie entre les différences lexicales des langues et les dates de leurs divergences.

Sur un même graphique (fig. 8) portant en abscisse les valeurs des pourcentages d'accord K, et en ordonnée les dates de divergences de deux langues (ou familles de langues) T, nous avons tracé deux courbes.

L'une, en trait plein, correspond à l'équation que nous avons établie (Henri Guiter 1977a, 1977b, 1979, 1983), $T = 1900 - 5070 \operatorname{colog} K$.

L'autre, en tirets, correspond à une relation de type swadeshien, pour laquelle nous avons choisi un coefficient qui nous permette de serrer de près les résultats de S. Embleton. C'est une simulation des valeurs qu'elle obtient pour les dates de divergences. $T = 1900 - 10750 \operatorname{colog} K$.

Dans ces deux équations nous remplacerons K par une dizaine de pourcentages de convergences lexicales fournis par nos propres mesures. Bien entendu, nous nous contentons de lire les valeurs de T sur les abaques. Il s'agit de séparations étudiées à la fois par S. Embleton et par nous-même, donc dans les domaines roman ou germanique.

Nous comparons les résultats de cette simulation à certains des résultats obtenus par S. Embleton, et à ceux que nous avons obtenus nous-même.

Cette double série de points est indiquée sur les courbes.

L'allure des deux courbes nous laissait prévoir que, pour $K > 60\%$ et $T > -475$, les dates fournies par la méthode Swadesh seraient plus basses que celles découlant de notre méthode.

Au point $K = 60\%$ et $T = -475$ les deux courbes se coupent; dès lors, la courbe Swadesh escalade allègrement les millénaires. Pour $K = 20\%$, elle atteint déjà $T = -5600$, et sort donc du cadre de notre graphique. Pour $K = 1\%$, elle atteindrait $T = -19600$!

Ceci nous explique la date de -4100 qui apparaît pour les langues wakash. Voici six petites tribus indiennes, dont l'effectif total atteignait

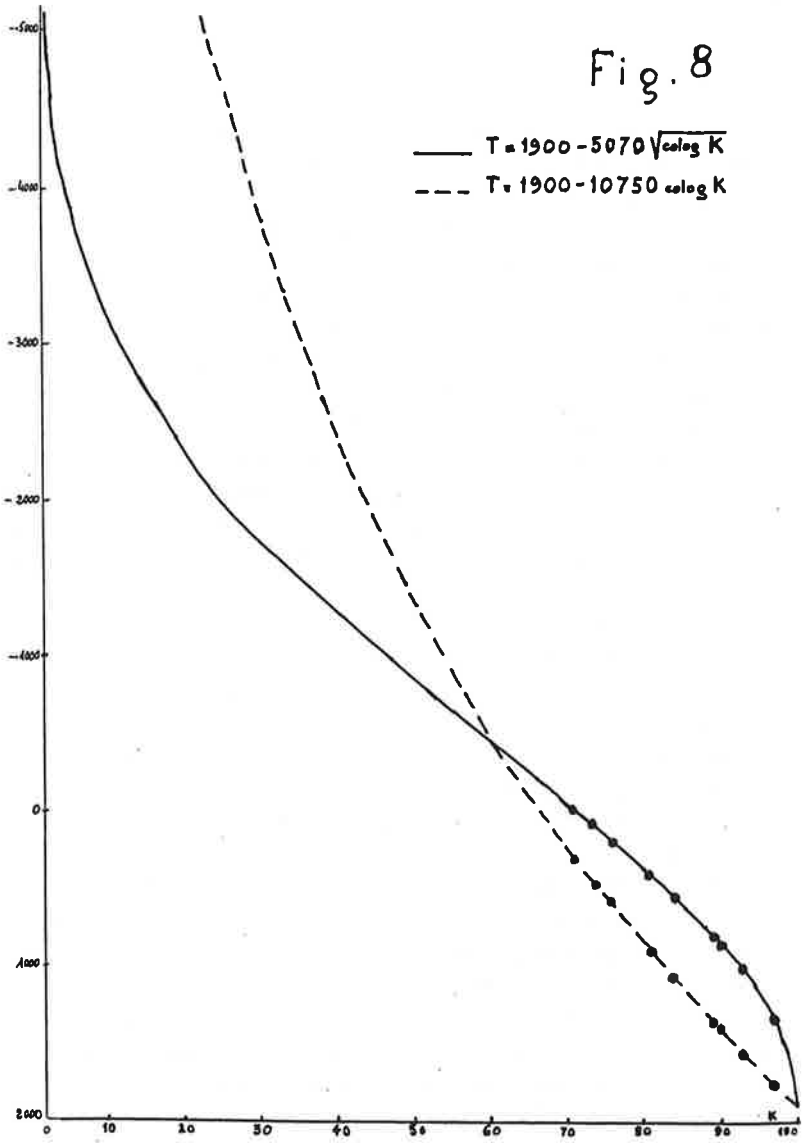


Fig. 8

Languages comparées	Résultats simulation	Résultats Embleton	Résultats Guiter
Norvégien/ Danois	1775	1768	1325
Portugais/ Espagnol	1550	1551	1000
Danois/ Suédois	1550	1540	1000
Islandais/ Dano-norvégien	1375	1224	825
Portugais/ Catalan	1350	1372	775
Frison/ Néerlandais	1350	1327	775
Italien/ Ibéro-roman	1050	952	525
Italien/ Français	900	952	375
Italien/ Roumain	650	544	200
Anglais/ Allemand	475	504	25
Scandinave/ German. occid.	300	295	-50

15.000 individus en 1780; ce nombre s'est bien réduit présentement (A. Meillet et M. Cohen 1952). Et il faudrait admettre que les langues de ces minuscules groupes humains, qui pêchent le poisson de l'Océan Pacifique au voisinage de Vancouver, se seraient isolées "depuis plus de quatre mille ans" (avant notre ère), pour reprendre l'expression d'un joyeux cantique de Noël! Evidemment, on ne peut rien prouver dans ce monde amérindien, où les données historiques font totalement défaut; c'est tout de même difficile à croire.

En définitive, à quoi tient la différence entre les équations de ces deux courbes, génératrices l'une d'arbres qui portent de bons fruits, l'autre d'arbres aux mauvais fruits? Elle tient essentiellement à l'introduction d'un symbole de racine carrée.

Ce n'est pas ici le lieu d'expliquer par quelles recherches nous sommes parvenu à cette équation; elles sont détaillées dans notre bibliographie (Henri Guiter 1977a, 1977b, 1989, 1983).

Le résultat étant acquis, on peut tout de même essayer d'en proposer une justification. M. Swadesh, et S. Embleton après lui, ont mesuré l'évolution du lexique en la rapportant au temps astronomique. Or le langage est un phénomène humain par excellence, et devrait donc être rapporté à un temps humain.

Lorsqu'on avance en âge, chacun a l'impression que le temps (le temps des horloges, le temps astronomique) passe de plus en plus vite. Il y a de plus en plus d'unités de temps astronomique à l'intérieur d'une unité de temps humain. Un individu change beaucoup plus entre 1 et 2 ans qu'entre 31 et 32 ans. Ces considérations et diverses observations personnelles nous ont conduit à admettre qu'il existe entre le temps humain et le temps astronomique une relation du second degré. Le phénomène joue en sens inverse si l'on remonte le temps au lieu de le descendre. Par conséquent, si le temps humain varie comme le carré du temps astronomique, inversement le temps astronomique varie comme la racine carrée du temps humain. Et si nous voulons exprimer les dates de divergences lexicale en "temps des horloges", il faudra introduire le signe "racine carrée" sur la fonction variable "colog K". C'est, d'ailleurs, pour cette raison que nous sommes passé des logarithmes aux cologarithmes: K est un pourcentage, donc un nombre compris entre 0 et 1; son logarithme est négatif, et ne saurait avoir de racine carrée réelle.

Bien entendu, cet essai d'explication psycho-physiologique ne s'impose nullement, et chacun est en droit de l'accepter ou de le rejeter. Ce n'est d'ailleurs pas lui qui nous a guidé dans notre recherche, et c'est *a posteriori* qu'il s'est présenté à notre esprit. Le seul fait qui demeure, c'est que l'introduction de la fonction "racine carrée" dans l'équation représentative, permet d'éliminer les absurdités historiques, qui avaient été précédement dénoncées à juste titre.

Notre critique des "arbres" de S. Embleton s'assortit donc d'une contreproposition positive, susceptible de désarmer les contempteurs de la glottochronologie,qui la taxaient de dangereuse absurdité. Par contre, elle montre le réel danger encouru, lorsqu'on prétend substituer le robot à l'homme dans les activités où le bon sens de celui-ci est indispensable.

Notes

1) S.M. Embleton, 1986.
2) Jean Séguy, 1973, p. 12.
3) Henri Guiter, 1984a.
4) Henri Guiter, 1977a, 1977b, 1979, 1983.
5) Henri Guiter, 1983, 1985.
6) Amado Alonso 1943.
7) Walther von Wartburg 1955.
8) Antoine Meillet 1949.
9) Carlo Tagliavini 1961.
10) José Joaquim Nunes, 1959.
11) Paul Russel-Gebbett 1965.
12) Eugenio Coseriu 1965.
13) Henri Guiter 1984b.
14) Henri Guiter 1985.
15) Antoine Meillet 1943.
16) Jacques Raison 1963.
17) Maurice Louis 1948.
18) E. Lavisse et A. Rambaud, t. I 1893.
19) Ramón Menéndez Pidal 1964.
20) E. Lavisse et A. Rambaud, t. III 1894.
21) A. Meillet et M. Cohen 1952.

Bibliographie

Alonso, A. (1943). "Partición de las lenguas románicas de Occidente", *Miscellania Fabra*: 81. Buenos Aires.

Coseriu, E. (1965). "Critique de la glottochronologie appliquée aux langues romanes". *Actes du X^e Congrès International de Linguistique Romane* I: 87. Paris.

Embleton, S. (1986). "Statistics in Historical Linguistics". *Quantitative Linguistics*, vol. 30, Bochum.

Guiter, H. (1977a). "Glottochronologie et langues occidentales". *Cahiers de l'Institut de linguistique* 2: 1. Louvain.

Guiter, H. (1977b). "A propos de glottochronologie". *Cahiers de l'Institut de Linguistique* 5-4: 3. Louvain.

Guiter, H. (1979). "Encore la glottochronologie". *Cahiers de l'Institut de Linguistique* 5-4: 3. Louvain.

Guiter, H. (1983). "Origines européennes et glottochronologie". *Quantitative Linguistics*, vol. 18: 136. Bochum.

Guiter, H. (1984a). "Datation de divergences romanes". *Revue de Linguistique Romane* 48: 269. Strasbourg.

Guiter, H. (1984b). "Essai de chronologie brittonique". *Actes du 107^e Congrès National des Sociétés Savantes* II: 77. Paris.

Guiter, H. (1985). "Corrélations de méthodes génético-linguistiques". *Cahiers de l'Institut de Linguistique* 11: 125. Louvain.

Lavisse, E., et Rambaud, A. (1893 – 1894). *Histoire Générale*, I: 585 et III: 418. Paris.

Louis, M. (1948). *Préhistoire du Languedoc Méditerranéen*. Nîmes.

Meillet, A. (1943). *Aperçu d'une histoire de la langue grecque*: 9. Paris.

Meillet, A. (1949). *Caractères généraux des langues germaniques*: 4. Paris.

Meillet, A., et Cohen, M. (1952). *Les langues du Monde*: 978. Paris.

Menéndez Pidal, R. (1964). *El idioma español en sus primeros tiempos*: 156. Madrid.

Nunes, J.J. (1959). *Crestomatia Arcaica*. Lisboa.

Raison, J. (1963). "Chronologie des tablettes cnossiennes". *Minos* VII, 2: 163. Salamanca.

Russel-Gebbett, P. (1965). *Medieval Catalan Linguistic Texts*. Oxford.

Séguy, J. (1973). "La dialectométrie dans l'Atlas Linguistique de la Gascogne". *Revue de Linguistique Romane* 37: 1. Strasbourg.

Tagliavini, C. (1961). *Grammatica comparata delle lingue germaniche*: 78. Bologna.

Wartburg, von W. (1955). "L'articulation linguistique de la Romania". *Actes du VII^e Congrès International de Linguistique Romane*: 37. Barcelona.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 201–218

Usage Trends in the Vocabulary of François Rabelais

Heath Tuttle
William Stanish
Cary, North Carolina

Ut silvae foliis pronos mutantur in annos,
Prima cadent: ita verborum vetus interit aetas,
Et iuvenum ritu florent modo nata vigentque.*

Introduction

What language is at any one moment in time, and how it changes across time are questions that have been pondered since ancient times. This study focuses on the suffixal derivatives that were used by the sixteenth-century French writer, François Rabelais, examining how the pattern of suffixal usage by him changed over the thirty-two-year period in which he wrote, specifically which suffixes showed significant trends of increasing or decreasing usage. For if thirty-two years is long enough to detect trends in his patterns of suffixal usage, some light may be shed on lexic dynamics in French, and perhaps upon that of other languages.

Methodology

Some seventy suffixes were active in derivation in sixteenth-century French. The frequencies of these derivatives in the Five Books of Rabelais, written in 1532, 1534, 1546, 1552, and 1564, were recorded.**

* G.B. Bonino, ed., L'Arte di Q. Orazio Flacco, Seconda ed. (Torino: Giovanni Chiantore, 1921), p. 10

** Charles Marty-Laveaux, ed., Œvres (Paris: Lemerre, 1868-1903), 6 vols.

The resulting 6,000+ words, sorted by suffix, were then dated, using standard historical-etymological dictionaries, and divided into two groups: those formed during the sixteenth century, and those formed earlier. Table 1 shows the frequencies of the pre-sixteenth-century works, counting every occurrence of a given derivative, for each of the Five Books.

Table 2 shows the corresponding results for sixteenth-century words. Table 3 shows the frequencies of the pre-sixteenth-century words, counting only the first occurrence in time of a given derivative, for each of the Five Books. Table 4 shows the corresponding results for sixteenth-century words.

The frequency of the derivatives of each suffix in each Book was divided by the total number of suffixal derivatives in that Book, using the total-occurrence totals from Tables 1 and 2, or the first occurrence totals from Tables 3 and 4, respectively. Since the total number of suffixal derivatives in a work is approximately proportional to the total number of words in it, this division had the effect of adjusting each frequency for the size of the Book. The resulting proportions were then tested for increasing or decreasing usage over time using a Mantel-Haenszel test of trend. The reported p-value inversely measures the strength of the evidence that there is a trend over time, so that the closer the p-value is to zero, the greater the evidence. P-values less than .05 are generally regarded as significant evidence of a trend. The results, using total-occurrence tallies, are shown in Table 5 for pre-sixteenth-century words, and in Table 6 for sixteenth-century words; and, using first-occurrence tallies, in Table 7 for pre-sixteenth-century words, and in Table 8 for sixteenth-century words. Suffixes were included in the tables only if their associated p-value was less than .05, and if there were at least ten occurrences of derivatives of the suffix over the Five Books. A smaller number of occurrences was regarded as insufficient evidence of a trend.

Discussion

A major implication of the latter four tables (5-8), where significant usage trends are listed, is the fact that the large majority of the some seventy suffixes active in the sixteenth century are not included in the tables, suggesting that the extent of their usage did not change

significantly: the numbers of their derivatives do not show marked increase nor decrease as Rabelais used them across the middle third of the sixteenth century.

Other, more specific trends are suggested too. For example, in Tables 5 and 6 (total-occurrence trends), *-ure* derivatives increase in their numbers, meaning that the number of *-ure* words, both of pre-sixteenth and sixteenth-century genesis, increased across the period. By contrast, *-aud* derivatives of pre-sixteenth-century origin decrease in number, while sixteenth-century derivatives of *-aud* increase in numbers. Thus older *-aud* words were falling out of favor, while newer *-aud* words were coming into favor.

In comparing the total-occurrence tables (5 and 6) with the first-occurrence tables (7 and 8), it may be noted that increasing usage in the total-occurrence tables for pre-sixteenth-century words is largely duplicated in the first-occurrence tables. Eight of eleven suffixes (e. g. *-icque*) showing increasing usage in the total-occurrence table for pre-sixteenth-century words (Table 5) also show increasing usage in the first-occurrence table for pre-sixteenth-century words (Table 7).

Conclusion

We have looked at a single writer's written vocabulary across time, segmenting it into a part already existing by his early childhood, and a part created by him or his contemporaries, to consider possible relations between those segments of vocabulary, to make more apparent the linguistic processes of a given period, and to look for usage trends of these two segments across time.

The authors extend great thanks to Andy Wilson and Ronnie Gale Emden for extensive access to their personal computer, and for their programming acumen. Discussions with Warren Sarle and Dave DeLong, and programming assistance from Lee Richardson, helped lead to the present analysis.

Table 1
Frequency of Pre-Sixteenth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-able	20	30	43	42	25
-ade/-ades	12	9	14	27	14
-age, -aige/-ages, -aiges	28	42	42	47	26
-aille/-ailles	6	9	12	12	13
-ailler	6	13	11	11	3
-aillon	0	2	0	1	0
-aire, -ayre/-aires	17	16	39	52	23
-al/-aulx, -aux	28	52	80	62	50
-ament	2	0	4	2	0
-amment	0	0	1	4	2
-ance/-anse/-ances	12	22	32	21	15
-ande/-andes	3	4	2	6	4
-ard/-ards	20	31	33	38	12
-art	12	22	20	28	23
-asse	3	3	4	7	2
-asser	12	14	11	15	12
-astre	2	3	2	0	4
-ateur/-ateurs	9	3	2	0	4
-atif	4	3	11	6	3
-ation	20	23	92	78	35
-atoire, -atoyre	0	3	11	6	3
-aud, -auld, -ault, aulx	24	28	32	32	10
-culer	2	3	2	1	0
-ean/-eans	0	0	0	0	1
-eau/-eaulx, -eaux	27	40	37	90	29
-ee/-ees	34	34	24	102	26
-eis, -ois, -oys	19	14	31	38	14
-elet	2	6	6	9	14
-elle/-elles	20	41	26	44	21
-ement	49	83	151	151	61

Table 1 (Continued)
Frequency of Pre-Sixteenth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-en	3	6	3	7	6
-er	333	623	612	627	316
-ere/-eres	6	13	9	6	14
-erie, -erye/-eries	8	29	16	19	15
-eron	4	1	2	3	1
-esque/-esques	1	2	2	3	4
-esse/-esses	8	12	6	19	15
-et/-ets	44	55	42	69	31
-ete (<Latin -itatem)	8	10	20	9	2
-ete (<Latin -itta)	4	9	6	16	7
-eter, -etter	14	15	15	12	8
-eton/-etons	2	3	1	3	0
-ette/-ettes	24	28	26	33	11
-eur, -our/-eurs, -ours	85	122	122	119	70
-eux, -eulx, -euse	19	41	67	74	34
-ficque, -fique/-fiques	3	8	17	27	34
-ian	1	5	2	12	1
-icque, -ique, -ic/-icques, ...	17	24	97	51	51
-ie, -ye/-ies, -yes	48	56	147	128	63
-ien/-yen	6	21	25	24	18
-ier, -yer/-iers (adj, nn, vb)	76	122	108	122	47
-iere, -yere	18	24	22	35	8
-if/-ives	8	9	33	20	11
-ile, -yle/-iles, -yles	7	10	22	21	11
-ille/-illes	12	23	9	22	4
-iller	1	4	2	6	1
-iment	0	3	4	0	0
-in/-ins	54	87	78	105	104
-ir, -yr/-irs (adj, adv, nn, vb)	91	128	138	127	86
-iser, -yser, -izer, -yzer	7	13	13	11	12

Table 1 (Continued)
Frequency of Pre-Sixteenth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-isme	4	5	7	4	5
-isson	0	4	0	4	2
-iste, -yste/-istes	5	8	6	3	5
-ite, -yte/-ites, -ytes (<-ita)	4	5	20	14	14
-ite, -yte/-ites, -ytes (<-itas)	15	16	51	44	24
-oir	42	88	121	85	30
-oire, -oyre/-oires	17	14	20	25	12
-on/-ons	65	91	438	107	65
-ot/-ots	19	25	30	53	21
-oter, -otter	3	8	11	9	5
-ouiller	0	5	6	7	5
-oyer, -oier	3	13	11	18	3
-u	19	31	44	49	20
-ure/-ures	10	20	29	24	26
Total	1511	2392	3258	3136	1638

Table 2
Frequency of Sixteenth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-able	1	2	5	4	2
-ade	1	4	6	23	11
-age, -aige/-ages, -aiges	3	2	15	12	8
-aille/-ailles	0	0	4	2	6
-ailler	0	0	0	0	1
-aillon	0	0	0	1	1
-aire, -ayre/-aires	2	2	1	1	1
-al/-aulx, -aux	3	1	3	2	0
-ament	1	0	0	0	0
-amment	0	0	1	0	0
-ance/-ances	0	0	1	1	1
-ande	0	0	0	1	0
-ard/-ards	3	4	3	0	3
-art	0	8	3	5	1
-asse	0	0	4	3	1
-asser	3	0	1	4	0
-astre	0	0	1	0	0
-ateur/-ateurs	0	0	0	1	4
-atif	1	2	0	0	0
-ation	1	0	6	2	2
-atoire, -atoyre	1	0	1	1	0
-aud/-ault, -aut/-aulx, -aux	7	8	4	3	21
-culer	2	0	0	0	0
-ean/-eans	0	0	0	1	0
-eau/-eaulx, -eaux	2	3	1	9	3
-ee/-ees	2	5	4	7	5
-eis, -ois, -oys	1	0	1	5	2
-elet	1	1	0	0	0
-elle/-elles	1	0	0	2	0
-ement	20	15	34	34	18

Table 2 (Continued)
Frequency of Sixteeth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-en	0	1	0	0	0
-er	16	9	36	37	15
-ere/-eres	1	0	1	0	0
-erie	5	7	3	6	6
-eron	1	0	0	1	1
-esque/-esques	1	0	1	0	0
-esse/-esses	3	0	1	5	8
-et/-ets	4	5	10	10	7
-ete (<Latin -itatem)	0	0	0	1	1
-ete/-etes (<Latin -itta)	0	1	0	2	0
-eter, -etter	5	1	1	0	0
-eton	0	2	0	0	0
-ette/-ettes	3	7	3	6	4
-eur/-eurs, -ours	13	9	9	8	6
-eux, -eulx, -euse	3	5	11	3	1
-ficque, -fique/-fiques	0	0	2	0	1
-ian	0	0	0	0	3
-icque, -ique/-icques ,-iques	3	2	11	7	5
-ie, -ye/-ies, -yes	3	2	4	3	2
-ien, -yen/-iens	1	0	1	1	3
-ier, -yer/-iers (adj, nn, vb)	9	10	15	19	9
-iere, -yere/-ieres, -yeres	3	2	5	5	5
-if/-ives	0	0	0	1	0
-ile, -yle/-yles	0	0	0	1	0
-ille, -ylle/-illes	1	1	0	0	1
-iller	0	1	0	3	0
-iment	1	0	0	0	0
-in/-ins	2	3	4	11	3
-ir, -yr/-irs (adj, adv, nn, vb)	0	1	0	2	0
-iser, -izer, -yser, -yzer	2	6	3	5	1

Table 2 (Continued)
Frequency of Sixteeth-Century Words, by Suffix
(Total Occurrences)

Suffix	Book				
	1	2	3	4	5
-isme, -ysme	0	1	2	0	1
-isson	0	0	0	0	0
-iste, -yste/-istes	2	4	6	2	0
-ite (<Latin -ita)	1	0	0	1	1
-oir	1	1	3	1	0
-oire, -oyre/-oires	1	0	0	5	0
-on/-ons	4	7	2	12	8
-ot/-ots	1	2	4	4	1
-oter, -otter	4	3	5	0	2
-ouiller	0	1	0	1	0
-oyer	0	1	5	0	0
-u	0	1	1	5	2
-ure/-ures	0	3	1	1	7
Total	150	156	249	293	195

Table 3
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

Suffix	Book				
	1	2	3	4	5
-able	11	8	15	15	9
-ade/-ades	10	6	8	15	5
-age, -aige/-ages, -aiges	17	11	15	10	3
-aille/-ailles	5	4	7	6	7
-ailler	4	2	2	1	1
-aillon	0	1	0	0	0
-aire, -ayre/-aires	11	10	24	15	8
-al/-aulx, -aux	16	21	32	16	14
-ament	2	0	3	1	0
-amment	0	0	1	4	1
-ance/-anse/-ances	8	14	14	7	4
-ande/-andes	3	0	0	1	0
-ard/-ards	14	4	8	1	0
-art	7	8	7	3	1
-asse	3	2	2	3	1
-asser	7	5	2	3	1
-astre	2	3	1	0	1
-ateur/-ateurs	7	3	12	14	3
-atif	3	3	9	3	2
-ation	19	19	57	32	14
-atoire, -atoyre	0	3	3	1	0
-aud, -audt, -ault, -aut/-aulx	10	7	5	2	2
-culer	2	1	0	0	0
-ean/-eans	0	0	0	0	1
-eau/-eaulx, -eaux	16	28	15	32	7
-ee/-ees	27	16	15	14	9
-eis, -ois, -oys	7	6	9	4	1
-elet	1	2	5	2	1
-elle/-elles	15	18	11	9	5
-ement	41	48	62	53	19

Table 3 (Continued)
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

Suffix	Book				
	1	2	3	4	5
-en	1	1	2	1	0
-er	230	259	215	131	49
-ere/-eres	4	4	3	0	6
-erie, -erye/-eries	8	19	11	4	6
-eron	4	1	0	2	1
-esque/-esques	1	2	2	1	2
-esse/-esses	7	7	2	3	0
-et/-ets	30	33	18	25	7
-ete (<Latin -itatem)	6	7	12	0	1
-ete/-etes (<Latin -itta)	3	6	2	5	1
-eter, -etter	11	5	5	3	2
-eton/-etons	2	2	1	1	0
-ette/-ettes	20	16	15	18	4
-eur, -our/-eurs, -ours	54	44	48	28	28
-eux, -eulx, -euse	16	28	35	26	5
-ficque, -fique/-fiques	3	3	5	1	4
-ian	1	2	1	2	1
-icque, -ique, -ic/-icques ...	15	18	58	21	21
-ie, -ye/-ies, -yes	35	28	82	41	26
-ien/-yen	3	7	4	3	2
-ier, -yer/-iers (adj, nn, vb)	62	60	44	32	13
-iere, -yere	15	11	6	9	1
-if/-ives	7	7	21	1	4
-ile, -yle/-iles, -yles	5	7	11	11	4
-ille/-illes	12	11	3	9	0
-iller	1	4	2	4	1
-iment	0	2	2	0	0
-in/-ins	35	34	30	35	27
-ir, -yr/-irs (adj, adv, nn, vb)	35	22	32	14	14
-iser, -yser, -izer, -yzer	6	7	5	5	3

Table 3 (Continued)
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

Suffix	Book				
	1	2	3	4	5
-isme	3	3	2	0	5
-isson	0	2	0	3	0
-iste, -yste/-istes	5	3	3	3	3
-ite, -yte/-ites, -ytes (< -ita)	4	3	16	5	7
-ite, -yte/-ites, -ytes (< -itas)	14	12	32	18	13
-oir	11	15	15	5	2
-oire, -oyre/-oires	12	4	8	3	2
-on/-ons	45	40	39	33	22
-ot/-ots	14	14	9	6	5
-oter, -otter	3	6	3	4	2
-ouiller	0	2	5	1	3
-oyer, -oier	2	5	5	5	0
-u	9	13	12	5	3
-ure/-ures	9	10	14	8	8
Total	1031	1042	1184	802	428

Table 4
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

Suffix	Book				
	1	2	3	4	5
-able	1	2	3	1	1
-ade	1	3	6	14	4
-age, -aige/-ages, -aiges	3	1	11	5	3
-aille/-ailles	0	0	3	1	4
-ailler	0	0	0	0	1
-aillon	0	0	0	1	0
-aire, -ayre/-aires	2	2	1	1	1
-al/-aulx, -aux	3	0	2	1	0
-ament	1	0	0	0	0
-amment	0	0	1	0	0
-ance/-ances	0	0	1	0	1
-ande	0	0	0	1	0
-ard/-ards	2	2	1	0	1
-art	0	8	1	1	1
-asse	0	0	3	2	0
-asser	2	0	1	2	0
-astre	0	0	1	0	0
-ateur/-ateurs	0	0	0	1	1
-atif	1	2	0	0	0
-ation	1	0	4	2	2
-atoire, -atoyre	1	0	1	1	0
-aud/-ault, -aut/-aulx, -aux	3	4	1	3	9
-culer	2	0	0	0	0
-ean	0	0	0	1	0
-eau/-eaulx, -eaux	2	2	1	7	2
-ee/-ees	2	4	2	5	2
-eis, -ois, -oys	1	0	1	4	0
-elet	1	1	0	0	0
-elle/-elles	1	0	0	2	0
-ement	17	12	20	15	7

Table 4 (Continued)
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

Suffix	Book				
	1	2	3	4	5
-en	0	1	0	0	0
-er	14	6	21	15	3
-ere/-eres	1	0	1	0	0
-erie	5	4	3	4	4
-eron	1	0	0	1	1
-esque/-esques	1	0	1	0	0
-esse/-esses	3	0	1	4	7
-et/-ets	4	4	5	5	4
-ete (<Latin -itatem)	0	0	0	1	0
-ete/-etes (<Latin -itta)	0	1	0	1	0
-eter, -etter	3	1	1	0	0
-eton	0	1	0	0	0
-ette/-ettes	3	5	3	3	0
-eur/-eurs, -ours	9	7	8	6	3
-eux, -eulx, -euse	2	4	10	2	1
-ficque, -fique/-fiques	0	0	2	0	1
-ian	0	0	0	0	3
-icque, -ique/-icques ,-iques	3	2	11	6	5
-ie, -ye/-ies, -yes	3	2	1	1	2
-ien, -yen/-iens	1	0	1	1	2
-ier, -yer/-iers (adj, nn, vb)	9	7	8	10	4
-iere, -yere/-ieres, -yeres	3	2	3	4	1
-if/-ives	0	0	0	1	0
-ile, -yle/-yles	0	0	0	1	0
-ille, -ylle/-illes	1	1	0	0	1
-iller	0	1	0	2	0
-iment	1	0	0	0	0
-in/-ins	2	3	3	8	3
-ir, -yr/-irs (adj, adv, nn, vb)	0	1	0	2	0
-iser, -izer, -yser, -yzer	2	5	1	2	1

Table 4
Frequency of Sixteenth-Century Words, by Suffix
(First Occurrences)

-isme, -ysme	0	1	1	0	0
-isson	0	0	0	0	0
-iste, -yste/-istes	2	4	3	1	0
-ite (<Latin -ita)	1	0	0	0	1
-oir	1	1	3	1	0
-oire, -oyre/-oires	1	0	0	4	0
-on/-ons	4	7	2	9	3
-ot/-ots	1	2	2	3	0
-oter, -otter	3	0	1	0	0
-ouiller	0	1	0	0	0
-oyer	0	1	1	0	0
-u	0	1	1	2	1
-ure/-ures	0	3	1	1	2
Total	131	122	164	172	93

Table 5
Pre-Sixteenth-Century Words Showing a Usage Trend
in the Sixteenth Century
(Total Occurrences)

Trend	Suffix	P-value	Words
Decreasing Usage	-er	.000	2511
	-ier, -yer/-iers	.000	475
	-oir	.005	366
	-eur, -our/-eurs, -ours	.008	518
	-ille/-illes	.008	70
	-aud, -auld, -ault, -aut/-aulx	.014	126
	-ette/-ettes	.023	122
Increasing Usage	-ficque, -fique/-ficques	.000	71
	-icque, -ique, -ic/-icques, ...	.000	240
	-in/-ins	.000	428
	-ation	.000	248
	-ie, -ye/-ies, -yes	.004	442
	-ite, -yte/-ites, -ytes (<-ita)	.004	57
	-ateur/-ateurs	.011	87
	-aire, -ayre/-aires	.012	147
	-ite, -yte/-ites, -ytes (<-itas)	.013	150
	-ure/-ures	.022	109
	-eux, -eulx, -euse	0.37	235

Table 6
Sixteenth-Century Words Showing a Usage Trend
in the Sixteenth Century
(Total Occurrences)

Trend	Suffix	P-value	Words
Decreasing Usage	-eur/-eurs, -ours	.008	45
	-iste, -yste/-istes	.045	14
Increasing Usage	-ade	.002	45
	-aille/-ailles	.005	12
	-esse/-esses	.013	17
	-ure/-ures	.031	12
	-aud, -ault, -aut/-aulx, -aux	.034	43

Table 7
Pre-Sixteenth-Century Words Showing a Usage Trend
in the Sixteenth Century
(First Occurrences)

Trend	Suffix	P-value	Words
Decreasing Usage	-er	.000	884
	-ier, -yer/-iers	.001	211
	-ard	.004	27
	-esse/-esses	.021	19
	-ille	.027	35
	-iere, -yere	.042	42
Increasing Usage	-icque, -ique, -ic/-icques, ...	.000	133
	-ie, -ye/-ies, -yes	.000	212
	-ation	.001	212
	-ite, -yte/-ites, -ytes (<-itas)	.003	89
	-ite, -yte/-ites, -ytes (<-ita)	.004	35
	-aille	.013	29
	-in	.014	161
	-able	.015	58
	-aire	.034	68
	-ouiller	0.36	11
	-ateur	.043	39

Table 8
Sixteenth-Century Words Showing a Usage Trend
in the Sixteenth Century
(First Occurrences)

Trend	Suffix	P-value	Words
Increasing Usage	-esse/-esses	.003	15
	-ade	.020	28
	-aud, -ault, -aut/-aulx, -aux	.024	20

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 219–233

Review

Research Design and Statistics for Applied Linguistics, by Evelyn Hatch & Hossein Farhady. Rowley, Mass.: Newbury House 1982, Pp. XVII + 290.

reviewed by

Rüdiger Grotjahn
Bochum

As the title of Hatch and Farhady's book indicates, it is written especially for people working in the field of applied linguistics. By applied linguistics the authors mean primarily second language acquisition research and language testing. Thus, this book is not intended for quantitative linguists primarily interested in the quantitative description and modelling of language. Nevertheless, I think that the contents of Hatch and Farhady's book as well as the following discussion will be of interest also to quantitative linguists, and in particular to those adhering to a broad conception of the field. (The reader interested in basic statistical techniques relevant to the quantitative analysis of linguistic data is also referred to my review of Butler (1985) in this volume.)

Since its publication the book by Hatch and Farhady has become quite popular amongst researchers, teachers and students in the field of applied linguistics. Quite a few applied linguists refer to it for their research designs and for their statistical analysis of the data obtained. It is also used in courses in research methodology and statistics for applied linguists. There are several possible reasons for this success. Firstly, it has filled a gap in the literature on applied linguistics, where a book on research design and statistics was badly missing. Secondly, it is a very elementary introduction to the topic – in wide parts rather a "cookery-book", giving recipes instead of a critical presentation of the topic. Since the vast majority of researchers in applied linguistics have had little or no systematic training in statistics and empirical methodology, a book of this kind may be rather appealing. Thirdly, reviews

such as that by Birch (1984), where it is characterized as "a very sound teaching textbook on research design, method, and evaluation (that) should prove invaluable to both students and teachers alike" (p. 64) might also have contributed to its success.

In the following I will argue that, although it has its merits, Hatch and Farhady's book is in most parts quite perfunctory, often misleading and not seldom even incorrect. Because of lack of space I cannot present all possible evidence in support of this argument but have to restrict myself to a discussion of major flaws in their book. Other minor flaws such as some misprints in formulas, computational errors, unexplained abbreviations, misleading wordings or inconsistencies in the mathematical notation, which may lead to some confusion, will not be commented upon. I hope the attentive reader will notice them himself.

In Chapter 1 entitled "What is research?", research is defined as "a systematic approach to finding answers to questions" (p. 1). The authors distinguish different steps in doing research and explain what they mean by systematic. Further, concepts such as hypothesis, internal validity and external validity are discussed. This chapter contains some valuable suggestions. However, the treatment of the concept of null hypothesis (and alternative hypothesis) may be somewhat misleading. According to the authors a null hypothesis is stated if one is uncertain which way to direct a hypothesis (p. 3). This is not correct because in significance testing a null hypothesis may also be formulated if one is sure about the direction. (The authors treat this topic more clearly in Chapter 8.) Furthermore, the authors seem to imply that a null hypothesis never predicts either a positive or a negative relationship between variables (p. 4). This is, however, only a very special case of the use of the null hypothesis in statistics. It is in my opinion more adequate to conceive of the null hypothesis as the hypothesis actually statistically tested and that one hopes to nullify (for a justification see e.g. Hays 1974, Chapter 9).

In Chapter 2 entitled "What is a variable?", the authors first briefly discuss the three most important measurement scales for variables (nominal, ordinal, interval). They then distinguish different types of variables such as dependent, independent, moderator, control, and intervening. Unfortunately, the fundamental problem of measurement

and scales is treated in a very cursory and even misleading way. They state for example: "School grade is also a nominal variable which could have 1 to 12 levels ... Again the numbers have no mathematical value; they only serve to identify which Ss belong in which level of the variable" (p. 13). This statement is incorrect because the numbers are not only used to *identify* the subjects (Ss), but tell us also something about the number of years the Ss have been attending school. Thus school grade has to be considered as measuring at least on an ordinal scale. Nevertheless, we can *treat* school grade as a nominal variable, but in doing so we loose important information.

In their discussion of the interval scale in the same chapter the authors state with regard to the problem of the equality of the intervals: "For example, if age is the variable we are researching, we can assume that a year is a year. But the value of a year may differ along the scale. For second language learning, the difference between each year in the 10- to 20-year range may be much larger in value than the year intervals between 60 and 70" (p. 14). Here the authors obviously confound the physical variable of age conceived as the time elapsed since birth (which constitutes a ratio scale and thus also an interval scale) with the effect different intervals of equal length of the variable age may have on another variable such as the rate at which a second language is learned.

Their treatment of the level of measurement of test scores on page 14 of Chapter 2 seems to imply that we always should consider test scores as measuring on an interval scale. However, if we doubt – and often we should – that a specific test really measures on an interval scale, we can also treat the scores as an ordinal variable and apply statistical procedures appropriate for ordinal variables.

Nowhere do Hatch and Farhady discuss *systematically* the important relation between scales of measurement and the appropriateness of statistical procedures. There are only some casual remarks in various chapters of their book. They mainly treat procedures especially suited for interval scale data and most of the numerous procedures developed especially for nominal or ordinal variables are neither discussed nor mentioned. This is very regrettable since most data in the social sciences (including applied linguistics) are measured only on a nominal or ordinal level. (The reader who is interested in procedures for

nominal and ordinal variables may refer, for example, to the classical work by Siegel (1956) or to the more modern, although less elementary treatment in Marascuilo and McSweeney (1977) or Meddis (1984). Some important non-parametric procedures, such as the U-test, are also treated by Butler (1985)).

In Chapter 3 entitled "Constructing research designs" the authors explain research designs such as pre-experimental, true experimental, quasi-experimental, ex post facto, and factorial. In spite of the title of Hatch and Farhady's book, the topic is not dealt with very extensively (although more extensively than in Butler 1985). I would like to make only one short remark: In their discussion of a time-series design, the authors state with regard to the figure presented on page 25: "If your results were like those in Line 2, it would indicate that your materials had a negative effect, since after the treatment their scores declined consistently." However, without a control group this interpretation is, although very plausible, not sufficiently substantiated because factors other than the treatment may also have lead to a decline such as in Line 2.

Chapter 4 briefly deals with "The research report format". The authors refer among other things to the APA format, a format proposed by the American Psychological Association which is used by most of the research journals in psychology and applied linguistics in North America. I think that this chapter contains some information valuable not only for American but also for European students.

In Chapter 5 entitled "Sorting and displaying the data", the coding of different types of data, the computation of some simple statistics (ratio, proportion, percent, rate, percent change, relative frequency, cumulative frequency) and the displaying of the data in the form of tables and figures is explained. Here, the fact that the scores 21, 23, 26, and 28 have zero frequencies is not taken into account in the frequency polygon on page 48. Furthermore, the treatment of the topic of symmetric and asymmetric curves (pp. 50ff.) is much too simplistic and even misleading. For example, not every symmetric bell-shaped curve is a normal curve, as the authors seem to suggest, and asymmetry is not only a question of "outliers".

In Chapter 6 entitled "Describing the data", statistics such as mode, median, mean, standard deviation, and variance are explained. What

in my opinion is badly missing is a *systematic* discussion of the appropriateness of the different statistics with regard to the different measurement levels. There is, for example, no indication that, to be strict, the mean, standard deviation and variance should only be calculated for interval scales. (On page 80 the mean, median, mode, standard deviation, variance, and range are characterized as measures for ordinal *and* interval data.) Furthermore, although grouped data are presented, the student is not told how to calculate the statistics in this case. One would also expect an indication as to how to calculate the mean of means or percentages – a topic where students (and even researchers) often make errors.

In Chapter 7 the concepts of standardized scores and normal distribution are discussed. There are several statements which are highly misleading or even incorrect. The authors claim for example: "The outcome of almost all human behavior is a normal distribution. No matter what kind of scale is used, no matter what kind of behavior is investigated or what type of data are gathered, the distribution of scores of large samples tends to be normal" (pp. 63f.). What is meant by large is explained on page 64 where we are told that "it is usually the case that if the N (the number of cases) is 30 or more, ... we will not violate the assumptions of the normal distribution drastically". However, only behavior that can be explained as the result of the combined effect of independent random variables tends to be normally distributed. Furthermore, the normal distribution is a continuous distribution. In applied linguistics we often have also *discrete* variables which follow for example a binomial or multinomial distribution. Such a binomial distribution, for example, may be *approximated* by a normal distribution, but with $N = 30$ this approximation is only acceptable if the binomial distribution is symmetric and is very poor if the distribution is highly skewed.

In discussing the concept of standardization the authors tell us nothing about the fact that standardizing scores from an observed distribution which is asymmetric does not lead to a normal distribution but to a distribution which is still asymmetric although it now has a mean of 0 and a variance of 1 as does the normal distribution. In such a case an interpretation of the z-scores in terms of the area of the normal distribution may be highly misleading. Hence the statement

"Between $\bar{X}$ and $\pm 2s$, 95% of the scores should occur. And between $\bar{X}$ and $\pm 3s$, 99% of the observations are accounted for." (p. 65) is only correct if $\bar{X}$ and s stand for μ and σ of the normal distribution and not for the mean and standard deviation of any observed distribution. (To be exact, in the last quotation, the numbers should read 1.96 and 2.58 and not 2 and 3.)

In Chapter 7 the authors explain for the first time how to use the computer program SPSS for statistical analysis. Unfortunately in this introductory section there are many statements which make sense only with reference to a special local computer system or operating system. Only one example: On page 73 the student is told (without any further comment or warning that this may be quite different with his local computer system): "To tell the computer you want to change the password, just punch the new password after the old one and separate the two with a comma." (I would be happy if it were that easy at my local computer site.)

Chapter 8 deals with the fundamental problem of probability and hypothesis testing. Here different types of null and alternative hypotheses (one-tailed, two-tailed, directional, nondirectional) are presented and the process of decision making with the help of significance testing is explained. Although the difficult issues discussed in this chapter necessarily need to be simplified in an introductory text, the presentation would in my opinion have been much clearer if the authors had introduced the fundamental concepts of "type I error" (also called "alpha error" or "error of rejection"), "type II error" (also called "beta error" or "error of acceptance") and "test power". With the help of these concepts the notions of null hypothesis and alternative hypothesis could have been defined in a more rigorous way and one could, for example, have made clear that acceptance of the null hypothesis does not necessarily mean that the null hypothesis is *correct*, but that it may actually be false and has only be accepted because of insufficient power of the test (insufficient number of Ss). Furthermore, statements such as the following, which is highly misleading because it does not take into account the possibility of a type II error, could have been avoided: "If we claim significance at the .001 level (and in social sciences that's almost bragging about how sure we are!), there is still 1 chance in 1,000 we could be wrong. While there are still chances that we are wrong, by

doing the analysis we have drastically reduced the possibilities of making a mistake." (p. 106) (Similar misleading statements can be found in various places in the book.)

What is also missing in this context is a discussion of the concept of "practical significance" (as opposed to that of statistical significance). If the number of Ss increases, the power of a statistical test increases too. Thus, if the number of Ss is only just large enough, even minimal deviations from the null hypothesis will lead to a statistically significant result, which may however be of little or no practical significance. For example, a correlation coefficient of 0.1 may be highly significant (for large N), but its practical significance measured by the amount of variance explained is almost zero. Therefore, judging the *importance* of a finding only on the basis of a significance test as do Hatch and Farhady in various places (cf. e.g. p. 135) is not correct. Since not only students but also some researchers in applied linguistics seem to know little or nothing about the various problems involved in significance testing, a more thorough treatment of this topic would have been very desirable. (A more thorough although not very elementary treatment can e.g. be found in Hays 1973 or Cohen 1977.)

In Chapter 9 the authors introduce the concepts of sampling distribution of means and sampling distribution of differences between means and explain how to use these distributions in hypothesis testing. Here it does not become clear enough that if the population variance is not known and must be estimated by the sample variance the t-test presented in the next chapter and not the z-test is the most adequate procedure. In the case of an unknown population variance the z-test presented in this chapter is only an approximation to the t-test, which with a relatively small sample size as that in the example on page 99 ($N = 36$) may lead to rejection of the null hypothesis in cases in which the t-test would lead to acceptance. A similar argument holds with regard to the z-test of the difference between two means presented on page 105. For this and other reasons it is in my opinion advisable to use the procedures presented in this chapter only with really large samples.

In Chapter 10 the t-test for one mean, two means from independent samples and two means from matched samples is presented. Unfortunately, although the authors tell us that the "t-test is one of the most

useful statistical procedures (and one of the most frequently used procedures) for research in Applied Linguistics" (p. 120), their presentation is very superficial and even incorrect in several respects. Since I can discuss only some major flaws, the reader is referred to the correct presentation of the *t*-test for example in Hays (1973) or Walker/Lev (1969). (The latter offers a more elementary treatment of the topic.)

On page 114 Hatch and Farhady tell us that in using the *t*-test we have to assume that "the scores on the independent variable are continuous and that there are only two levels to the variable" and that "the variances of the scores in the population are equal, and the scores are normally distributed." Furthermore, on page 119 they claim that "the scores on the independent variable should be measured on an interval scale." This is incorrect. Firstly, it is not the independent but the dependent variable that has to be continuous and has to be measured on an interval scale. The independent variable has to be discrete (dichotomous) and has to be measured on a nominal scale. Secondly, the variances of the scores in the population need not be equal. On the contrary, the formula presented on page 113, which is the only formula for the *t*-test of two independent random samples given by Hatch and Farhady, is especially suited if the population variances *cannot* be assumed to be equal. If the variances are assumed to be equal, the formula used by the SPSS program and mentioned but not further commented upon on page 126 should be used. This formula may yield quite different results if sample sizes and/or variances differ considerably. Furthermore, with different sample sizes and unequal variances the number of degrees of freedom should not be computed as indicated by Hatch and Farhady but as in the SPSS example on page 127. (Hatch and Farhady do not mention this problem at all.) Thirdly, only for small samples do the scores need to be normally distributed. Even in the case of smaller sample sizes the *t*-test is quite robust with regard to violations of the assumption of normality especially if in the two-sample test the population variances and the sample sizes are equal. Therefore, in the case of small samples, we should try to provide for approximately equal sample sizes when we do a *t*-test.

There are still some other flaws to be mentioned. (1) On page 124 the authors deal with *matched* samples, but erroneously present the formula for the standard error of differences between means from two

independent samples. (2) Hatch and Farhady do not show how to test the important assumption of equality of variances. Since in the SPSS example on page 127 the *F*-test is used to test this assumption, this test, which is very simple to apply but which is not unproblematic, should have been commented upon at least briefly. (3) The reader is not told what to do when the assumptions on which the *t*-test is based are not met. Here, a reference to or better a section about the non-parametric *U*-test, which is a very efficient procedure and which can in contradistinction to the *t*-test also be used with ordinal data, should have been included. (4) Only the statistical significance but not the practical significance of the outcome of a *t*-test is discussed. Since even a very small difference between two means will become statistically significant provided the sample size is large enough, we need a measure which indicates the amount of variance in the dependent variable accounted for by the independent variable. Such a measure can be found for example in Hays (1973: 417). (The point biserial correlation coefficient discussed in Chapter 15 could also be used but this possibility is not mentioned by Hatch and Farhady.) If the measure yields a small value, even a highly significant *t*-value should be considered to be without much practical significance. If in the case of relatively small sample sizes the value obtained is high but the *t*-test not statistically significant, this may be interpreted as an indication that there may be an association between independent and dependent variable worth being investigated on the basis of larger samples.

In Chapter 11 the authors present both one-way analysis of variance (ANOVA) and preplanned and post hoc comparisons for testing the differences of two or more means. In Chapter 12 ANOVA of factorial designs is introduced and the important topic of statistical interaction is discussed. Compared to the chapter about the *t*-test this section of the book is much more sound and will prove valuable in particular to readers not very familiar with this topic. Nevertheless, I would like to make at least the following points: (1) There is no discussion of the assumptions underlying the statistical models presented. It should for instance have been mentioned that equality of variances has to be assumed and that this assumption can be tested, but that with equal sample sizes this assumption is not so important. (2) ANOVA is presented without any further comment as a test of interval and ordinal

data. Applying ANOVA to ordinal data may however be problematic. (3) There is no indication as to what to do when the assumptions are not met. Here at least a reference to suitable nonparametric tests as an alternative to ANOVA would have been helpful. (4) As with the t-test the issue of practical significance should have been discussed and measures been presented.

Chapter 13 treats the well-known χ^2-test for nominal data. As in most sections of the book the authors do not attempt to teach an understanding of the rationale behind the methods and formulas presented but first of all try to explain how to *apply* them. Although such an approach is certainly often necessary in a book primarily aiming at beginners, there are not a few cases (e.g. the formula for the expected frequency on page 168 or the rationale behind the correction factor and the corrected chi-square formula on page 171) where a better understanding could have been easily achieved. With regard to the correction factor the authors state for example: "If you do a one-way χ^2 and have only 1 d.f. [...] you will need to correct the estimate so that it fits the χ^2 distribution for d.f.'s over 1. [...] If the observed value is larger than the expected value, subtract .5 from the observed value. If the observed value is less than the expected value, add .5 to the observed." (p. 171) Here the authors should have explained *why* to add or subtract just .5 from the observed value and mentioned that this is a *correction for continuity* used to obtain a better approximation to a χ^2 distribution with 1 d.f. (and not as erroneously argued to the χ^2 distribution for d.f.'s over 1). Some other points: (1) I see no justification why the section "Working with the computer" is missing in this particular chapter although there exists a corresponding SPSS subprogram, namely CROSSTABS. (2) There is no discussion of measures of association such as the φ statistic or Cramér's V, which are calculated by CROSSTABS and which can be used to evaluate the practical significance of an obtained χ^2-value. (3) It should at least have been mentioned that there are (modern) alternatives to the χ^2-test, which are often more appropriate. (According to Hatch and Farhady "the χ^2 is probably the best statistical procedure to use" (p. 172); for a good, although not very elementary presentation of alternative procedures see Bishop, Fienberg and Holland 1975).

Chapter 14 deals with the method of implicational scaling – a

method applied for example in sociolinguistics and interlanguage studies but not treated in most (elementary) statistics books. The topic is dealt with in a rather sound way and problems with implicational scaling are discussed at some length. I think that this chapter should prove to the applied linguist to be one of the most valuable parts of Hatch and Farhady's book.

In Chapter 15 the important topic of correlational analysis is discussed in more than twenty pages. The following coefficients are presented: Pearson product moment correlation, point biserial correlation and Spearman rank-order correlation. The discussion of how to interpret correlations and the section about factors affecting correlation coefficients should be mentioned as positive features of this chapter. Although some important issues, for example that of part-whole correlation, which in my opinion should have been treated at some length in a textbook such as this, are not touched upon, I think that the quality of correlational analysis in applied linguistics will significantly improve if what the authors tell us is taken into consideration.

However, as in almost all chapters of this book, the treatment of the statistical aspects of the topic is unsatisfactory. A few points: (1) The formulas for the Pearson product moment correlation coefficient presented on page 198f. are *not* algebraically equivalent as argued by Hatch and Farhady. They will only then yield the same value if the formula based on z-scores is divided by $N-1$ and not by N. (By the way one more example of the lack of care in the presentation of statistics can be found on the top of page 199 where one and the same symbol is used for the cross product of the deviation scores and for the covariance.) (2) On page 204 the authors state: "When one of the variables in the correlation is nominal (e.g., sex, grade), the point biserial correlation is used to determine the relationship between the levels of the nominal variable and the continuous variable. The nominal variable does not have to be normally distributed. For example the number of male vs. female or foreign students vs. native speakers can have quite a skewed distribution." This is highly misleading, if not incorrect. The point biserial coefficient is only used when one of the variables is genuinely dichotomous. Furthermore, a true nominal variable is a discrete variable and therefore *cannot* be normally distributed. (The authors seem to equate symmetry with normality. If a variable is really normally

distributed but dichotomized, not the point biserial coefficient but the biserial coefficient should be used.) (3) With regard to the Spearman correlation coefficient we are not told what to do when there are ties in the rankings. (The problem of ties is even not mentioned at all.) Furthermore, the reader should be warned that nowadays most statistics book use the symbol "ρ" to refer to the Pearson correlation coefficient for populations and not, as Hatch and Farhady do, to the Spearman correlation coefficient, the latter being often designated by "r_s" (e.g. in the SPSS manual). (4) In order to provide a better understanding of the topic, it should have been mentioned that the formulas for the point biserial correlation and for the Spearman correlation are nothing other than algebraic simplifications of the formula for the Pearson product moment correlation. (5) The formula for testing the statistical significance of r with the help of the t-statistic on page 207 (and page 285) is incorrect. The correct formula for the t-statistic or the equivalent F-statistic can be found in Guilford and Fruchter (1978: 142) or Blalock (1979: 418).

In Chapter 16 the methods of simple linear regression, multiple linear regression, and partial correlation are presented. There are again some flaws in the presentation of the mathematical basis. (1) The two formulas for the standard error of estimate presented on page 224 and cited again on page 228 are *not* equivalent as the authors argue. (The formula based on the correlation coefficient is not an unbiassed estimate.) (2) On page 226, since $d.f._{reg} = K$, $d.f._{res}$ should equal $N - K - 1$ and not $N - 1 - 1$ and on the top of page 229 $d.f._{reg}$ should be equal to K and not to $K - 1$. (3) The formula given in point 14 on page 229 as well as in the appendix on page 286 is incorrect in several respects. (SPSS calculates the standard error of b and not that of standardized beta. For the correct formula see Nie et al. (1975: 326)). (4) On page 222 not the value for r, 0.92, but for the slope, 1.40, should be inserted in the regression equation. (5) Important assumptions such as that of collinearity are not discussed at all. (6) In the SPSS example on page 227ff., in calculating the correlation between the variable GRAM (which refers to a subtest, i.e. to a *part* of the whole test) and the variable TOTAL (which refers to the *whole* test), GRAM has been also correlated with itself – which leads to a spuriously inflated correlation between the two variables. This problem, which is generally

referred to as "part-whole correlation" (see e.g. Guilford and Fruchter 1978: 331 and 465), should at least have been mentioned (cf. also my remarks on Chapter 16).

Chapter 17 deals with the important issue of reliability and validity. Again, it is primarily the treatment of the statistical aspects that is not satisfactory. Some examples: (1) Contrary to Hatch and Farhady's argument, the formula presented on the bottom of page 247 is neither the Kuder-Richardson formula KR-20 nor Cronbach's α but a standardized α (a generalization of the Spearman-Brown formula presented on the same page). Furthermore, it is not true that KR-20 is also referred to as Cronbach's α. Rather, KR-20 is a special case of α for dichotomous data (cf. Hull and Nie 1981: 256). (2) It should have been mentioned that the Kuder-Richardson formula 21 on page 248 should only be used if the item difficulties are the same. (3) There is no warning that estimating the reliability by the method of internal consistency presupposes that the items are experimentally independent. If this condition is not met, as is the case in cloze tests, KR-20 or Cronbach's α overestimate the reliability. (In the literature on cloze tests for example this fact has been quite often overlooked.)

In Chapter 18 the topic of factor analysis, including the relationship of factor analysis, reliability and validity, as well as an SPSS example are dealt with. Because this very complex and difficult topic has been treated on only ten pages, only some major aspects could be discussed and numerous simplifications were necessary. There are however also some instances where the treatment is very misleading, if not even incorrect. On page 262 for example the authors state with regard to the Varimax rotated factor matrix: "Notice that it is quite different from the unrotated factor matrix. Here the factors are orthogonal (i.e., they are uncorrelated with each other and independent of each other)." This statement implies that the factors of the unrotated matrix are not independent of each other, which is certainly not true because the unrotated factors (initial factors) are orthogonal as well (cf. Nie et al. 1975: 470). Nevertheless I think that in spite of some shortcomings this chapter will be useful as a very elementary introduction to factor analysis in applied linguistics.

Finally, I would like to give a more global appreciation of Hatch and Farhady's book. On the basis of what has been said, the following

conclusions can be drawn: (1) There are numerous flaws in the book. Some of the flaws may be due to the necessity of simplification. The vast majority, however, cannot in my opinion be so easily excused. (2) There should be a more thorough treatment of the relation between scales and statistics and of the logic of statistical inference. (3) Important nonparametric tests, such as the U-test, should have been presented or at least mentioned. Furthermore, a section dealing with the important topic of sampling methods should have been included. (4) The treatment should be less dogmatic and patronizing. Instead, the various assumptions and problems involved should be more thoroughly discussed. (5) The presentation could often be more concise. Even in a textbook for beginners one could for example very well do without the numerous lengthy presentations of elementary calculations such as subtraction or addition of two simple numbers (e.g. on page 222 and 248). This would give room for the teaching of really important issues. (6) There should be precise references to the relevant literature. The bibliography, which is sadly limited, should be extended and also more recent publications should be included. (7) Hatch and Farhady's book is to a great extent rather a statistical "cookery-book". The authors should first of all have tried to teach an *understanding* of the methods presented. This would have been easily possible as the example of other elementary textbooks shows (see the references given below and my review of Butler 1985 in this volume).

I think that the rather detailed discussion of flaws has been necessary to help the unsophisticated reader to avoid mistakes when using Hatch and Farhady's book. Furthermore it should have become clear that the applied linguist should not rely primarily on this book but (in addition) should consult well-established textbooks for psychologists or sociologists such as Guilford and Fruchter (1977), Hays (1973), Mueller, Schuessler and Costner (1977), Walker and Lev (1969) or Kerlinger (1973) – the last named offering also an extensive introduction to research methodology. What is in my opinion the most positive feature of Hatch and Farhady's book is the fact that is is written especially for applied linguists and includes numerous examples from this field. Furthermore it is very elementary. A book with such a conception was missing in applied linguistics. I hope that this extensive review will help ensure that both a revised edition of Hatch and Farhady's book

as well as other future publications on research design and statistics for (applied) linguistics will be less subject to criticism than the present work.

References

Birch, D. 1984. "Review of E. Hatch & H. Farhady, Research design and statistics for applied linguistics," *Applied Linguistics* 5, 63-65.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. 1975. *Discrete multivariate analysis: Theory and practice.* Cambridge, Mass.: MIT.

Butler, C. 1985. *Statistics in linguistics.* Oxford: Basil Blackwell.

Cohen, J. 1977. *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Guilford, J.P., and Fruchter, B. 1978. *Fundamental statistics in psychology and education.* Tokyo: McGraw-Hill Kogakusha.

Hays, W.L. 1973. *Statistics for the social sciences.* London: Holt, Rinehart and Winston.

Hull, C.H., and Nie, N.H. 1981. *SPSS update 7-9: new procedures and facilities for releases 7-9.* New York: McGraw-Hill.

Kerlinger, F.N. 1973. *Foundations of behavioral research.* London: Holt, Rinehart and Winston.

Marascuilo, L.A., and McSweeney, M. 1977. *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA.: Brook/Cole.

Meddis, R. 1984. *Statistics using ranks: a unified approach.* Oxford: Basil Blackwell.

Mueller, J.H., Schuessler, K.J., and Costner, H.L. 1977. *Statistical reasoning in sociology.* Boston: Houghton Mifflin.

Nie, N.H. et al. 1975. *SPSS: Statistical Package for the Social Sciences.* New York: McGraw-Hill.

Siegel, S. 1956. *Nonparametric methods for the behavioral sciences.* New York: McGraw-Hill.

Walker, H.M., and Lev, J. 1969. *Elementary statistical methods.* New York: Holt, Rinehart and Winston.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 235–245

Review

Quantitative Psychology: Some Chosen Problems and New Ideas, by Maria Nowakowska. Amsterdam: North-Holland 1983. XXX + 943 pp. (Advances in Psychology 15)

reviewed by

Rüdiger Grotjahn
Bochum

Before starting a detailed discussion of the book under review, it should be stated that Nowakowska's book is quite exceptional. It is to a great extent interdisciplinary in character and covers a very wide range of topics on nearly 1000 pages. This makes it relatively difficult for any reviewer to give a fair account of all its aspects. As a consequence and due to the present reviewer's varying degree of familiarity with the topics dealt with, a number of aspects which are possibly worth more detailed discussion will be mentioned in the following only briefly or not at all.

Chapter 1 (114 pages) gives a rather detailed outline of the theory of psychological tests. After a brief discussion of the object of measurement in psychology and of the notion of trait, the statistical foundations of the 'classical' theory of mental test scores are presented and concepts such as reliability, validity, homogeneity, parallelism and experimental independence are discussed in considerable detail. In this context topics such as prediction and test bias are also touched upon. The outline follows Lord and Novick's classical work rather closely. The presentation is critical and the assumptions on which the theory is based are very clearly stated. Extensions of the theory in form of the theory of generic true scores and generalizability as developed by Cronbach and others are briefly discussed.

The chapter ends with a very brief presentation (only four pages) of latent trait theories such as the Rasch model. The title, namely "An outline of the contemporary theory of psychological tests", is somewhat misleading since most of the chapter is devoted to a discussion of

classical test theory, dealing with more modern approaches either very briefly, as is the case of the theory of generalizability or latent trait models (for a good treatment of these topics cf. e.g. de Gruijter and van der Kamp 1984), or not at all, as for example approaches based on the concept of criterion referenced measurement (for this approach cf. e.g. Hambleton et. al. 1978 or Klauer 1987). Therefore, the chapter should rather be considered as a good and also very readable introduction to classical test theory – some knowledge of (elementary) statistics presupposed.

Chapter 2, entitled "Factor Analysis: Arbitrary Decisions within Mathematical Model", is by far the shortest chapter in the book (only 68 pages). It is devided into two main sections. The first section contains a good methodological discussion of models of (exploratory) factor analysis. Multidimensional scaling as an alternative to factor analysis is also briefly discussed. It is pointed out "that multidimensional scaling is based on weaker assumptions, using only the order information contained in the data, and not the numerical values (which are usually subject to much higher error)" (p. 144), and that this results in greater applicability. It is shown that applications of factor analysis involve various arbitrary decisions, which bias the results by subjectivity. The decisions concern for example the number of factors to be extracted, the type of rotation and the final interpretation of factors. Furthermore, it is pointed out that the results are also dependent upon the sample and the measurement tools used. It should be noted that the author primarily deals with exploratory factor analysis and that most of the criticism does not apply to confirmatory factor analysis as a statistical method for testing hypotheses about multivariate observations (for a good treatment of this approach see McDonald 1985).

In the second section of Chapter 2 the abstract methodological considerations of the first section are illustrated by a detailed analysis of an example of extensive application of factor analysis, namely the personality theory of Cattell. Although there in no coherent presentation of empirical data to support the author's criticism, I think that the methodological discussion of Cattell's work is sound and of interest not only to researchers in the field of personality.

Chapter 3 (102 pp.) deals with some psychological problems of construction and application of questionnaires. The chapter is devided

into three main parts. The first part deals with models of response to questionnaire items. The author starts with the construction of a (fuzzified) model of variability and ambiguity of questionnaire items. Subsequently, she analyzes the well-known psychometric paradox that more stable (less ambiguous) items tend to have less discriminating power. On the basis of this analysis she concludes very correctly:

> "From the point of view of the purpose of measurement, to satisfy some psychometric criteria, however important, cannot serve as a means for itself, since the main issue is to collect the richest possible psychological information about the subjects" (p. 204).

In the following section the factors determining the answer to a questionnaire item are studied on the basis of sample of 56 students. In this study a questionnaire consisting of 17 scales for evaluating questionnaire items is presented. Then 28 items are chosen from the 16 Factor Personality Questionnaire of Cattell. Each subject first had to respond to an item and then to evaluate the item and his response on the 17 scales. Finally, on the basis of a factor analysis the author proposes a model of answering a questionnaire item. In view of the criticism of factor analysis presented in Chapter 2 as well as other possible arguments against factor analysis as a method for obtaining information about the mechanisms underlying observable behavior (cf. p. 207 and e.g. Carroll 1983 for a critique), I think that the validity of the model has still to be shown.

In the final section of part I responding to a question in an interview (as well as responding to a test item) is conceived as a multiple-criterion decision making process. This process is conceptualized as a series of trial answers carried out mentally. A mathematical model is proposed which seems to describe the process well and also allows the reconstruction of sociological interview data by means of simulation.

In the second main part of Chapter 3 the author presents some alternatives to the standard approach to test construction. As a first alternative she discusses the concept of so-called contextual tests, which is based on the theory of unfolding scales. As a second alternative she deals with dynamic questionnaires, which are a special case of adaptive or tailored testing models. Here the items are presented in an order depending on the previous responses of the subject. The termination of testing is determined by a specific stopping rule. Hence, the total

number of items presented is not fixed but a random variable. Finally, the author proposes a random walk model for dynamic questionnaires without, however, discussing this model systematically. In my opinion it has several advantages compared to classical test theory to consider the total number of items presented to be a random variable. This holds true in particular for computerized testing (cf. e.g. Weiss 1983).

In the third main part of Chapter 3, it is shown how the results of test theory in combination with the theory of fuzzy sets can be used to measure the distance between certain fuzzy concepts, namely those measured by questionnaires. This approach, which is in my opinion not yet sufficiently worked out, is considered by the author to be a possible path of development of both measurement theory and test theory.

Chapter 4 (157 pp.) is entitled "Formal Semiotics: Representations, Observability and Perception". The author, who has published rather extensively on this topic, reports for the most part the results of her own studies in this chapter. Not only a theory is presented, but in some sense also a huge research program comprising many domains of cognition such as perception of natural language, information processing, logic and philosophy, computer science, etc.

In the first main section, the analysis deals with the interesting topic of multimedial communication, i.e. communication taking place simultaneously on several media (e.g. verbal, gestures, facial expressions). The author proposes a formal model of multimedial communication languages and discusses concepts such as expressibility, which relates to the possibility of expression when some of the media cannot be used. The units of the multimedial languages are represented by the theory of random graphs devised by the author, which takes into account the possibility of random appearance and disappearance of edges and nodes.

The next main section deals with control processes in perception. The model proposed breaks with the conception of perception as a passive process. It is assumed that several mechanisms participate in the perception process and that the latter is a result of their interaction. Three main mechanisms are modelled, namely those of spontaneous and purposeful inspection as well as that of transportation in comparisons. It should be noted that in this model perception is also

connected with memory as well as with problems of observability and change.

In the next main section the topic of observability is treated. The basic concepts of the theory of observability proposed are (random) fuzzy events, joint observability, masks and filters. Examples of (random) fuzzy events are notions such as inflation or knowledge. Joint observability relates to the fact that two attributes may be both observable, but not jointly, a fact that limits the scope of legitimate inference considerably. Masks describe the temporal patterns of variables to be observed as well as the sets of variables to be observed at any given time. The concept of filter relates to the possibility that although the observer decides to observe a variable, this variable may be unobservable (i.e. filtered in some way).

In the subsequent sections of Chapter 4 semantic aspects of representations of objects (or situations) are studied. The concepts 'system informationally complete' and 'algebra of description of (fuzzy) objects' are introduced. Furthermore, verbal copies are characterized with regard to their exactness and faithfulness, the latter referring to a multidimensional notion of truth and corresponding in some sense to the validity of a description. Chapter 4 closes with sections devoted to formal semiotic systems, signs and information as well as with a very short section on statistical semiotics. There are in my opinion very interesting ideas in this chapter. Nevertheless I think that the theory presented is still not very well worked out and that mathematics is here used mainly to formalize ideas rather than to construct models leading to real theoretical progress.

Chapter 5 (254 pp.) deals with "Selected topics in measurement theory". Measurement theory is conceived as a branch of mathematical psychology dealing with the existence and uniqueness of psychological measurement scales as well as with their construction, in particular of scales relating to perception. After discussing concepts such as ordinal measurement and standard sequences, an abstract presentation of the measurement problem is given based on the concepts of relational system and homomorphism. Then the following important measurement situations are discussed at some length: extensive measurement (exemplified by measurement of risk and the construction of a qualitative probability scale), difference measurement (in particular with regard to

the concept of just noticeable differences) and conjoint measurement (both additive and multinomial). Subsequently the topic of utility and subjective probability including the well-known axiom system by Luce and Krantz is dealt with.

In the next section some problems in utility theory are dicussed. As regards risk, a model is suggested that explains some apparently paradoxical (irrational) behavior, namely situations where increase in probability of success of an option (other conditions being equal) causes a person to reject the option. This is considered to be a consequence of specific incentives associated with risk taking. Thus the model explains situations in which the standard subjective expected utility (SEU) model does not apply. Since the latter is generally thought of as a standard of rationality, the theorems proved in this section offer in a certain sense some rationalization of irrationality. In this context a model is proposed which explains some phenomena of social behavior. Certain plausible postulates of utility are suggested taking into account some observed regularities of prosocial and antisocial behavior and a new type of non-linear utility function is introduced. The most important feature of this model lies in the fact that the utility index relates to the 'changes in utility' rather than to the utilities of the final states. In this way the author also tries to take into account the well-known fact of intransitivity of preferences.

The next section of Chapter 5 is devoted to an application of measurement theory to the construction of a theory of time, both objective and subjective. Such a theory is of fundamental importance for a wide variety of (empirical) disciplines including the study of cognitive processes and computer science. The formal theory presented takes the influence of the variable 'memory' upon the perception of time into account and thus allows us to decribe subjective distortions of time perception. Although the relevant theoretical and experimental literature on the psychology of time is hardly touched upon, I think that this section contains a series of valuable ideas which should be taken into account in the further development of (psychological) measurement theory.

Chapter 5 also contains a section on scaling. If we consider measurement to consist of assigning numbers to objects so as to reflect

adequately the property to be measured, measurement theory is concerned with the formal conditions which must be met in order for such an assignment to exist, whereas scaling theory deals with the techniques of assigning numbers to objects so as to reflect the property to be measured. Three basic types of scaling technique are discussed: (1) pair comparison, (2) magnitude scaling, and (3) rating and category scales. The topic of this section is illustrated by the results of an experiment carried out by the author herself concerning the perception of the power of members of various committees. The objective power is defined by the index of social power proposed by Shapley and Shubik, which is based on the concept of minimal winning coalition, that is the minimal number of persons sufficient to ensure the final decision regardless of the votes of the remaining members. The question as to whether the choices are consistant with this index is studied and interesting misperceptions of social power are ascertained.

Chapter 5 ends with a discussion of the problem of linguistic representation of knowledge about objects or situations. This problem, which is of central importance not only for psychology but also for linguistics or artificial intelligence (e.g. in the context of expert systems), is analyzed in the book several times from different points of view. In this section it is dealt with under the heading of "linguistic measurement" within the framework of classification theory. For the description of the dynamics of classification, formal classification languages are introduced. It is shown that it is possible to account for temporal characteristics of dynamic classification processes by making the assumption that "inter-classification" times are geometrically distributed and then estimating the waiting time for the mth classification of an object on the basis of this assumption. The section ends with a brief discussion of the role of analogies and metaphors in the construction of a theory of linguistic measurement.

Chapter 5, which is the longest chapter in the book, is certainly one of the most valuable parts of Nowakowkska's contribution to quantitative psychology. It is a good outline of contemporary measurement theory and contains many interesting ideas. Nevertheless one important objection against the theory presented should at least be mentioned. Measurement is conceived by the author as the representation of an empirical relational system by a numerical relational system. In

this context it does not become clear enough that psychological measurement is far more than that, namely a multi-level modelling process with a syntactical, semantical as well as pragmatical dimension. This important aspect, which has far-reaching implications for psychological theory formation, has been most clearly pointed out by Gigerenzer (1981).

In Chapter 6 (247 pp.), the last chapter of the book, a formal theory of actions is presented. The central idea of this theory – a preliminary version has already been developed in earlier publications by the author – is to treat composite actions as strings of elementary actions. Section 1 first deals with the basic deterministic case. Later on a version is presented in which the notion of admissibility of a string of actions is fuzzified. In the last part of Section 1 the model is further generalized by assuming that the transitions to the next state of an action are (at least partially) random, that is probabilistic. The probabilities are considered to be of a stationary Markov type.

The next two sections deal with temporal relations between events. To take these temporal characteristics into account the notion of 'history' is introduced and it is shown how the framework developed can be used to analyze knowledge representations and to characterize the process of development (progress).

Section 4 discusses issues that arise when one wants to attain multiple goals. An algebraic theory of conflicts is developed leading to a rich typology of conflicts. Subsequently the theory is applied to motivation problems in the form of a kind of motivational calculus.

In Section 7 the theory discussed so far, which concerns actions of a single person, is extended to cover the case of simultaneous actions of a group of persons. In the next section this framework is modified so as to characterize formally the concept of system synthesis in systems theory. Here the question to be answered is: Under what conditions should one treat a number of systems which operate simultaneously as one joint system?

The next sections briefly deal with topics such as preference systems and their relations to decision theory, generative grammars of planning, description of operation of organizations, the problem of ethical valuations.

Section 13 concerns the important problem of aggregation of evaluations by different persons and with respect to different criteria. A fuzzified version of the aggregation problem is formulated which appears to be similar to that of preferences under uncertainty. Finally an index of homogeneity of evaluations in a group is introduced.

In the next section the framework developed in Section 13 is applied to model alienation and dynamics of social change. On the basis of the proposed model some very interesting hypotheses concerning social change are formulated. I think it would be worthwhile to explore these hypotheses empirically.

In the last section an outline of how the formal theory of action developed so far may be applied to educational and developmental psychology is given. To this end development is described in terms of action languages and stadial languages, and the education process is treated as a certain process of stochastic control with incomplete information.

In sum, the last chapter of the book contains many interesting ideas relevant not only to psychology but also for example to (quantitative) linguistics. Some aspects of the model appear to be already relatively well developed, whereas others have been only touched upon. I think that researchers interested in this topic will find a number of valuable suggestions for further investigations.

Before I now attempt a more global appreciation of Nowakowska's book some minor details: (1) For a book of nearly 1000 pages an index would be very helpful. (2) There are not enough precise references to the relevant literature. These would be helpful for readers wanting to deepen their understanding of specific points. (3) To illustrate the use of the formal models presented, more data should be included and discussed at some length.

In sum, this interdisciplinary book is full of ideas of interest not only to the psychologist but also to other social scientists including linguists. Some chapters deal with standard topics such as for example Chapters 1 and 5, which present a clear mathematical exposition of the basic ideas of the theory of psychological tests and of measurement theory and constitute a good introduction to these topics for readers with some background in mathematics. Others discuss topics far less standard in quantitative psychology as for example formal semiotics

or formal theory of actions where much of the theory has been developed by Nowakowska herself. As the subtitle of the book "Some Chosen Problems and New Ideas" indicates, the book is clearly selective. As a consequence some important topics such as mathematical learning theory or hypothesis testing are only touched upon or not treated at all. The reader aiming at a more comprehensive understanding of quantitative psychology should therefore in addition consult publications such as Coombs, Dawes and Tversky (1970), Sydow and Petzold (1982), Townsend and Ashby (1983), Rettler (1985) and Hager (1987), to name only a few. To conclude, I think that in spite of some slight reservations Nowakowska's book can be clearly recommended and will be of much value not only to the quantitative psychologist but to the quantitative linguist as well.

References

Carroll, J.B. 1983. "Studying individual differences in cognitive abilities: through and beyond factor analysis." In Dillon, R.F., Schmeck, R.R. (eds). *Individual differences in cognition.* New York: Academic Press 1983: 1-33.

Coombs, C.H., Dawes, R.M., and Tversky, A. 1970. *Mathematical psychology. An elementary introduction.* Englewood Cliffs, N.J.: Prentice Hall.

De Gruijter, D.N.M., and Van der Kamp, L.J.Th. 1984. *Statistical models in psychological and educational testing.* Lisse: Swets & Zeitlinger.

Gigerenzer, G. 1981. *Messung und Modellbildung in der Psychologie.* München: Reinhardt.

Hager, W. 1987. "Grundlagen einer Versuchsplaung zur Prüfung empirischer Hypothesen in der Psychologie." In Lüer, G. (ed). *Allgemeine Experimentelle Psychologie.* Stuttgart: Fischer 1987. 43-264.

Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D.B. 1978. "Criterion referenced testing and measurement: A review of technical issues and developments." *Review of Educational Research* 48: 1-47.

Klauer, K.J. 1987. *Kriteriumsorientierte Tests.* Göttingen: Hogrefe.

McDonald, R.P. 1985. *Factor analysis and related methods.* Hillsdale, N.J.: Erlbaum.

Rettler, H.J. 1985. *Zur statistischen Prüfung psychologischer Hypothesen.* Braunschweig: Technische Universität (Braunschweiger Studien zur Erziehungs- und Sozialarbeitswissenschaft 16).

Sydow, H., and Petzold, P. 1982. *Mathematische Psychologie. Mathematische Modellierung und Skalierung in der Psychologie.* Berlin: Springer.

Townsend, J.T., and Ashby, F.G. 1983. *The stochastic modeling of elementary psychological processes.* Cambridge: Cambridge University Press.

Weiss, D.J. (ed). 1983. *New horizons in testing: latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Schulz, K.-P. (ed.):
Glottometrika 9.
Bochum: Brockmeyer 1988, 247–259

Review

Statistics in Linguistics, by Christopher Butler. Oxford: Basil Blackwell 1985. Pp. X + 214.

reviewed by

Rüdiger Grotjahn
Bochum

The book under review is intended for students and researchers in general and applied linguistics with no previous knowledge of statistics and only very basic mathematical knowledge. It describes basic statistical techniques relevant to the quantitative analysis of linguistic data and contains many worked examples and exercises to which detailed answers are provided. The data discussed refer to various fields of general and applied linguistics (including language acquisition and language testing) and are in many cases taken from actual projects in linguistics. The areas covered are for example: sentence length/word length in different texts; time taken to utter a sentence/to detransform a sentence in its simplest form; intensity of stressed and unstressed syllables; scores on a language test; distribution of pause length; relation between certain linguistic features and the politeness of an utterance; syntactic change in Middle English.

In contradistinction to Hatch and Farhady (1982) for example, who very often give recipes instead of teaching an understanding of the statistical methods presented (cf. my review in this volume), Butler makes a considerable effort to provide the reader with an at least rudimentary understanding of the statistical procedures discussed. I can only agree when he argues:

> "Many courses on applications of statistics concentrate far too heavily on the methods themselves, and do not pay sufficient attention to the reasoning behind the choice of particular methods. I have tried to avoid this pitfall by discussing the 'why' as well as the 'how' of statistics. [...] While recognising that most linguists (including myself) will have neither an interest in the more theoretical side of

the subject nor the mathematical background necessary for a full discussion, I feel that it is highly unsatisfactory for readers or students simply to be presented with a formula, with no explanation whatever of how it is arrived at. Where I thought it appropriate, I have attempted to give an idea of the rationale behind the various methods discussed in the book." (p. IX)

It is this attempt at teaching an understanding of the statistical procedures which in my opinion makes Butler's book clearly superior to, for example, Hatch and Farhady (1982) as an introduction to statistics for (applied) linguists.

In the following I will briefly comment on all twelve chapters of Butler's book. But first a general remark: Since most chapters are very short (the mean chapter length is only 14 pages), Butler's treatment is necessarily very selective and various simplifications were necessary. Therefore, some of the following comments should be considered not as criticism but rather as suggestions for some additional pages in a revised edition.

Chapter 1 deals with "Some fundamental concepts in statistics". In the first section the difference between samples and (finite and infinite) populations as well as between various sampling methods (random, quasi-random, block, stratified random, disproportionally stratified, multi-stage) are discussed with special reference to linguistic data. Subsequently, the distinction between parameters and statistics (estimates) and between the descriptive and inferential functions of statistics is briefly commented upon. Although sampling procedures are discussed with special reference to linguistic data, the relevant literature in quantitative linguistics is not taken into account at all. It should, for example, have been mentioned that almost all lexical samples and most samples of other linguistic entities as well are by necessity biassed in several respects. To give only one example: It has been shown that the smaller a sample is compared to the population, the more the mean relative lexical frequency in the sample overestimates the mean lexical probability in the population. Thus, since lexical samples are normally rather small compared to the population, we systematically overestimate at least part of the lexical probabilities (for a discussion of this and other problems in sampling linguistic data see e.g. Orlov 1982).

In the last section of Chapter 1 different types of variables such as dependent, independent, continuous, discrete, as well as different

levels of measurement of a variable are distinguished. It is pointed out that different statistical procedures are appropriate for different levels of measurement and that in current research practice this fact is very often not sufficiently taken into account.

In Chapter 2, entitled "Frequency distributions", the distribution of raw and grouped data and the displaying of the data in the form of histograms and frequency polygons is dealt with. The concepts of skewness and kurtosis (degree of peaking) are introduced and different kinds of skewness and kurtosis are conceptually distinguished (coefficients of skewness and kurtosis are not presented).

Chapter 3 deals with "Measures of central tendency and variability". The concepts of mean, median and mode are introduced and it is shown how to calculate these coefficients for both ungrouped and grouped data. Furthermore, it is shown that the choice of an appropriate measure of central tendency depends on the level of measurement of the variable concerned, the shape of the frequency distribution, and finally on whether the measure is suitable for intended further statistical analyses. Subsequently, some measures of variability, namely the range, the interquartile range, the quartile deviation, the mean deviation, the variance and standard deviation are briefly discussed. In most cases the rationale underlying a specific measure is pointed out and some simple derivations are given as well. What is in my opinion missing in this introductory treatment is an indication as to how to calculate the means of means or of percentages – a topic where students and even researchers often make errors.

In Chapter 4 the main properties of the normal distribution are discussed and illustrated with the help of several graphs. It is then shown how to use some properties of the normal distribution to quickly check for normality of an observed distribution. (A χ^2-test of fit is given in Chapter 9.) I think that in this chapter the author should have also told us that standardizing scores from an observed distribution with the help of the z-transformation leads to a distribution with mean 0 and variance 1, but does not influence the shape of the distribution and that therefore in case of a skewed distribution an interpretation of the z-scores in terms of the area of the normal distribution may be highly misleading. (By the way, in the first formula for the z-score on page 47 it is not $\bar{x}$ but the population parameter μ which should be inserted.)

Chapter 5, entitled "Sample statistics and population parameters: Estimation", covers the following topics: The sampling distribution of the mean; standard error of the mean and confidence limits; the t distribution; confidence limits for proportions; estimating the sample size required in order to estimate a population parameter with a given degree of accuracy. Here it does not become completely clear that in case we estimate the population variance by the sample variance, it is the t distribution and not the normal distribution which is the appropriate statistical model. If the sample size is large, we can *approximate* the t distribution by the normal distribution. For $N = 30$, the approximation is quite good but not so good that it should always be used in significance testing, as Butler seems to imply (cf. also my review of Hatch and Farhady 1982 in this volume). I think this latter point would have become clearer if Butler had compared the shape of the normal distribution with that of the t distribution for different degrees of freedom. Furthermore, with regard to the sampling distribution of a proportion which according to Butler is approximately normal if the product of N and p (or of N and $1 - p$ if this is smaller) is at least 5 (p. 62), it can be shown that it is much safer not to use the above criterion but the following one: $np(1-p) \geq 9$. (The same holds by the way for the test of the difference between two proportions on page 94.)

Chapter 6 is entitled "Project design and hypothesis testing: basic principles". In the section on project design Butler discusses the following topics: experimental vs. observational (correlational) studies; sources of unwanted variation and their minimisation; independent groups design vs. repeated measures design vs. matched subjects design. Although the whole topic of project design is treated in only 4 pages, the basic principles are clearly pointed out. In the section on hypothesis testing the following topics are dealt with: null hypothesis and alternative hypothesis; significance levels; types of errors in significance testing (type I error vs. type II error); directional (one-tailed) and non-directional (two-tailed) tests; parametric vs. non-parametric tests; criteria for choosing a test. The presentation of significance testing – a conceptually quite difficult topic – is very clear and didactically well thought out, and in my opinion far better than for example that of Hatch and Farhady (1982). What is missing is a discussion of the concept of "practical significance" (as opposed to that of statistical

significance). Even a statistically highly significant result may be of no practical significance, for even a very small deviation from the null hypothesis, i.e. a deviation with no or only small importance with regard to practical decisions, will always achieve statistical significance if the number of subjects is only just large enough. This fact would have become clearer if Butler had introduced the concept of "test power" and, in spite of his remark that "the factors affecting the value of β are beyond the scope of our discussion" (p. 72), had briefly commented upon these factors. I know from my own teaching experience in courses in applied statistics that a non-technical discussion of this fundamental problem would have been easily possible (for a more extensive discussion of the topic of significance testing see also my review of Hatch and Farhady 1982 in this volume).

Chapter 7, which is entitled "Parametric tests of significance", deals with the z-test for means and proportions from independent samples and with the t-test for both independent and correlated samples. Although this chapter is much better than the corresponding chapter in Hatch and Farhady (1982), there are some points which should be commented upon.

(1) On pages 75 and 78 Butler argues that the principal assumption underlying parametric tests is that the populations have a normal distribution. This holds true only for the parametric tests discussed by Butler, but not for parametric tests in general.

(2) At the bottom of page 78 Butler states with regard to the z-test of two means at the .05 level that this means asking the question "If we claim that the means of the populations from which the samples were drawn differ, shall we have at least a 95 per cent chance of being right?" Here Butler obviously confounds the type I error with the type II error because a significance level of .05 means that we calculate the probability of the obtained z-value under the assumption that the null hypothesis holds; that is, we claim that the population means are the *same*. To answer the question asked by Butler, however, means to calculate the *power* of the test.

(3) It does not become clear that if we have to estimate the standard deviations, the difference between two means divided by the standard error of the difference is *not* normally distributed

but follows a t distribution which can be approximated by the normal distribution provided the sample sizes are large. (I would be more conservative than Butler and suggest sample sizes of at least 50.)

(4) It should be mentioned that even in the case of smaller sample sizes the t-test is quite robust with regard to violations of the assumption of normality if the population variances and the sample sizes are equal. Therefore, in the case of small samples, we should try to provide for approximately equal sample sizes.

(5) In testing the significance between two proportions, the reader should be told that $H_0 : p_1 = p_2$, which implies $\sigma_1 = \sigma_2$, and that therefore the z-test and not the t-test is appropriate.

(6) A measure for the practical significance of the outcome of a t-test should be given (cf. e.g. Hays 1973: 417 for such a measure).

In Chapter 8 entitled "Some useful non-parametric tests" the Mann-Whitney U-test, the Wilcoxon signed-ranks test and the sign test are presented. I think this chapter will prove very valuable because (applied) linguists will often have to deal with data not meeting the assumptions involved in the use of parametric tests. One short comment with regard to the U-test: Butler's explanation of the rationale underlying this test holds only for equal sample sizes. If $N_1 \neq N_2$, not the *sum of ranks* for each sample will be similar under the null hypothesis and different under the alternative hypothesis as we are told on page 98f. but the *mean rank*, i.e. the sum of ranks divided by the sample size.

Chapter 9 is devoted to the well-known χ^2-test. As a first application of this test Butler shows how it can be used to check whether a set of observed data comes from a normally distributed population (χ^2-test for goodness of fit). In my opinion the method presented is problematical in several respects (cf. Büning and Trenkler 1978: 84ff.).

(1) The test is only suitable where we have a fairly large sample. However, if the sample is large, the assumption of normality does not play an important role in the parametric tests discussed by Butler.

(2) In testing the goodness of fit, in contradistinction to ordinary significance testing, we are interested that the null hypothesis will *not* be rejected. Therefore, if the test fails to reject the

null hypothesis, one cannot conclude that the data actually follow the theoretical model. Thus, if we erroneously accept the null hypothesis, we commit a type II error with an unspecified probability (cf. Hartung, Elpelt and Klösener 1984: 139).

(3) The result of the χ^2-goodness of fit test depends on the grouping of the data, which is arbitrary to a certain degree.

(4) If there are only a few classes or if the expected frequencies are small, the approximation to the χ^2-distribution may be poor (cf. e.g. Cochran 1952, Tate and Hyer 1973). (On page 201, Butler applies the χ^2-test to distributions with 4 and 5 classes respectively.)

Because of this and other possible criticisms, I think the problem of testing the fit of an observed set of data to a normal distribution should either not be treated in an introductory text or a different method should be used, namely a modified form of the Kolmogorov-Smirnov-test, which has several advantages over the χ^2-test (cf. Lilliefors 1967, Conover 1971).

With regard to the χ^2-test for independence and association, I would like to make the following points:

(1) In Exercise 1 on page 198 the χ^2-test is applied to a combined sample of $N = 15,437$. With such a large sample, we obtain a highly significant result almost by necessity even if there is only a very small difference between observed and expected frequencies. This fact, which also applies to other data presented in this chapter, should at least have been mentioned (cf. my remarks on the problem of practical significance).

(2) A reference should have been made to the φ-coefficient discussed in Chapter 11, which can be used to measure the practical significance of an obtained χ^2-value.

(3) In Exercise 3 on page 199 a "family" of seven χ^2-tests is applied to one and the same sample. Here it should have been pointed out that carrying out one and the same test several times on one and the same sample leads to an increase in the probability of making a type I error (cf. Miller 1981) and that there are methods more adequate than the one presented.

(4) It should have been mentioned that there are modern alternatives to the χ^2-test which are often more appropriate (for a good,

although not very elementary presentation see Bishop, Fienberg and Holland 1975).

Chapter 10 entitled "The F distribution and its uses" first deals with the F-test for the homogeneity of two variances. Subsequently, a very short introduction to the analysis of variance (ANOVA) technique is given. This chapter contains several flaws and is in my opinion clearly inferior to the corresponding chapter in Hatch and Farhady (1982).

(1) In the section on the F-test for the homogeneity of two variances, it should be mentioned that in contradistinction to the t-test for the difference of two means the F-test is very sensitive to violations of the assumption of normality. Thus, if we are not sure if the assumption of normality is met, it may be better not to carry out an F-test before doing a t-test (in particular, if the sample sizes are approximately equal).

(2) On page 132 Butler tells us that in the analysis of variance the assumption of homogeneity of variances can be tested by the F-test for two variances. However, in the case of more than two variances the use of the F-test for two variances leads to an increase in the type I error. Therefore, other methods such as the Bartlett test should be used (cf. Winer 1971: 205-210). However, in case the assumption of normality is not met, it may be better not to test the assumption of homogeneity of variance at all (especially if the sample sizes are approximately equal), since most tests for the homogeneity of (several) variances are very sensitive to departures from normality (cf. above my remarks on the t-test).

(3) Butler tells us to use the z-test or the t-test previously discussed to find out which means differ significantly in case ANOVA has yielded a significant result (cf. e.g. p. 129). However, multiple z-tests or multiple t-tests lead to an increase in the type I error (as is the case with multiple F-tests for the homogeneity of variances). Therefore, procedures such as the Scheffé test should be used, which is discussed in almost all statistics books dealing with ANOVA. (There are not a few publications in (applied) linguistics where this fact is not taken into account either.)

(4) In the first formula on page 131 the squared deviations of each individual group mean from the grand mean must be weighted

by the corresponding group sizes. Hence the formula should read: $SS_b = \Sigma N_{group}(\bar{x}_{group} - \bar{x}_{grand})^2$.

Chapter 11 entitled "Correlation" covers the following topics: the nature of correlation; visual representation of correlation by means of scattergrams; the Pearson product-moment corrrelation coefficient; the Spearman rank correlation coefficient; the φ-coefficient. I think this chapter is good as a short first introduction to the important topic of correlation. A more extensive discussion of how to interpret correlations and of factors affecting correlation coefficients can be found for example in Hatch and Farhady (1982) or Diehl and Kohr (1982). Some short comments:

(1) The use of the Pearson correlation coefficient does not assume, as Butler tells us, that each set of scores is normally distributed (p. 143). Normality has to be assumed only if we test an obtained coefficient for significance.

(2) The important concept of covariance should have been introduced. With the help of this concept the rationale underlying the Pearson coefficient could have been very easily explained.

(3) The reader should have been told how to interpret the Pearson coefficient in terms of the amount of variance explained. In this context it could have been shown that a highly significant correlation coefficient may be without any practical significance measured by the amount of variance explained.

(4) Since Butler as a rule uses Greek letters for population parameters (a convention followed by most modern statistics books), the use of ρ to designate the Spearman coefficient of a sample (cf. p. 146) might be confusing. (In most modern statistics books ρ designates the Pearson coefficient of a population; the Spearman coefficient for a sample is then often designated by r_s.)

(5) It does not become clear that the Spearman coefficient can be used whenever the relationship between two variables is monotonic increasing or monotonic decreasing. If the relationship is in addition linear, the Pearson coefficient is more appropriate.

(6) On page 149 the negative sign of the φ-coefficient obtained should have been commented upon. (If we interchange the columns "Failure" and "Success" in Table 11.5 without making

a corresponding modification in the formula for φ, the sign of φ becomes positive.)

(7) The reader is not told how to calculate the correlation in a contingency table larger than four by four. Here it could have been shown that φ is merely a special case of the coefficient V introduced by Cramér, which can be used with contingency tables of *any* size (for a definition of V see e.g. Blalock 1979: 305). With the help of V it could for example have shown that in Exercise 1 on page 198 the correlation is almost zero ($V = .07$) in spite of the highly significant χ^2 – value obtained.

(8) There is no indication as to how to interpret the numerical value of a φ-coefficient obtained. Here it should have been shown that the value of φ strongly depends on the marginal frequencies in a given contingency table. Thus a numerical value of, say, .3 may indicate a rather weak correlation in one contigency table and a rather strong correlation in another one. (To make φ-coefficients from different tables comparable, an obtained φ-coefficient may be divided by its theoretical maximum (see Cole 1949 and the discussion in Clauß and Ebner 1979: 283ff.)).

(9) With regard to Exercise 4 on page 153 the reader should have been told that carrying out three significance tests on the same set of data inflates the error of the first kind and that therefore a non-parametric ANOVA would have been more appropriate.

In the final chapter entitled "Statistics, computers and calculators" the Statistical Package for the Social Sciences (SPSS) and the Minitab statistical package are very briefly discussed. (The reader interested in a more extensive treatment of the topic is referred by Butler to his book *Computers in Linguistics* (1985)). I think this chapter is good as a very first introduction for readers not at all familiar with this topic. Especially important is Butler's final caveat on page 167 from which I would like to cite at least the following passage:

> "There is, of course, every reason why we should take advantage of computers or pocket calculators to reduce the tedium and the likelihood of error involved in calculting statistics from first principles. However, it cannot be too strongly emphasized that it is essential to understand the principles on which the calculation and the use of statistics is based."

In view of the fast-spreading use (and abuse) of statistical packages such as SPSS in (applied) linguistics I feel that this caveat cannot too often be repeated.

I would now like to give a more global appreciation of Butler's book. Let me start with some criticism: (a) As has been shown, there are some real flaws in the presentation of statistics (e.g. in Chapter 10), which can however easily be corrected. (b) Although Butler primarily deals with linguistic data, he completely fails to take into account the relevant literature in quantitative linguistics (cf. my comments on Chapter 1). (c) The bibliography is extremely limited (only 10 items). Since Butler's treatment of the topic is necessarily very selective and simplifying, precise references to the relevant literature would be of considerable help to readers wanting to deepen their understanding of the topic.

In my opinion, the most positive feature of the book is that the author continuously attempts to provide the reader with an understanding of the formulas and the statistical methods presented. (Nevertheless, there are some cases (e.g. the formula for the φ-coefficient or the correction for continuity on pages 108 and 122) where such an attempt has not been made although it would have been easily possible.) Furthermore, the sometimes quite difficult content is presented in a very clear and didactically sound manner.

A second positive feature is Butler's concern for pointing out the important relationship between the measurement level of the data and the statistical method to be chosen – an aspect not sufficiently taken into account for example by Hatch and Farhady (1982).

A third positive feature is that Butler presents in addition to the χ^2-test, some other important non-parametric methods. Since in (applied) linguistics the assumptions on which parametric methods are based are often not met, non-parametric methods should be treated at some length in an introductory textbook (and not neglected almost completely, as Hatch and Farhady 1982 do). Moreover, even if the assumptions for the application of a parametric procedure are met, non-parametric methods are often almost as powerful as the corresponding parametric methods. Finally, non-parametric methods are as a rule mathematically much simpler than parametric procedures. (For example, the formula for the U-test on page 99 could have been

quite easily derived with the help of fairly elementary algebra.) This is another reason why non-parametric methods should be emphasized in an introductory text aiming at an understanding of the formulas and methods presented.

In sum, I think that Butler's book, in spite of some flaws in the presentation of the statistical methods and in their application to linguistic data, can be recommended as a first introduction to the application of statistics in (applied) linguistics. This holds in particular for readers interested not in a cookery-book-like presentation but in an understanding of the methods presented.

References

Bishop, Y., Fienberg, S., and Holland, P. 1975. *Discrete multivariate analysis: Theory and practice.* Cambridge, Mass.: MIT.

Blalock, H.M.J. 1979. *Social Statistics.* Tokyo: McGraw-Hill Kogakusha. (second revised edition).

Büning, H., and Trenkler, G. 1978. *Nichtparametrische statistische Methoden.* Berlin & New York: de Gruyter.

Butler, C. 1985. *Computers in linguistics.* Oxford: Basil Blackwell.

Clauß, G., Ebner, H. 1979. *Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen.* Thun and Frankfurt, M.: Harri Deutsch.

Cochran, W.G. 1952. "The χ^2 test of goodness of fit." *Annals of Mathematical Statistics* 23: 315-345.

Cole, L. 1949. "The measurement of interspecific association." *Ecology* 30: 411-424.

Conover, W. 1971. *Practical nonparametric statistics.* New York: Wiley.

Diehl, J.M., and Kohr, H.-U. 1982. *Deskriptive Statistik.* 4th. ed. Frankfurt: Fachbuchhandlung für Psychologie.

Hartung, J., Elpelt, B., and Klösener, K.-H. 1984. *Statistik. Lehr- und Handbuch der angewandten Statistik.* München: Oldenbourg.

Hatch, E., Farhady, H. 1982. *Research design and statistics for applied linguistics.* Rowley, Mass.: Newbury House.

Hays, W.L. 1973. *Statistics for the social sciences.* London and New York: Holt, Rinehart and Winston.

Lilliefors, H. 1967. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *Journal of the American Statistical Association* 62: 399-402.

Miller, R. 1981. *Simultaneous statistical inference.* 2nd ed. New York: Springer.

Orlov, Ju.K. 1982. "Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik)." Orlov, Ju. K., Boroda, M.G. and Nadarejšvili, I.Š. (eds). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer: 1-55.

Tate, M., and Hyer, L. 1973. "Inaccuracy of the X^2-test of goodness of fit, when expected frequencies are small." *Journal of the American Statistical Association* 68: 836-841.

Winer, B. 1971. *Statistical principles in experimental design.* 2nd. ed. New York: McGraw-Hill.

CURRENT BIBLIOGRAPHY

ABBREVIATIONS

Festschrift Muller: Muller, Ch.
Méthodes quantitatives et informatiques dans l'étude des textes.
Colloque International CNRS Université de Nice, 5-8 juin 1985. Genève - Paris, Slatkine-Champion, 1986, 948 pp.

Tuldava 1987: Tuldava, Ju.A. (Ed.)
Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov [Quantitative linguistics and automatic text analysis].
Tartu, Acta et Commentationis Universitatis Tartuensis Nr. 774, 1987.

GENERAL

1. BERTRAND, C.: Enseignement des méthodes quantitatives et computationnelles en sciences humaines à l'Université de Liège. Festschrift Muller, 89-98.

2. BONNAUD-LAMOTTE, D.: L'ordinateur a-t-il servi nos études littéraires contemporaines? Bilan et perspective de l'URL 5. Festschrift Muller, 107-116.

3. BRAINERD, B.: Three issues related to the theme of this conference. Festschrift Muller, 125-132.

4. DENOOZ, J.: Le centre informatique de philosophie et lettres de l'Université de Liège. Festschrift Muller, 295-302.

5. FORTIER, P.A.: Les données du trésor de la langue française: une occasion pour les chercheurs. Festschrift Muller, 399-408.

6. GENTILHOMME, Y.: Théorie des microsystèmes. Mise en oeuvre dans les sciences du langage. Festschrift Muller, 433-444.

7. JOLIVET, R.: Aspect statistique de la structuration linguistique. Festschrift Muller, 507-518.

8. KRYLOV, Ju.K., JAKUBOVSKAJA, M.D.: On some possibilities of applying the quantum-mechanical formalism when constructing stochastic models of text generation and recognition. Symposium in Grammars of Analysis and Synthesis and Their Representation in Computational Structures. Tallinn 1983, 49-51.

9. MASSONIE, J.Ph.: Q-occurrences libres. Festschrift Muller, 609-624.

10. MULLER, Ch.: Méthodes quantitatives et informatiques dans l'étude des textes. Colloque International CNRS Université de Nice, 5-8 juin 1985. Genève - Paris, Slatkine-Champion, 1986, 948 pp.

11. NEŠITOJ, V.V.: Forma predstavlenija rangovych raspredelenij [About the form of representing rank distributions]. Tuldava 1987, 123-134.

12. NEŠITOJ, V.V.: Issledovanie rangovych raspredelenij [An investigation of rank distributions]. Naučno-techničeskaja informacija, Ser. 2, 1985/2, 16-20.

13. ROPER, J.P.G.: The use of microcomputer in literary and linguistic research. Festschrift Muller, 727-736.

14. TOURNIER, M.: Bilan critique. Festschrift Muller, 885-889.

15. TULDAVA, Ju.A.: K voprosu o klassifikacii i interpretacii lingvističeskich raspredelenij [On classification and interpretation of linguistic distributions]. Tuldava 1987, 155-168.

16. TULDAVA, Ju.A. (Ed.): Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov [Quantitative linguistics and automatic text analysis]. Tartu, Acta et Commentationis Universitatis Tartuensis Nr. 774, 1987.

17. WOODS, A., FLETCHER, P., HUGHES, A.: Statistics in Language Studies. Cambridge, Cambridge University Press 1986, 322 pp.

PHONOLOGY

18. DELBECQUE, N., DEBROCK, M.: Le e muet: vérification sur un corpus de français parlé. Festschrift Muller, 255-272.

19. DUGAST, D.: Les affinités phonémiques électives dans les textes en français. Festschrift Muller, 341-352.

20. GRŽEBIČEK, L. [= Hřebíček, L.]: Fonologičeskaja struktura tureckogo slova [The phonological structure of the Turkish words]. Novoe v zarubežnoj lingvistike. Vypusk XIX. Problemy sovremennoj tjurkologii. Sostavlenie A.N. Barulina, Moskva, Progress, 1987, 48-58.

21. LEHFELDT, W.: Quantitative methods in phonology. Phonologica 1984, London 1987, 157-163.

WRITING

22. CATACH, N.: Réflexions en vue d'une modernisation automatique des graphies anciennes. Festschrift Muller, 167-174.

MORPHOLOGY

23. CALZOLARI, N., CECCOTTI, M.L., ROVENTINI, A.: Some generalisations on classes of derivatives. Festschrift Muller, 159-166.

24. GERD, A.S.: Ėtalonnye tipy morfologičeskich paradigm drevneslavjanskich tekstov [Standard types of morphological paradigms in Old Slavonic texts]. Tuldava 1987, 55-72.

25. NOE, A.: Un indice pour la localisation des formes: le coefficient de variation. Festschrift Muller, 681-688.

Syntax

26. BREUILLARD, J.: Pour une macrostylistique de la phrase. Festschrift Muller, 133-142.

27. DIMON, P., LEJOSNE, J.C.: Analyse syntaxique en traitement automatique des langues naturelles. Exemples sur l'allemand et le néerlandais. Festschrift Muller, 315-328.

28. HUG, M.: Le rôle des données quantitatives dans l'étude syntaxique d'une langue. Festschrift Muller, 481-494.

29. OUAMARA, A.: Cooccurrence et syntaxe: réseau lexico-syntaxique. Festschrift Muller, 689-698.

Semantics

30. ARAPOV, M.V.: Upotrebitel'nost' i mnogoznačnost' slova [Usualness and polysemy of words]. Tuldava 1987, 15-28.

31. CIBOIS, Ph.: Les préalables nécessaires à l'analyse automatique de contenu. Festschrift Muller, 195-206.

32. POLIKARPOV, A.A.: Polisemija: sistemno-kvantitativnye aspekty [Quantitative aspects of polysemy]. Tuldava 1987, 135-154.

Lexicology

33. BEAUCHEMIN, N.: Homographie et solutions pratiques en lemmatisation minimale. Festschrift Muller, 25-36.

34. BRANCA, P.: La lemmatisation de l'arabe. Festschrift Muller, 117-124.

35. CONDE, C., JACOBI, D.: L'indexation lexicale des contextes des termes pivots dans un corpus intertextuel de vulgarisation scientifique. Festschrift Muller, 207-220.

36. DELCOURT, Ch.: De la statistique lexicale à la statistique segmentale. Festschrift Muller, 273-284.

37. DEVONS, N.: Concerning corpus-derived lexical frequency ratings. Festschrift Muller, 303-314.

38. LENOBLE, M.: Statistique lexicale et critique littéraire: le mariage impossible? Festschrift Muller, 565-574.

39. LYNE, A.A.: In praise of Juilland's D: a contribution to the empirical evaluation of various measures of dispersion applied to word frequencies. Festschrift Muller, 587-598.

40. MANASJAN, N.S.: O mere blizosti teoretičeskogo i empiričeskogo raspredelenij (na materiale raspredelenij dlin leksičeskich edinic v tekste) [A measure of proximing between theoretical and empirical distributions (applied to the distribution of lengths of lexical units in texts)]. Tuldava 1987, 103-114.

41. NAJOCK, D.: Bootstrap experiments for the evaluation of expected values and variances of vocabulary sizes. Festschrift Muller, 657-670.

42. NEŠITOJ, V.V.: Kriterii odnorodnosti leksičeskogo sostava častotnogo slovarja [The criterion of lexical homogenity of a frequency dictionary]. Vesnik Beloruskago dzjaržajuago un-ta Ser. 4, 1984/1, 54-56.

43. NISSAN, E.: On the architecture of onomaturge. An expert system inventing neologisms. Festschrift Muller, 671-680.

44. SALONI, Z.: Une expérience d'utilisation de l'ordinateur dans la lexicographie: l'index schématique a tergo des mots-formes de la langue polonaise contemporaine. Festschrift Muller, 771-778.

45. SHOHAM, K.: SQL: Un modèle pour une base de données lexicologiques. Festschrift Muller, 797-806.

46. STINDLOVA, J.: Statistical data from the dictionary of Czech literary language grammatical data. Festschrift Muller, 807-820.

47. THOIRON, Ph.: Indice de diversité et mesure de la richesse

lexicale. Festschrift Muller, 831-840.

48. TOURNIER, M.: Dans l'ombre portée des sigle confédéraux: un mirage lexicométrique (CGT et CFDT en 1972) Festschrift Muller, 841-854.

49. TULDAVA, Ju.A.: Problemy i metody kvantitativno-sistemnogo issledovanija leksiki (na materiale èstonskogo jazyka) [Problems and methods of quantitative systemic investigation of lexicon (applied to Estonian)]. Tartu, Tartuskij gosudarstvennyj universitet 1984.

Text Analysis

50. ABI FARAH, A.: La reconnaissance automatique de l'auteur inconnu d'un texte arabe. Festschrift Muller, 13-18.

51. ALEKSEEV, P.M.: Lingvističeskie raspredelenija (Elementy količestvennogo analiza teksta) [Linguistic distributions (Elements of quantitative text analysis)]. Leningrad, Leningradskij gosudarstvennyj pedagogičeskij institut 1985.

52. ALEKSEEV, P.M.: O rangovych raspredelenijach v kvantitativnoj tipologii teksta [On rank frequency distributions in quantitative text typology]. Tuldava 1987, 3-14.

53. BEAUDE, J.: Le lexique de l'imagination dans l'œuvre de Malebranche. Festschrift Muller, 37-42.

54. BENSON, J.D., BRAINERD, G., GRAEVES, W.S.: A quantifical approach of field of discourse. Festschrift Muller, 55-64.

55. BENZECRI, J.P.: L'analyse multidimensionnelle des textes. Festschrift Muller, 65-78.

56. BERNI CANANI, U.: Information retrieval et analyse de textes. Festschrift Muller, 79-88.

57. BOIVIN, L., BRATLEY, P.: Un système de dépistage de textes entiers. Festschrift Muller, 99-106.

58. BORODA, M.G., PEŠKOVSKIJ, V.E.: Ritmika associativnogo potoka: k probleme količestvennogo analiza [Rhythm of the associative stream: the problem of its quantitative analysis]. Tuldava 1987, 49-54.

59. CERCONE, N., MURCHISON, C.: Suggestions for the design of an expert system for literary research. Festschrift Muller, 175-184.

60. CORNS, Th.N.: Literary theory and computer-based criticism. Current problems and future prospects. Festschrift Muller, 221-228.

61. CUMMINGS, M.: Analysis of old English text through logic programming. Festschrift Muller, 229-240.

62. DAUTREY, Ph., MULLER, P.: Etude des structures temporelles dans le discours politique. Festschrift Muller, 241-254.

63. DUCHASTEL, J., PLANTE, P.: Le Déredec et l'analyse de textes par ordinateur. Festschrift Muller, 329-340.

64. EVRARD, E.: Quelques variations quantitatives dans l'hexamètre latin. Festschrift Muller, 361-372.

65. FARINA, L.F.: XT-LEXED: a quanti-qualitative method for text analysis and the IBM PC models. Festschrift Muller, 373-380.

66. FIALA, P.: Inventaires distributionnels et opérateurs textuels dans le Rivage des Syrtes de Julien Gracq. Structures syntaxiques et faits stylistiques. Festschrift Muller, 381-390.

67. FLIKEID, K.: Le français acadien: analyse comparative de textes de langue parlée. Festschrift Muller, 391-398.

68. FRAUTSCHI, R.L.: Diégèse et métalepse dans quatre textes de Diderot. Quelques évidences quantitatives. Festschrift Muller, 409-420.

69. GABR, R.: La désignation de l'homme et de la femme dans le Coran: étude statistique. Festschrift Muller, 421-432.

70. GLUŠAK, T.S., BOL'ŠAKOV, I.I.: Stiledifferencirujuščie po-

tencii otglagol'nych suščestvitel'nych bezaffiksnogo tipe v nemeckom jazyke [On style discriminating power of suffixless deverbative nouns in German]. Tuldava 1987, 73-80.

71. IDE, N.M.: Patterns of imagery in William Blake's the four Zoas. Festschrift Muller, 495-506.

72. KRYLOV, Ju.K.: Stacionarnaja model' porožđenija svjaznogo teksta [A stationary model of coherent text generation]. Tuldava 1987, 81-102.

73. LAFON, P.: Pour une nouvelle unité de segmentation des textes à partir de l'inventaire des segments répétés. Application à la résolution générale du congrès de la CFDT en 1976. Festschrift Muller, 531-540.

74. LAURETTE, P.: Visualisation, structure et système. Information graphique et textes littéraires. Festschrift Muller, 541-554.

75. LAURIAN, A.M.: Stylistique et automatisme: à nouvelle approche nouvelles définitions. Festschrift Muller, 555-564.

76. LEBEDEV, A.N.: Zakonomernosti povtorenija slov v reči [Laws of repetition of words in speech]. Psichologičeskij žurnal 4, 1983, 11-22.

77. LUONG, N.X., NOVI, M.: Représentations arborées de données textuelles. Festschrift Muller, 575-586.

78. MILIC, L.T.: The apriori question in stylistics. Festschrift Muller, 637-644.

79. MULLER, Ch.: Méthodes quantitatives et informatiques dans l'étude des textes. Colloque International CNRS Université de Nice, 5-8 juin 1985. Genève - Paris, Slatkine-Champion, 1986, 948 pp.

80. PERRET, U.: Linguistique judiciaire souteneue par l'ordinateur. Festschrift Muller, 699-704.

81. PICCHI, E., CALZOLARI, N.: Textual perspective through an

automatized lexicon. Festschrift Muller, 705-716.

82. POSWICK, Fr.R.F.: Informatique et bible - 1985. Festschrift Muller, 717-726.

83. ROTHKEGEL, A.: Knowledge representation and text analysis. Festschrift Muller, 737-746.

84. RUTTEN, Ch.: Aristote, métaphysique, lambda 8. Essai de stylométrie. Festschrift Muller, 747-758.

85. SALEM, A.: L'utilisation des inventaires des segments répétés dans les études typologiques sur des corpus de textes. Festschrift Muller, 759-770.

86. SCHMIDT, K.M.: Concept versus meaning: the contribution of computer-assisted content analysis and conceptual text analysis to this disputed area. Festschrift Muller, 779-798.

87. TADDEI-ELMI, G., ALBA, L., CAMMELLI, A., ROMANO, G.A.: Thesaurus et relations paradigmatiques de sens. Une expérience de thesaurus juridique. Festschrift Muller, 821-830.

88. TUOMINEN, S.L.: L'analyse discriminative en stylistique. Festschrift Muller, 855-862.

89. WEITZMAN, M.: Les caractéristiques numériques indépendantes de la longueur du texte. Festschrift Muller, 863-878.

SOCIOLINGUISTICS

90. HŘEBÍČEK, L.: Quantities of Social Communication With General Applications to Islam and Social Morphogenesis. Prague 1986, 197 pp.

91. MARTEL, P.: Richesse lexicale et variables sociologiques. Festschrift Muller, 599-608.

Documentation

92. AIT HAMLAT, A.: Applications des méthodes statistique à l'indexation automatique de documents. Festschrift Muller, 19-24.

93. BEHAR, H.: Un projet de banque de données d'histoire des faits littéraires. Festschrift Muller, 43-54.

94. BURNARD, L.: From archive to database. Festschrift Muller, 143-158.

95. DENDIEN, J.: Un système de gestion de base de données textuelles. Festschrift Muller, 285-294.

96. GOFFIN, R.: Eurodicautom: la banque de données terminologiques de la commission des communautés européennes: 1975-1985. Festschrift Muller, 445-454.

97. GOLDFIELD, J.: An approach to literary computing in French on the UNIX system. Festschrift Muller, 455-466.

98. HARRIS, M.D.: Improving efficiency in storage & access of natural language text. Festschrift Muller, 467-480.

99. HOCKEY, S.: Remarks on software. Festschrift Muller, 881-884.

100. KAMMER, M.: On problems of literary data banks. Festschrift Muller, 519-530.

101. MATTHEWS, E., RAHTZ, S.: Designing and using a database of Greek personal names. Festschrift Muller, 625-636.

102. NAIS, H., DERNIAME, O., HENIN, M., GRAFF, J., MONSONEGO, S.: Constitution d'une base de données textuelles et lexicographiques du Moyen français. Festschrift Muller, 645-656.

Language Teaching

103. DUNN-LARDEAU, B.: Maîtrise du français écrit et applications pédagogiques de l'ordinateur. Festschrift Muller, 353-360.

Music

104. CHARNASSE, H.: Emploi de l'informatique pour la transcription et l'édition de textes musicaux. Festschrift Muller, 185-194.

Quantitative Linguistics

Aim and Scope: Application of Mathematical Methods in Research on Linguistics, Literature and Related Areas.

Modes of Publication: Irregular intervalls, circa 4 volumes per year.

Available volumes:

Vol. 1: Altmann, G. (ed.), *Glottometrika* 1. 1978; VII+231 pp., DM 24,80

Vol. 2: Grotjahn, R., *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* 1979; V+296 pp., DM 34,80

Vol. 3: Grotjahn,R. (ed.), *Glottometrika* 2. 1980; V+218 pp., DM 24,80

Vol. 4: Strauss, U., *Struktur und Leistung der Vokalsysteme.* 1980; III+158 pp., DM 19,80

Vol. 5: Matthäus, W. (ed.), *Glottometrika* 3. 1980; 236 pp., DM 29,80

Vol. 6: Grotjahn, R., Hopkins, E. (eds.), *Empirical research on language teaching and language acquisition.* 1980; 231 pp., DM 29,80

Vol. 7: Altmann, G., Lehfeldt, W., *Einführung in die quantitative Phonologie.* 1980; X+401 pp., DM 29,80

Vol. 8: Altmann, G., *Statistik für Linguisten.* 1980; III+239 pp., DM 29,80

Vol. 9: Hopkins, E., Grotjahn, R. (eds.), *Studies in language teaching and language acquisition.* 1981; 220 pp., DM 29,80

Vol. 10: Skorochod'ko, E.F., *Semantische Relationen in der Lexik und in Texten.* 1981; 218 pp., DM 29,80

Vol. 11: Grotjahn, R. (ed.), *Hexamater studies.* 1981; VI+263 pp., 29,80

Vol. 12: Rieger, B. (ed.), *Empirical semantics I. A collection of new approaches in the field.* 1981; XIII+375 pp., DM 49.80

Vol. 13: Rieger, B. (ed.), *Empirical Semantics II. A collection of new approaches in the field.* 1981; XII+442 pp., DM 49.80

Vol. 14: Lehfeldt, W., Strauss, U. (eds.), *Glottometrika* 4. 1982; 200 pp., DM 29,80

Vol. 15: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen.* 1982; 330 pp., DM 49,80

Vol. 16: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf's law.* 1982; 262 pp., DM 39,80

Vol. 17: Arapov, M.V., Cherc, M.M., *Mathematische Methoden in der historischen Linguistik.* 1983; 171 pp., DM 24,80

Vol. 18: Brainerd, B. (ed.), *Historical linguistics.* 1983; II+236 pp., DM 29,80

Vol. 19: Winkler, P. (ed.), *Investigations on the speech process.* 1983; III+312 pp., DM 34,80

Vol. 20: Köhler, R., Boy, J. (eds.), *Glottometrika* 5. 1983; 228 pp., DM 29,80

Vol. 21: Goebl, H. (ed.), *Dialectology.* 1984; IV+335 pp., DM 49,80

Vol. 22: Alekseev, P.M., *Statistische Lexikographie.* 1984; 157 pp., DM 24,80

Vol. 23: Schwibbe, G., *Intelligenz und Sprache: Zur Vorhersagbarkeit des intellektuellen Niveaus kontentanalytischer Indikatoren.* 1984; 200 pp., DM 24,80

Vol. 24: Piotrowski, R.G., *Text - Computer - Mensch.* 1984; IX+422 pp., DM 49,80

Vol. 25: Boy, J., Köhler, R. (eds.), *Glottometrika* 6. 1984; 195 pp., DM 24,80

Vol. 26: Rothe, U. (ed.), *Glottometrika* 7. 1984; 176 pp., DM 29,80

Vol. 27: Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A., *Mathematische Linguistik.* 1985; 514 pp., DM 64,80

Vol. 28: Piotrowski, R.G., Popeskul, A.N., Chazinskaja, M.S., Rachubo, N.P., *Automatische Wortschatzanalyse.* 1985; 187 pp., DM 29,80

Vol. 29: Andersen, S., *Sprachliche Verständlichkeit und Wahrscheinlichkeit.* 1985; 194 pp., DM 29,80

Vol. 30: Embleton, Sh.M., *Statistics in historical linguistics.* 1986; VIII+194 pp., DM 29,80

Vol. 31: Köhler, R., *Zur sprachlichen Synergetik: Struktur und Dynamik der Lexik.* 1986; 201 pp., DM 29,80

Vol. 32: Fickermann, I. (ed.), *Glottometrika* 8. 1987; 211 pp., DM 29,80

Vol. 33: Wildgen, W., Mottron, L., *Dynamische Sprachtheorie. Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese.* 1987; III+423 pp., DM 59,80

Vol. 34: Grotjahn, R., Klein-Braley, C., Stevenson, D.K. (eds.), *Taking their measure: The validity and validation of language tests.* 1987; X+274pp., DM 39.80

Vol. 35: Schulz, K.P. (ed.), *Glottometrika* 9. 1988; 271 pp.

In preparation:

Piotrowski, R., Lesochin, M., Luk'janenkov, K., *introduction of elements of mathematics to linguistics.*

Tuldava, Ju., *Probleme und Methoden der quantitativen Analyse der Lexik.*

Hammerl, R. (ed.), *Glottometrika* 10.

Mizutani, Sh. (ed.), *Japanese contributions to quantitative linguistics.*

Altmann, G., *Wiederholungen in Texten.*

Altmann, G., Hammerl, R., *Diskrete Wahrscheinlichkeitsverteilungen* I,II.