QUANTITATIVE LINGUISTICS

Vol. 38

GLOTTOMETRIKA 10

edited

by

R. Hammerl

Studienverlag Dr. N. Brockmeyer
Bochum 1989

QUANTITATIVE LINGUISTICS

CIP-Titelaufnahme der Deutschen Bibliothek

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

10 (1989)
(Quantitative linguistics ; Vol. 38)
ISBN 3-88339-700-8
NE: GT

ISBN 3-88339-700-8

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

CONTENTS

Annotations

Hammerl, R. (ed.), Glottometrika 10, 1988

Ein Beitrag zu Köhler's Modell der sprachlichen Selbstregulation

R. Hammerl, Kielce
J. Maj, Kielce

1. Einleitung

Die zunehmende Anwendung mathematischer Methoden in der quantitativen Linguistik hat in den letzten Jahren dazu geführt, daß viele Sprachgesetzmäßigkeiten (Gesetze) gefunden und in Form mathematischer Modelle dargestellt werden konnten. Es ist offensichtlich klar, welchen Vorteil die mathematische Beschreibung im Gegensatz zur verbalen Beschreibung der untersuchten Sachverhalte besitzt.

Andererseits macht es aber die Vielfalt der gefundenen Sprachgesetze, die ja sehr verschiedene sprachliche Probleme betreffen, nicht ohne weiteres möglich, ein allgemeines Sprachmodell zu bilden, aus dem sich alle schon gefundenen Sprachgesetze und deren Wechselwirkungen objektiv ableiten lassen. Das heißt jedoch nicht, daß die Erstellung eines solchen Sprachmodells unmöglich wäre. Erste wichtige Vorarbeiten in dieser Richtung wurden von Köhler (1986) geleistet.

Die natürlichen Sprachen stellen im Vergleich zu komplizierten technischen und technologischen Objekten Systeme dar, die nur in einem beschränkten Maße einer deterministischen Modellbildung zugänglich sind, da ja hier keine a priori physikalischen, chemischen usw. Abhängigkeiten zwischen den Elementen dieser Systeme vorliegen. Außerdem sind diese Systeme nur über konkrete Sprachprodukte (Texte) zugänglich, deren Zahl unendlich und deren Struktur stark von individuellen menschlichen Faktoren belastet ist, so daß die Struktur dieser Texte nur stochastisch beschrieben werden kann.

Aus dieser Sicht knüpfen wir an die wichtigen Vorarbeiten Köhler's (1986) an, die der Erstellung eines Basismodells der Lexik gewidmet sind. Dabei werden nur ausgewählte Probleme dieser Arbeit von Köhler diskutiert, die uns am wichtigsten erschienen und für die Weiterentwicklung dieses Modells von erstrangiger Bedeutung sind, wobei unsere Kritik stets

konstruktiv sein wird. Wir gehen dabei davon aus, daß die Erstellung eines Sprachmodells, welches die Struktur und Dynamik der Lexik beschreiben soll, nur unter Anwendung von Methoden möglich ist, die in der Systemtheorie und Kybernetik entwickelt wurden. Aus diesem Grunde werden wir zunächst die wichtigsten Grundlagen der systemtheoretischen Modellierung vorstellen.

2. Zur systemtheoretischen Modellierung

Das Aufstellen eines mathematischen Modells mit der Struktur- und Parameterangabe wird in der Systemtheorie als Identifikation bezeichnet. Grundsätzlich unterscheidet man eine theoretische und eine experimentelle Systemidentifikation. Die Spezifik des zu untersuchenden Sachverhaltes zwingt den Forscher oft zur Anwendung der experimentellen Systemidentifikation, obwohl die theoretische Identifikation viele Vorteile hat.

In der Systemtheorie untersuchte Objekte werden als Systeme (black box) betrachtet, wobei allgemein die Eingangsgrößen u_i, die Ausgangsgrößen x_i und das System S (mit einer bestimmten Struktur und bestimmten Parametern) unterschieden werden. Dies verdeutlicht Abbildung 1.

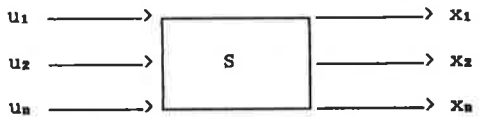


Abb. 1. Grundelemente systemtheoretischer Modellierung

2.1. Theoretische Systemidentifikation

Hier werden aus theoretischen Überlegungen über die Spezifik des untersuchten Systems Modelle erstellt und später überprüft. Dabei sind zunächst mathematische Abhängigkeiten zwischen den wesentlichsten existierenden Größen (Eingangsgrößen, Ausgangsgrößen, Zustandsgrößen = Hilfsgrößen bei der Aufstellung der Zusammenhänge zwischen Eingangs- und Ausgangsgrößen) abzuleiten. Das Ergebnis muß ein Komplex mathematischer Abhängigkeiten sein, welches über die gegenseitigen Zusammenhänge aller interessierenden Größen Auskunft gibt. Die geordnete Form dieser Abhängigkeiten, die für die Analyse und Synthese des Systems geeignet ist, wird gewöhnlich in Form einer Übertragungsfunktion (= dem Modell) G(p) dargestellt. G(p) beschreibt die Struktur und die Parameter des Systems eindeutig. Ein solches Modell läßt die Untersuchungen aller Änderungen des Systems in verschiedenen Situationen und in verschiedener Zeit zu, wobei die jeweiligen Ausgangsgrößen x beobachtet werden. Somit sind Untersuchungen des Systems im Bildbereich (die Zeit wird hier als Konstante, d.h. als Variable in Form des Laplace-p-Operators angesehen) und im Zeitbereich (die Variable Zeit stellt hier eine objektive, unabhängige Variable dar) möglich.

Bei der Simulation kann man das Verhalten des Systems bei verschiedenen Eingangssignalen (= dynamische Größen) untersuchen. Ein spezielles Eingangssignal, welches oft bei der Interpretation des Verhaltens der Systeme und deren Klassifikation Anwendung findet, ist das Signal $\delta(t) = 1$, dessen zeitlicher Verlauf in Abbildung 2 gezeigt wird.

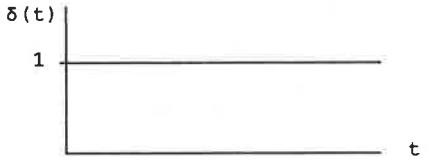


Abb. 2

Die Antwort eines Systems S auf ein solches Eingangssignal ist die sog. Übergangsfunktion h(t) als Ausgangsgröße. Beide Größen, d.h. $\delta(t)$ und h(t), werden wir im Abschnitt 3.3 bei der Simulation des Verhaltens eines speziellen Systems verwenden.

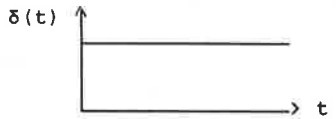
2.2. Experimentelle Systemidentifikation

Oft ist es jedoch unmöglich, aufgrund von theoretischen Überlegungen die Abhängigkeiten zwischen den interessierenden Größen zu modellieren (diese liegen entweder nicht vor oder führen zu sehr komplizierten Modellen). Notwendig ist dann immer die Anwendung der experimentellen Systemidentifikation. In Abhängigkeit von der Spezifik des zu untersu-

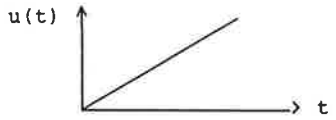
chenden Systems können verschiedene experimentelle Identifikationsmethoden Anwendung finden. Generell untersucht man die Abhängigkeiten zwischen den Eingangs- und Ausgangsgrößen. Man spricht von aktiven Experimenten, wenn man die Möglichkeit hat, künstliche Eingangssignale anzuwenden, im andern Falle hat man es mit passiven Experimenten zu tun. Aktive Experimente werden in Abhängigkeit von der Art des künstlich gewählten Eingangssignal weiter unterschieden:

(a) Experimente mit deterministischen Signalen (zu jedem Zeitpunkt ist eindeutig der Wert dieser Signale festgelegt)

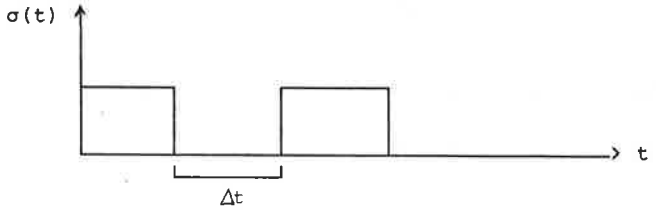
- z.B. Sprungfunktion δ(t)



- Rampenfunktion u(t)

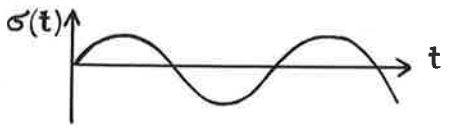


- Impulsfunktion σ(t)



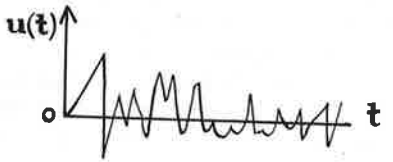
(b) Experimente mit deterministischen periodischen Signalen

- z.B. Sinusfunktion σ(t)



(c) Experimente mit stochastischen Signalen

- stochastische Funktionen u(t)



$E(u(t)) = 0.$

In passiven Experimenten können nur solche Eingangssignale verwendet werden, die tatsächlich in der Praxis als Eingangssignale dieser untersuchten Systeme auftreten. Ein Spezialfall davon ist die statistische experimentelle Identifikation, die stochastische natürliche Signale voraussetzt. Die statistische Identifikation kann nach folgendem allgemeinen Ablaufschema vorgenommen werden:

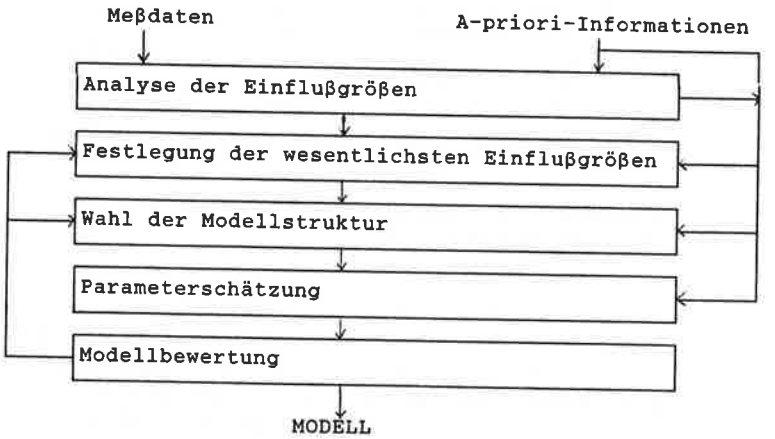


Abb. 3. Ablaufschema - experimentelle statistische Identifikation

Führt man aktive Experimente durch, d.h. verwendet man spezielle künstliche Eingangssignale, so erhält man am Systemausgang eine spe-

zielle Antwort (Ausgangssignal) des Systems. Auf der Grundlage dieser Systemantwort kann der Charakter des Systems, dessen Struktur und Parameter bestimmt werden. Die entsprechenden Methoden findet der Leser in der Fachliteratur (z.B. Reinisch 1974; Isermann 1971; Thoma 1971).

Zusammenfassend kann also festgestellt werden, daß die experimentelle Identifikation die Aufgabe hat, den gemessenen Werten der Eingangsgrößen u und den Ausgangsgrößen x des Systems mit einem vorgegebenen Zielkriterium Θ ein Modell M des Systems S so anzupassen, daß Θ minimiert wird. Dies illustriert Abbildung 4. Θ stellt hierbei die Struktur des Systems dar, A die Matrix der Parameter des Systems und z den Vektor der Störgrößen.

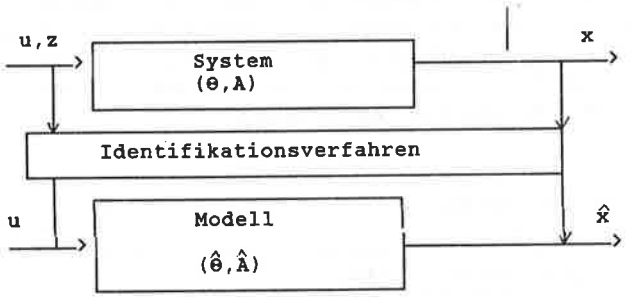


Abb. 4. Experimentelle Systemidentifikation

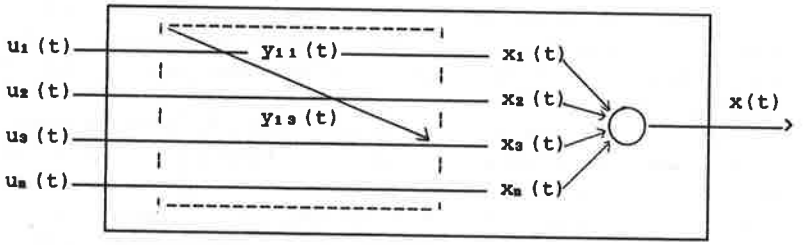
2.3. Zusammengesetzte Methoden der Systemidentifikation

Natürliche Systeme lassen in vielen Fällen weder die Anwendung der reinen theoretischen noch experimentellen Systemidentifikation über aktive Experimente zu. Somit können lediglich passive Experimente der experimentellen Identifikation angewandt werden. Aber in komplizierten Systemen ist die Aussonderung der wichtigen Eingangsgrößen recht schwer, und oft können diese Signale nicht von wirkenden Störgrößen unterschieden werden.

Aus diesem Grunde resultiert die Notwendigkeit einer einführenden theoretischen Erkennung des Systems und eine spätere Berücksichtigung der somit gewonnenen Informationen als A-priori-Information innerhalb einer experimentellen Identifikation.

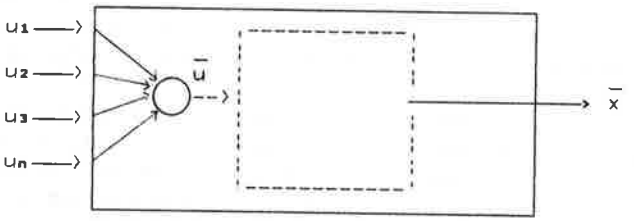
Oben wurden die Eingangs- und Ausgangsgrößen schon als Signale bezeichnet, d.h. als dynamische Größen, die zeitabhängig sind. Die entsprechenden Systeme bezeichnet man als dynamische Systeme im Gegensatz zu statischen Systemen, die einen konkreten Spezialfall dynamischer Systeme für $t = t_0 =$ konstant darstellen. Die oben genannten Identifikationsverfahren gelten grundsätzlich für dynamische Systeme.

In sprachlichen Untersuchungen werden oft statistische statische Systeme untersucht, die jedoch nichts über das Verhalten von Elementarsignalen (Elementargrößen) aussagen, sondern nur über das Wirken deren Summe. Dynamische Systeme dagegen geben die Möglichkeit der Analyse von Elementarsignalen und deren Summe in Abhängigkeit von der Zeit. Dies verdeutlichen Abbildungen 5 und 6.



$u_1(t) = u_2(t) = \ldots = u_n(t_0) = 0$ für $t_0 = 0$
(zugänglich sind die Größen $u_n(t)$, $y_{nn}(t)$, $x_n(t)$ und $x(t)$

Abb. 5. Dynamische Systeme



(zugänglich sind u_n, $\bar{u}$ und $\bar{x}$)

Abb. 6. Statische statistische Systeme

Die Anwendung statischer Modelle in der Systemtheorie ist demzufolge nur dann möglich, wenn das System keine Dynamik enthält und x(t) für alle Zeitpunkte konstant ist (für dasselbe Eingangssignal). Somit kann ein statisches Modell nicht der Ausgangspunkt für die Erarbeitung eines dynamischen Modells sein und von einem statischen Modell kommt man somit auch nicht durch dessen "Verbesserung" (Köhler 1986: 153) zu einem komplizierten dynamischen Modell.

Das zu untersuchende System der natürlichen Sprachen zeichnet sich, wie es schon oben angedeutet wurde, durch seine große Kompliziertheit aus, die sich darin äußert, daß sehr viele sowohl sprachliche als auch außersprachliche Parameter und Variablen Einfluß auf dessen Gestaltung, Spezifik und Entwicklung haben, die oft noch individuellen Charakter tragen.

Bei der Modellierung des Systems der natürlichen Sprachen muß der Sprachträger (Mensch) außerhalb dieses Modells berücksichtigt werden, wobei jedoch die Beziehungen zwischen Eingangsgröße, Ausgangsgröße und Störgrößen aufgezeigt werden müssen.

Dieses System läßt weder die reine theoretische noch die experimentelle Identifikation zu, sondern nur die Kombinierung beider Methoden. Die Spezifik dieses Systems und der Charakter der hier von uns untersuchten Größen (lexikalische Einheiten) bedingen die Anwendung einer speziellen Methode bei der Identifikation des zu untersuchenden Systems. Da man kein aktives Experiment durchführen kann (da man am System keine Manipulationen vornehmen kann), können nur passive Experimente realisiert werden. Dabei ist es theoretisch möglich, die Erkenntnisse über bisherige Sprachgesetze, die ausschließlich statistische Gesetze sind und oft mit stetigen Funktionen beschrieben werden, anzuwenden. Eine Anwendung der statischen statistischen Sprachgesetze ist nicht möglich, da es uns hier um eine dynamische Modellierung gehen muß, die – wie schon oben angedeutet wurde – erst wiederum in einem Spezialfall als statische Modellierung erscheint.

Von den Sprachgesetzen, die Aspekte eines dynamischen Systems beschreiben, ist für unsere späteren Analysen vor allem das Piotrowski-Gesetz geeignet, welches in der Interpretation Altmanns (1983) einen sehr allgemeinen Charakter annehmen kann. Das Ziel der Identifikation ist somit die Ableitung der Übertragungsfunktion G(p), welche in unseren weiteren Untersuchungen unter Anwendung systemtheoretischer Methoden auf der Grundlage von Abhängigkeiten zwischen Eigenschaften lexikalischer Einheiten, die mit dem verallgemeinerten Piotrowski-Gesetz (mit stetigen Funktionen) beschrieben werden können, ermittelt werden soll. Da die untersuchten lexikalischen Eigenschaften diskreten Charakter besitzen,

wäre es angebracht, die Identifikation des zu untersuchenden lexikalischen Systems als diskretes System vorzunehmen, was wiederum die Anwendung eines anderen systemtheoretischen Methodeninventars erfordern würde. Da jedoch bisher keine diskreten dynamischen Sprachgesetze der Lexik vorliegen, soll in dieser Arbeit, wo es uns vorwiegend auf das Aufzeigen eines Lösungsweges für die Identifikation des lexikalischen Systems (hinsichtlich bestimmter lexikalischer Eigenschaften) als dynamisches System ankommt, auf die Anwendung dieser Methoden verzichtet werden. Dies soll Gegenstand einer selbständigen, größeren Arbeit sein.

Untersucht man dynamische Systeme, d.h. die Entwicklung bestimmter Eigenschaften dieser Systeme (die Änderung deren Zustände) in Abhängigkeit von der Zeit (stationäre Systeme) nach den oben vorgestellten Identifikationsverfahren, so wird stets vorausgesetzt, daß sich die Struktur und die Parameter dieser Systeme in der Zeit nicht ändern. Da diese Bedingungen für das lexikalische System sicher nicht eingehalten werden können, muß überprüft werden, welche Parameter und welche Elemente der Struktur der Systeme verändert werden (adaptive Systeme), was Informationen über die Evolution der Systeme liefert (stabiler Systemzustand – instabiler Systemzustand). Dann kann auch untersucht werden, welche äußeren Faktoren (Bedürfnisse) die beobachteten Systemveränderungen bewirken.

Eine zweite Voraussetzung für die Anwendung der besprochenen Identifikationsverfahren ist die Annahme von der Linearität des Systems, was bedeutet, daß einem konkreten Wert des Eingangssignals ein und nur ein Ausgangssignalwert (für einen konkreten Zeitpunkt t_0) entspricht. Diese Bedingung kann nicht von vornherein als erfüllt angesehen werden (vor allem auch nicht für das hier zu untersuchende lexikalische System). Aus diesem Grunde untersucht man zunächst Subsysteme hinsichtlich dieser Bedingung, erstellt für die Subsysteme $G_i(p)$ und setzt dann G(p) für das gesamte System aus diesen Teilwerten $G_i(p)$ über einfache Operationen zusammen.

Ist die Bedingung der Linearität des Systems nicht erfüllt, finden leicht modifizierte Methoden Anwendung bei der Suche der Übertragungsfuntion G(p).

Die Bedingung der Linearität ist immer dann bei der statistischen Modellierung von Systemen erfüllt, wo man in einem gewissen Arbeitspunkt und dessen Umgebung linearisiert, obwohl das untersuchte System selbst nichtlinear sein kann.

In unseren weiteren Untersuchungen eines lexikalischen Subsystems setzen wir voraus, daß diese Bedingung erfüllt ist, obwohl wir nicht in der Lage sind, diese Bedingung durch Analysen empirischer Daten zu

überprüfen (da bisher keine entsprechenden empirischen Untersuchungen über die Abhängigkeiten lexikalischer Eigenschaften in der Zeit durchgeführt wurden).

3. Diskussion des Modells zur Struktur und Dynamik der Lexik von Köhler

3.1. Allgemeine Charakteristik des Modells

Köhler (1986) untersucht ein statisches Subsystem der Lexik, indem er die gegenseitigen Abhängigkeiten von 4 lexikalischen Eigenschaften in Abhängigkeit von außersprachlichen Bedürfnissen und innersprachlichen Parametern analysiert.

Hierbei handelt es sich um die Abhängigkeiten zwischen der Eigenschaft der Polylexie (PL) und der Länge (L) lexikalischer Enheiten, der Polytextie (PT) und der Polylexie, der Frequenz (F) und der Polytextie, der Länge und der Frequenz.

Die Modellstruktur für den Zusammenhang dieser Größen beruht auf der Annahme, daß die relative Veränderungsrate einer Größe umgekehrt proportional zum entsprechnden Wert der anderen Größe ist, d.h. allgemein gilt:

$$\frac{y'}{y} = -\frac{N}{x}, \qquad (1)$$

wo x die unabhängige Variable, y die abhängige Variable ist und N einen Proportionalitätsfaktor darstellt. Die Lösung dieser Differentialgleichung führt Köhler auf folgende Abhängigkeiten:

$$PL = a_1 \cdot L^{b_1} \qquad (2)$$

$$PT = a_2 \cdot PL^{b_2} \qquad (3)$$

$$F = a_3 \cdot PT^{b_3} \qquad (4)$$

$$L = a_4 \cdot F^{b_4} \qquad (5)$$

Die Abhängigkeiten (2) – (4) konnten in empirischen Untersuchungen – zunächst nur an deutschem Material – überprüft und umfassend bestätigt werden, nicht aber Abhängigkeit (5).

Köhler untersucht ein dynamisches System mit einem statischen statistischen Modell, welches nicht Resultat einer dynamischen Modellierung und der Feststellung G(p) = konstant ist und auch nicht von der dann notwendigen Voraussetzung gestützt wird, daß die Veränderungen des untersuchten lexikalischen Subsystems in der Zeit nicht signifikant sind (d.h. G(p) ist eine Konstante). Im Gegenteil, die Variable Zeit, die im Modell nicht berücksichtigt wird, kommt bei der Modellierung durch die HIntertür wieder in Erscheinung, was z.B. an folgendem Textfragment, welches zur Modellierung der Abhängigkeit zwischen L und F dient, verdeutlicht werden soll: "...daß Ausdrücke...durch steigende Gebrauchshäufigkeit *in naher Zukunft* eine solche Verkürzung erfahren oder aber durch kürzere ersetzt werden... Die Kürzung eines Wortes kann also je nach Länge *beschleunigt* oder *gebremst* werden" (Köhler 1986:144; Hervorhebung durch R.H. und J.M.).

Die Nichtberücksichtigung der Variablen Zeit bei der Untersuchung dieser Abhängigkeit ist die Hauptursache dafür, daß das Köhlersche Modell diese Abhängigkeit nich umfassend beschreiben kann. Da die Köhlerschen Abhängigkeiten statische Abhängigkeiten sind, d.h. nur für einen bestimmten Zeitpunkt $t = t_0$ gelten, sagen sie nichts über die (zeitlichen) Veränderungen konkreter lexikalischer Einheiten bezüglich der untersuchten Eigenschaften aus, sondern nur über einen durchschnittlichen Entwicklungszustand der untersuchten lexikalischen Einheiten (da angenommen werden kann, daß sich die untersuchten lexikalischen Einheiten zum Zeitpunkt $t = t_0$ in einem verschiedenen zeitlichen Entwicklungszustand befinden). Deshalb ist auch eine Interpretation wie "Die Veränderung der Kürzungsrate *einer* lexikalischen Einheit ist eine Funktion der Länge und der Frequenz." (Köhler 1986:144, Hervorhebung durch R.H., J.M.) nicht gerechtfertigt. Die Köhlerschen Abhängigkeiten sagen nur etwas über die durchschnittlichen Differenzen bezüglich der Länge innerhalb einer Gruppe von lexikalischen Einheiten zu einem bestimmten Zeitpunkt aus. Das statische Modell muß statisch und nicht dynamisch interpretiert werden.

Köhler (1986:73) bemerkt bei der Modellierung der Abhängigkeit zwischen L und F, daß hier ein Rückkopplungseinfluß der Länge auf die Frequenz, also deutlich ein zeitlich versetzter Einfluß, zu berücksichtigen

ist, was jedoch bei der Ableitung des entsprechenden mathematischen Modells außer acht gelassen wurde. Er präzisiert auch nicht, welchen Charakter dieses Rückführungsglied hat. Aus der Systemtheorie sind hierfür das Trägheitsglied und das Totzeitglied bekannt, deren zeitliche Charakteristik in den folgenden Abbildungen gezeigt wird.

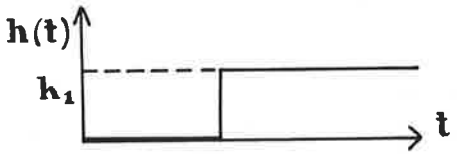


Abb. 7. Abhängigkeit des Ausgangssignals h(t) von der Zeit t für das Totzeitglied

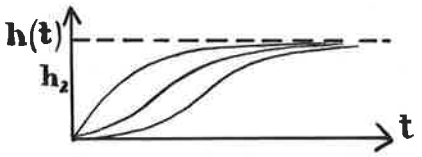


Abb. 8. Abhängigkeit des Ausgangssignals h(t) von der Zeit für das Trägheitsglied

Eine Außerachtlassung der Rückführung bei der Modellierung der Abhängigkeit zwischen F und L hat zur Folge, daß das erstellte Modell keine Möglichkeit zur Selbstregulierung besitzt.

Obwohl Köhler gezeigt hat, daß die Funktion $L = aF^b$ die Abhängigkeit der entsprechenden empirischen Daten adäquat beschreibt, so kann das jedoch noch nicht als Beweis für die Richtigkeit dieses theoretischen Modells gelten, vor allem dann nicht, wenn die Ableitung dieses Modells nicht korrekt ist.

Es kann gezeigt werden, daß eine verschiedene Zahl anderer mathematischer Funktionen, die wohl auch teilweise theoretisch begründet werden können, die Abhängigkeit der empirischen Daten adäquat beschreibt.

An anderer Stelle bemerkt Köhler (1986:86), daß man bei Kenntnis des zeitlichen Verlaufs einer der untersuchten lexikalischen Eigenschaften und des von Köhler vorgestellten Basismodells (vgl. Gleichung (2) - (6)) somit gleichzeitig das dynamische Verhalten sämtlicher "struktureller Eigenschaften in ihrem zeitlichen Verlauf" vorhersagen kann. Eine solche Schlußfolgerung ist aus systemtheoretischer Sicht abzulehnen. Für jede lexikalische Eigenschaft müßte deren Zeitabhängigkeit in einem dynamischen Modell modelliert werden, um dann die Abhängigkeiten zwischen je zwei lexikalischen Eigenschaften in einem dynamischen Modell untersuchen zu können.

Später werden wir noch ein Modell diskutieren, welches die Abhängigkeiten zwischen drei Eingangs- und einer Ausgangsgröße bezüglich der von Köhler untersuchten 4 lexikalischen Eigenschaften betrifft. Dies stellt eine Erweiterung der soeben vorgestellten Methode der dynamischen Modellierung dar.

Das von Köhler vorgestellte statische Modell der Lexik weist gegenüber einem dynamischen Modell noch weitere wesentliche Nachteile auf, da man den (zeitlichen) Verlauf der jeweiligen Abhängigkeiten nicht in einem Simulationsexperiment untersuchen und mit dem Verhalten des entsprechenden Systems in der Realität vergleichen kann; in diesem Sinne ist nur ein dynamisches Modell nachprüfbar.

3.2. Vorschlag für die Erstellung eines dynamischen Lexikmodells

In unseren weiteren Untersuchungen in diesem Abschnitt werden wir versuchen, den methodischen Weg für die Erstellung eines dynamischen Modells zu skizzieren, welches die Abhängigkeiten zwischen der Länge L lexikalischer Einheiten und deren Frequenz F betrifft.

Weder empirische Untersuchungen noch bisherige theoretische Voraussetzungen legen eindeutig fest, welche der beiden Variablen L oder F als abhängige bzw. unabhängige Variable bei der Beschreibung deren Abhängigkeit anzusehen ist. Zumindest an dieser Stelle müssen und können wir uns noch nicht eindeutig festlegen. Unten werden wir dieses Problem noch aus systemtheoretischer Sicht diskutieren.

Aus diesem Grunde werden wir 2 lexikalische Subsysteme und die ihnen entsprechenden Modelle M_1 und M_2 untersuchen (vgl. Abb. 9 a und b).

Abb. 9a. Dynamisches Modell für die Untersuchung der Abhängigkeit der Frequenz von der Länge

Abb. 9b. Dynamisches Modell für die Untersuchung der Abhängigkeit der Länge von der Frequenz

Wir werden später versuchen, das Verhalten beider Modelle zu simulieren.

Voraussetzung für die Ableitung der Modelle M_1 und M_2 ist die Kenntnis der Abhängigkeiten der Länge von der Zeit, d.h. L(t), und der Frequenz von der Zeit, d.h. F(t). Wir verfügen jedoch nicht über empirische Daten, die eine Modellierung beider Abhängigkeiten erlauben würden. Aus diesem Grunde wenden wir das von Altmann verallgemeinerte Piotrowski-Gesetz an (vgl. Altmann 1983:59 ff.), welches eine "hypothetische Aussage über den zeitlichen Verlauf der Veränderungen einer beliebigen sprachlichen Entität" liefert.

Da das Piotrowski-Gesetz somit Veränderungen beliebiger sprachlicher Größen in der Zeit beschreibt, ist es auch möglich, dieses Gesetz auf die Beschreibung der zeitlichen Veränderung der Frequenz und der Länge lexikalischer Einheiten zu übertragen.

Nach dem Piotrowski-Gesetz für unvollständige Veränderungen (Altmann 1983:60 ff.) ist der Zuwachs neuer Formen u' durch

$$u' = bu(c - u) \tag{6}$$

festgelegt, wo b und c Konstanten sind.

Wird z.B. ein Wort der Länge L durch steigende Frequenz auf L-m gekürzt (z.B. "Demonstration" auf "Demo"), so kann man dieses Gesetz auf die zeitliche Veränderung der Länge L anwenden.

Ändert sich dagegen die Frequenz lexikalischer Einheiten, d.h. steigt sie zunächst an und fällt sie später wieder ab, so kann man das Piotrowski-Gesetz für eine reversible Veränderung (Altmann 1983:61 ff.) anwenden. Dann gilt:

$$x' = (b - ct)x(1 - x), \tag{7}$$

wo wiederum b und c Konstanten darstellen.

Nicht zufällig wurden in den Gleichungen (6) und (7) die Symbole u (bzw. u') und x (bzw. x') verwendet, da wir zuerst Modell M_1 analysieren werden, wo L die Eingangsgröße (in systemtheoretischen Arbeiten wird diese Eingangsgröße mit dem Symbol u bezeichnet) und F (dies entspricht dem Symbol x) die Ausgangsgröße ist.

Wie wir schon oben bemerkt haben, werden Modelle in der Systemtheorie mit Hilfe der sogenannten Übertragungsfunktion G(p) beschrieben, wobei gilt:

$$G(p) = \frac{L(x(t))}{L(u(t))}, \tag{8}$$

wobei der Ausdruck im Nenner die Laplace-Transformation des Ausgangssignals und der Ausdruck im Zähler die Laplace-Transformation des Eingangssignals bei verschwindenden Anfangsbedingungen [x(t=0) = 0, u(t=0) = 0) darstellt.

Setzt man in Gleichung (7) für x(t) F(t) ein und in Gleichung (6) für u(t) L(T), so gilt:

$$L'(t) = b_1 L(t)(c_1 - L(t)) \tag{9}$$

$$F'(t) = (b_2 - c_2 t)F(t)(1 - F(t)). \tag{10}$$

Auf beide Funktionen muß nun die Laplace-Transformation angewandt werden, welche ein mathematisches Hilfsmittel darstellt, das die notwendigen Rechnungen wesentlich vereinfacht und leicht eine Beschreibung des untersuchten Modells erlaubt. Es gibt eindeutige Transformationsregeln für den Übergang aus dem Zeitbereich (Gleichung (9) und (10)) in den Bildbereich (Gleichung (11) und (12)), die in der Fachliteratur umfassend beschrieben werden (vgl. z.B. Reinisch 1974:278 ff.); aus diesem Grunde führen wir hier nur das Ergebnis dieser Transformation von Gleichung (9) und (10) an.

$$pL(p) = b_1 L(p)(c_1 - L(p)) \tag{11}$$

$$pF(p) = (b_2 - c_2/p^2)\, F(p)(1 - F(p)). \tag{12}$$

Durch Umformung erhält man:

$$F(p) = \frac{p^3 - b_2 p^2 + c_2}{-b_2 p^2 + c_2} \tag{13}$$

$$L(p) = \frac{b_1 c_1 - p}{b_1}. \tag{14}$$

Die entsprechende Übertragungsfunktion für Modell M_1 lautet dann:

$$G(p) = \frac{F(p)}{L(p)} = \frac{b_1 p^3 - b_1 b_2 p^2 + b_1 c_2}{b_2 p^3 - b_1 c_1 b_2 p^2 + c_2 b_1 c_1 - c_2 p}. \tag{15}$$

Wie wir schon angedeutet haben, beschreibt die Übertragungsfunktion die Struktur des Modells. Für selbstregulierende Systeme ist es außerordentlich wichtig, zu erkennen, wie die Rückkopplung wirkt, d.h. an welcher Stelle und über welche Größen sich die Ausgangsgrößen auf die Eingangsgrößen einwirken. Um diese Frage lösen zu können, werden wir Gleichung (15) umformen (der Zähler und Nenner wird mit $b_2 p^{-3}$ erweitert); so erhält man bei der Auflösung nach F(p):

$$F(p) = \left(\frac{b_1}{b_2} - b_1 p^{-1} + \frac{b_1 c_2}{b_2} \cdot p^{-3}\right) E(p) \tag{16}$$

mit

$$E(p) = L(p) - \left(-c_1 b_1 p^{-1} - \frac{c_2}{b_2} \cdot p^{-2} + \frac{b_1 c_1 c_2}{b_2} \cdot p^{-3}\right) E(p). \tag{17}$$

Die Struktur von Modell 1 mit der in Gleichung (15) angegebenen Übertragungsfunktion kann nun in Form eines Blockschaltbildes dargestellt werden (vgl. Abb. 10).

den (vgl. Abb. 10).

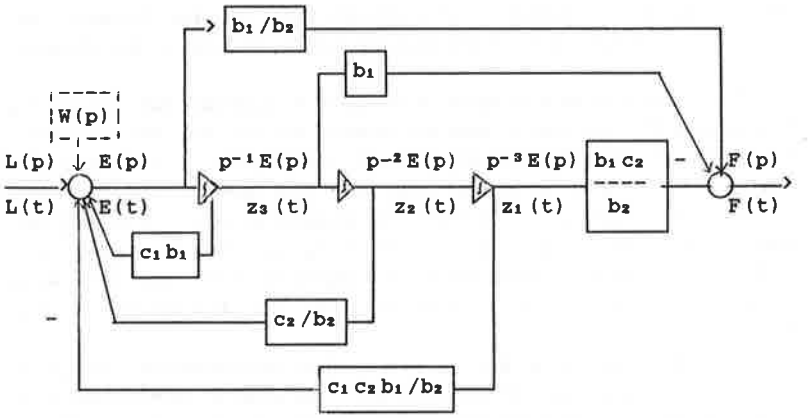


Abb. 10. Blockschaltbild von Modell M_1

Wie Abbildung 10 zeigt, ist in der Abhängigkeit zwischen L und F tatsächlich eine Rückkopplung vorhanden, die Faktoren $c_1 b_1$ z.B. wirken positiv und der Faktor $c_1 c_2 b_1 / b_2$ negativ stabilisierend. Der letzte Faktor ist verantwortlich für die Selbstregulierung des Modells und demzufolge auch des entsprechenden Systems. Unter Anwendung der vorgestellten Methode haben wir somit eine direkte Möglichkeit, eine Aussage darüber zu machen, wie die Konstanten c_1, c_2, b_1 und b_2 der Zeitverläufe von L und F das Verhalten des ganzen Systems beeinflussen. Die Größen $z_1(t)$, $z_2(t)$ und $z_3(t)$ werden Zustandsgrößen genannt, über die L(t) mit F(t) verbunden ist; somit sind natürlich auch die Zustandsgrößen untereinander verbunden, was wir an einem Beispiel zeigen wollen.

$\dot{z}_3(t)$ entspricht $z_2(t)$ bzw. $z_2(t)dt$ entspricht $z_3(t)$, wenn man zusätzlich die rückwärts- bzw. vorwärtsführenden Abzweigungen berücksichtigt. Von $z_3(t)$ führt ein Weg über den Faktor $c_1 b_1$ zurück, d.h. es ist hier schon $z_3(t) - c_1 b_1$ zu beachten, es führt aber auch noch ein Weg mit $-b_1$ nach vorn; von $z_2(t)$ führt nur ein Weg mit dem Faktor c_2/b_2 zurück (zu beachten ist somit $-c_2/b_2$). Somit kann folgende Gleichung aufgestellt werden:

$$\dot{z}_3(t) - b_1c_1 - (-b_1) = z_2(t) - c_2/b_2 \quad . \qquad (17)$$

Auf analoge Weise kann man alle anderen Zustandsgrößen und im Endeffekt auch Eingangs- und Ausgangsgröße über ein System von Differentialgleichungen, die es dann zu lösen gilt, verbinden. Somit kann dann der Einfluß aller Parameter und Zustandsgrößen auf den Zusammenhang zwischen Eingangs- und Ausgangsgröße exakt modelliert und simuliert werden

Die Größe E(p) stimmt stark vereinfacht ausgedrückt mit der aus der Regelungstechnik bekannten Regelabweichung überein, die dort als Differenz zwischen Ist- und Sollwert definiert ist. Wenn im untersuchten System ein Selbstregulierungsmechanimus auftreten soll, so muß auch E(p) auftreten. E(p) enthält den Einfluß von Störgrößen (z.B. den Einfluß anderer lexikalischer Eigenschaften, der durch das Wirken anderer, noch nicht berücksichtigter Bedürfnisse hervorgerufen wird; die Bedürfnisse können auch als zeitveränderliche Sollwerte, d.h. Führungsgrößen W(p) wirken).

Über die Anwendung des entsprechenden mathematischen Apparates (z.B. Bode-Diagram) ist man in der Lage, eine genaue Interpretation des Einflusses der einzelnen Parameter auf den Prozess der Selbstregulierung (Streben gegen den Sollwert) vorzunehmen, die Grenzwerte für die Stabilität des Systems zu bestimmen und den Einfluß der Parameter auf die Geschwindigkeit der Regulierung (z.B. durch Ermittlung der Einschwingzeit, der bleibenden Regelabweichung, d.h. der Abweichung vom Sollwert, die das System im Selbstregulierungsprozess nicht kompensiert).

Vergleicht man Abb. 10 mit den von Köhler (1986:66 ff., besonders S. 74) entwickelten Blockschaltbildern, so ist wohl ein qualitativer Unterschied nicht zu versehen. Erstens ist Köhler nicht in der Lage, einen genauen und aus einem mathematischen Modell resultierenden Einfluß der Parameter auf die Abhängigkeiten zwischen Eingangs- und Ausgangsgröße anzugeben, zweitens wurden Zustandsgrößen und Regelabweichung nicht berücksichtigt und drittens sind die aufgeführten Blockschaltbilder in mehreren Fällen nicht äquivalent mit der von Köhler angegebenen mathematischen Beschreibung der Abhängigkeiten im Blockschaltbild (es bleibt für uns z.B. völlig unklar, auf welche Weise Köhler (1986:74) die Richtungen der gegenseitigen Abhängigkeiten der 4 untersuchten Eigenschaften und der berücksichtigten Parameter ermittelt).

Hat man die Übertragungsfunktion G(p) abgeleitet, so ist man auch in der Lage, die Übertragungsfunktion h(t) abzuleiten, die eine Antwort des Systems (zeitlicher Verlauf des Ausgangssignals auf eine sprungförmige Änderung des Eingangssignals) z.B. in einem Simulationsexperiment ist. h(t) kann man unter Anwendung der Laplace-Umkehrtransformation L^{-1} ableiten. Hierbei gilt:

$$h(t) = L^{-1}\left[\frac{1}{p} - G(p)\right] \qquad (18)$$

Die Übertragungsfunktion h(t) wird uns besonders in den folgenden Simulationsexperimenten interessieren.

3.3. Simulationsexperimente

3.3.1. Modell M_1

Mit Hilfe einer Computersimulation wird zunächst der zeitliche Verlauf der Frequenz als Ausgangsgröße als eine Antwort auf eine charakteristische Sprungfunktion δ(t) = konstant bei verschiedenen Parametern untersucht.

Zunächst formen wir Gleichung (15) etwas um (über eine Polynomdivision), da es für die Simulation aus technischen Gründen notwendig ist, daß die höchste Potenz von p im Zähler kleiner ist als im Nenner. Somit erhält man:

$$G(p) = -\frac{b_1}{b_2} + \frac{p^2(b_1^2c_1 - b_1b_2) + p(c_2b_1/b_2) + b_1c_2(1 - b_1c_1/b_2)}{p^3b_2 - p^2b_1c_1b_2 - pc_2 + c_2c_1b_1}$$

Mit

$A = -b_1/b_2$ $\qquad$ $r_3 = b_2$

$d_2 = b_1^2c_1 - b_1b_2$ $\qquad$ $r_2 = b_1c_1b_2$

$d_1 = c_2b_1/b_2$ $\qquad$ $r_1 = c_2$

$d_0 = b_1c_2(1 - (b_1c_1/b_2))$ $\qquad$ $r_0 = c_2c_1b_1$

gilt

$$G(p) = A + \frac{p^2 d_2 + p d_1 + d_0}{p^3 r_3 + p^2 r_2 + p r_1 + r_0}. \qquad (19)$$

Unter Vorgabe bestimmter Werte für b_1, b_2, c_1, c_2 (woraus leicht die Werte für A, d_2, d_1, d_0, r_3, r_2, r_1, r_0 errechnet werden können) wurde nun eine Computersimulation durchgeführt. Da es für unsere Simulation vorteilhaft war, die letzteren Parameter zu verwenden, werden wir nur diese anführen. Dabei werden wir bewußt solche Parameterwerte auswählen, für die das System in Abhängigkeit von der Zeit stabil ist, an der Grenze der Stabilität bzw. instabil.

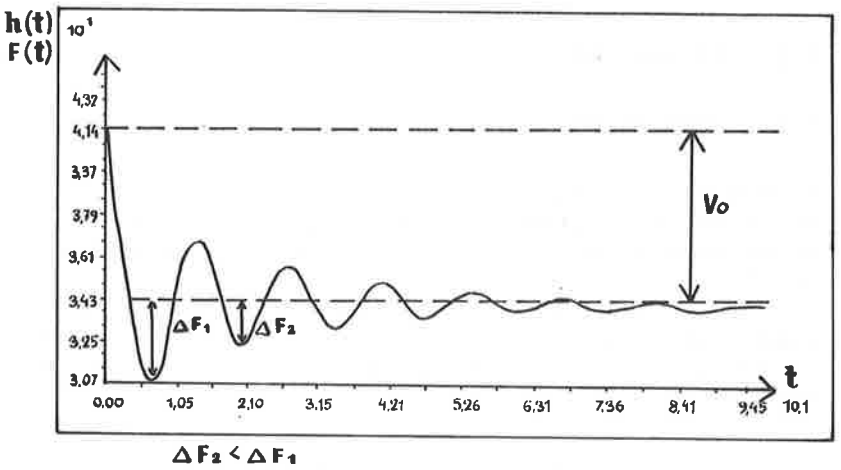


Abb. 11. Simulation des Modells M_1 (Stabilität)
(A = 15.0; d_0 = -3.52; d_1 = -0.8; d_2 = -9.0; r_0 =1.0;
r_1 = 0.8; r_2 = 5.0; r_3 = 2.0; δ(t) = ΔL = 3.0)

Ein System ist dann stabil, wenn die Ausgangsgröße x(t) für t --> ∞ einen vorgegebenen Wert ε nicht übersteigt, d.h. x(t) < ε.

Ein System liegt an der Grenze der Stabilität, wenn entsprechend x(t) = ε für t --> ∞ ist.

Instabil ist ein System dementsprechend dann, wenn x(t) --> ± ∞ für t --> ∞.

Wie man aus Abbildung 11 ablesen kann, werden die Veränderungen der Frequenz ΔF_1 bei einer Vergrößerung der Länge ΔL = 3.0 mit wachsendem t immer kleiner, bis F einen bestimmten Wert annimmt. Es kann somit ein Wert ε gefunden werden, für den F(t) < ε gilt, d.h. das System nimmt einen stabilen Zustand an. Dabei ist zu bemerken, daß die Zeitachse t mit beliebigen Einheiten (z.B. Tage, Jahre, Jahrhunderte) belegt werden kann, was jedoch für unsere Untersuchungen, in denen die verwendeten Parameter nicht aus konkreten empirischen Untersuchungen gewonnen wurden, von untergeordneter Bedeutung ist.

Eine tiefgründige Interpretation der in Abbildung 11 dargestellten Abhängigkeit wird auch dadurch erschwert, daß bei der Simulation eine stetige Funktion verwendet wird.

Da das Piotrowski-Gesetz sowohl auf einzelne lexikalische Einheiten als auch auf bestimmte Mengen lexikalischer Einheiten anwendbar ist, so kann auch Abbildung 11 hinsichtlich dieser beiden Fälle interpretiert werden. Die von uns gewählten Parameter geben jedoch diesbezüglich keine Informationen. Deshalb kann hier nur allgemein gesagt werden, daß lexikalische Einheiten, auf die die gewählten Parameter zutreffend sind, zwei entgegengesetzt gerichteten Wirkungskräften des Modells (und bei Richtigkeit des Modells auch des Systems) unterliegen, wenn deren Länge um 3 Einheiten größer wird. Eine der Kräfte wirkt in Richtung der Aufrechterhaltung der alten, längeren Formen, die andere in Richtung der Verwendung der neuen, kürzeren Formen. So erklären wir uns die Schwingungen der Frequenz in Abhängigkeit von der Zeit. Die Abnahme der Amplitude könnte dann davon zeugen, daß ein bestimmter Gleichgewichtszustand angestrebt wird. V_0 gibt dann an, um welchen Wert sich die Frequenz in diesem Fall verkleinert hat.

Abbildung 12 zeigt das Verhalten des Systems bei denselben Parametern, wobei lediglich das gewählte Eingangssignal δ(t) = ΔL = -3.0 und A = -15.0 beträgt. Wir untersuchen somit einen Fall, wo die Länge lexikalischer Einheiten um 3 Einheiten verkleinert wird, und beobachten die dadurch hervorgerufenen Änderungen der Frequenz als Ausgangssignal in Abhängigkeit von der Zeit.

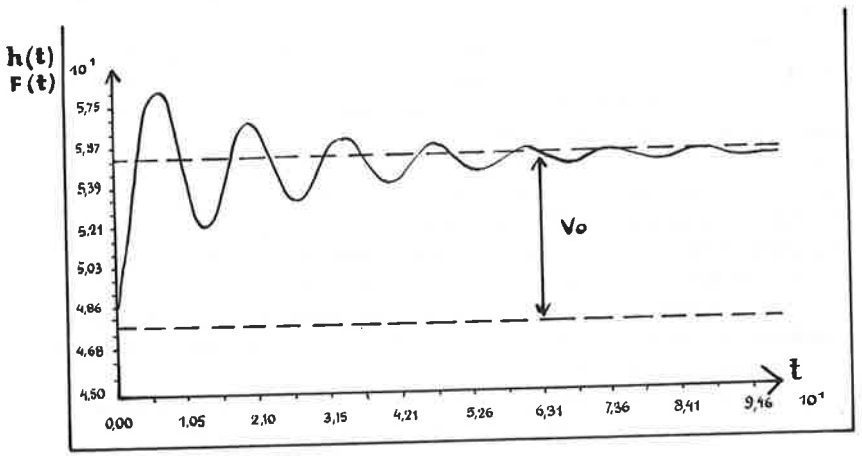


Abb. 12. Simulation des Modells M_1 (Stabilität)
($\delta(t) = \Delta L = -3.0$; $A = -15.0$; alle anderen Parameterwerte - siehe Abb. 11)

Analog dem in Abb. 11 dargestellten Systemverhalten nimmt das System auch hier nach einer gewissen Zeit einen stabilen Zustand an. Wie zu erwarten war, ruft eine Verkürzung der Länge eine Zunahme der Frequenz hervor.

Für den in Abb. 13 dargestellten Fall schwingt die Frequenz mit gleicher Amplitude, d.h. beide wirkenden Kräfte haben den gleichen Betrag. Obwohl am Eingang des Modells eine Vergrößerung der Länge der lexikalischen Einheiten vorgegeben wurde, nimmt die Frequenz am Ausgang zunächst zu. Diese Tatsache als auch die Schwingung mit gleicher Amplitude werden durch die vorgegebenen Parameterwerte verursacht. Damit wurde jedoch nicht gezeigt, daß es lexikalische Einheiten gibt, die diese Parameterwerte erfüllen, sondern nur, daß das System bei der Existenz solcher lexikalischen Einheiten an der Grenze der Stabilität liegt. Es ist jedoch klar, daß ein solches Systemverhalten für t --> ∞ in der Realität nicht möglich ist, d.h. das untersuchte lexikalische Subsystem muß diesen Systemzustand überwinden, falls er für eine längere Zeitdauer auftreten würde.

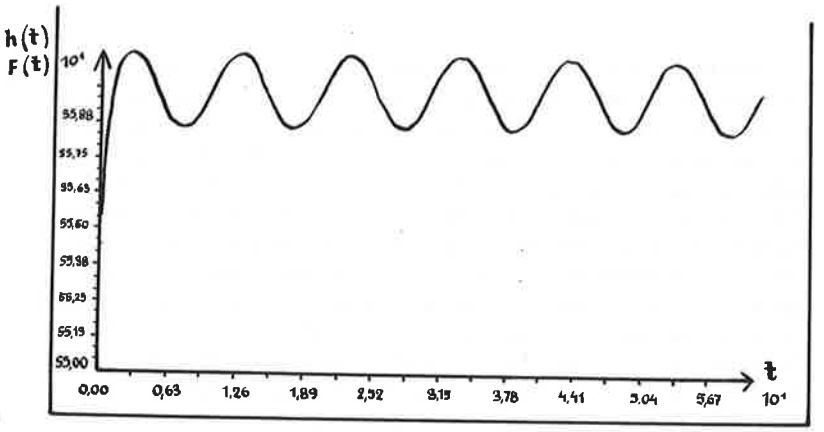


Abb. 13. Simulation des Modells M_1 (an der Grenze der Stabilität)
($A = 10.0$; $d_0 = 3.52$; $d_1 = 0.8$; $d_2 = 9.0$; $r_0 = 2.0$; $r_1 = 0.8$; $r_2 = 5.0$; $r_3 = 2.0$; $\delta(t) = \Delta L = 5.0$)

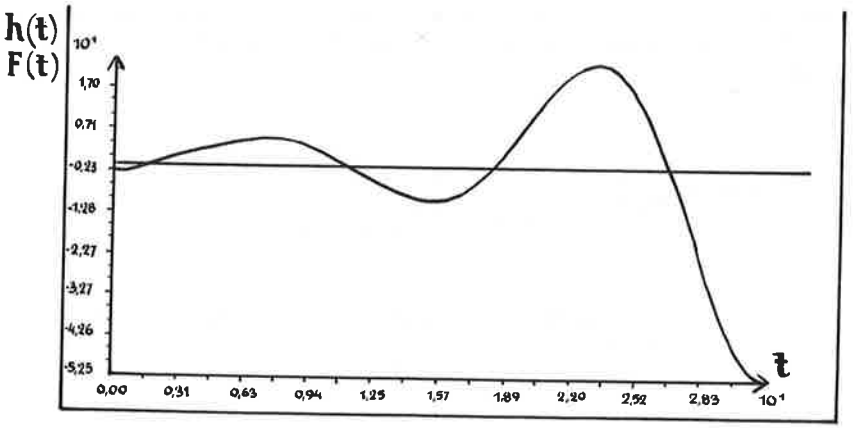


Abb. 14. Simulation der Modells M_1 (Instabilität)
($A = -1.0$; $d_0 = 3.52$; $d_1 = 0.8$; $d_2 = 9.0$; $r_0 = 1.0$; $r_1 = -0.8$; $r_2 = 5.0$; $r_3 = 2.0$; $\delta(t) = \Delta L = 1.0$)

Wie man aus Abb. 14 ersehen kann, bewirkt eine Zunahme der Länge zunächst eine Zunahme der Frequenz und dann eine Abnahme bis auf F = 0. Das Modell sagt voraus, daß eine solche lexikalische Einheit nach einer gewissen Zeit wieder mit einer noch höheren Frequenz auftreten kann. Inwieweit dies jedoch in der Realität möglich ist, kann hier nicht diskutiert werden. Der Verlauf dieser Abhängigkeit würde sicher anders aussehen, wenn man mit diskreten Funktionen und mit bestimmten Wertebereichsbeschränkungen (z.B. $F \geq 0$, $L \geq w$, wo w eine natürliche Zahl ist) arbeiten würde. Abb. 14 kann auch noch unter einem anderen Aspekt diskutiert werden: Wie man sieht, ist die Häufigkeit der lexikalischen Einheiten, die die geforderten Parameterwerte erfüllen, relativ klein. Wenn die aus diesem Grunde schon relativ langen Einheiten zusätzlich verlängert werden, dann erhält man Einheiten, deren Verwendung in der sprachlichen Kommunikation uneffektiv und unökonomisch ist, d.h. diese Einheiten werden aus dem lexikalischen System verdrängt.

3.3.2. Modell M_2

Wie schon erwähnt wurde, wird für Modell M_2 F(t) als Eingangsgröße und L(t) als Ausgangsgröße gewählt. Die Übertragungsfunktion für diesen Fall lautet:

$$G(p) = \frac{L(p)}{F(p)} = \frac{p^3b_2 - p^2b_1c_1b_2 - pc_2 + b_1c_1c_2}{p^3b_1 - p^2b_1b_2 + b_1c_2}$$

$$= \frac{-b_2}{b_1} + \frac{p^2(b_2 - b_1c_1b_2) - c_2p + c_2(b_1c_1 - b_2)}{p^3b_1 - p^2b_1b_2 + b_1c_2}$$

$$= A + \frac{p^2d_2 + pd_1 + d_0}{p^3r_2 + p^2r_1 + r_0}. \qquad (20)$$

Das entsprechende Blockschaltbild zeigt Abb. 15.

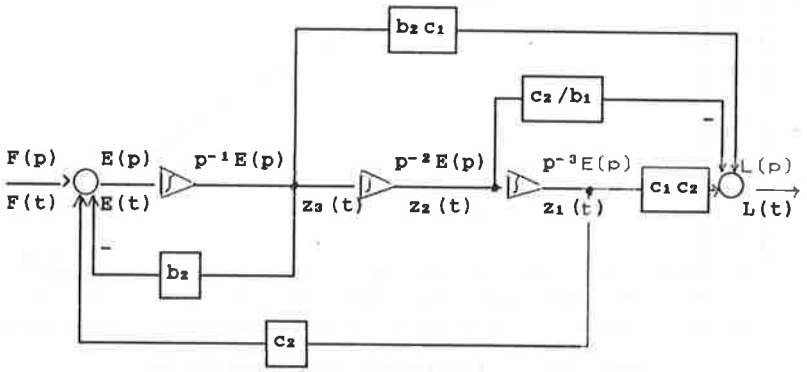


Abb. 15. Blockschaltbild von Modell M_2

Abbildung 15 zeigt, daß für diese Abhängigkeit keine negative Rückführung zwischen Ausgang und Eingang besteht. Man kann zwar für c_2 negative Werte für die Simulation wählen, um diese negative Rückkopplung herzustellen, wobei sich aber gleichzeitig das Vorzeichen anderer Faktoren ändert (z.B. im Faktor c_2/b_1, der auf dem Weg zwischen $z_2(t)$ und L(p) liegt), was den Rückkopplungseffekt rekompensiert. Außerdem fällt in Abbildung 15 das Fehlen einer Rückkopplung zwischen $z_2(t)$ und dem Eingang L(t) auf. Diese strukturellen Bedingungen lassen die Schlußfolgerung zu, daß dieses Modell (System) bei allen möglichen Kombinationen von Parameterwerten instabil ist. Zur genaueren Darstellung und Analyse der Stabilitätsbedingungen müßte man z.B. den sogenannten Phasen- und Amplitudengang und das Hurwitz-Kriterium anwenden (vgl. Philippow 1977:319 ff.), worauf jedoch hier verzichtet werden soll.

Dieses Ergebnis kann folgendermaßen interpretiert werden:

(a) In der Abhängigkeit zwischen L(t) und F(t) ist L(t) die unabhängige und F(t) die abhängige Variable und nicht umgekehrt, wenn man voraussetzt, daß das Piotrowski-Gesetz diese Abhängigkeiten adäquat beschreibt.

(b) Unter der Annahme, daß F(t) die unabhängige Variable ist, kann man das Piotrowski-Gesetz nicht zur Modellierung der Abhängigkeit der Länge L von der Frequenz F heranziehen.

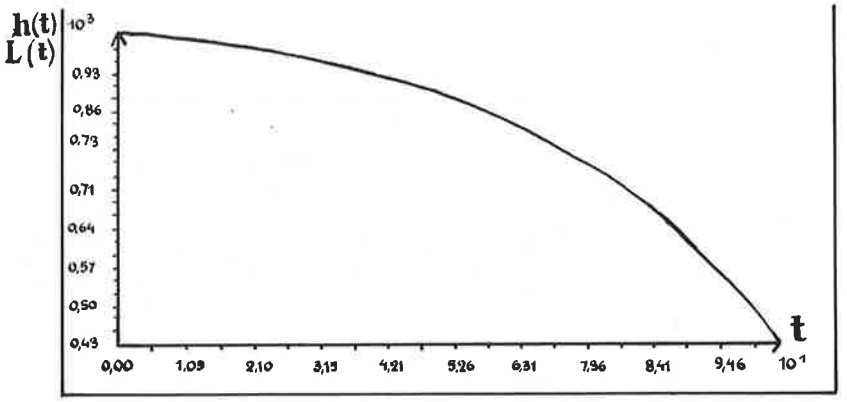


Abb. 16. Simulation des Modell M_2
(A = −20.0; d_0 = −1.56; d_1 = 0.58; d_2 = 6.25;
r_0 = −1.45; r_1 = −6.25; r_2 = 2.5; δ(t) = ΔF(t) = −50.0)

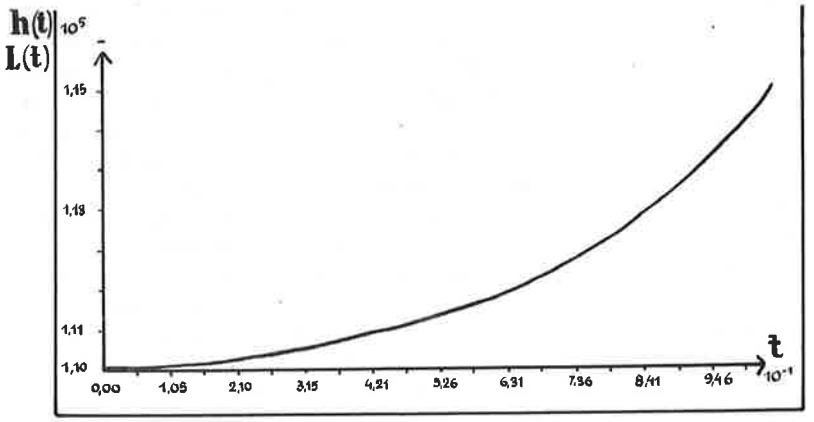


Abb. 17. Simulation des Modells M_2
(A = 220.0; δ(t) = ΔF(t) = 500.0;
andere Parameterwerte – s. Abb. 16)

Diese Interpretation läßt jedoch außer acht, daß

(a) L(t) und F(t) in der Realität nur diskrete Werte annehmen und somit der Beschreibung mit einer diskreten Funktion bedürfen,

(b) L(t) und F(t) bestimmten Wertebereichsbeschränkungen unterliegen.

Die durchgeführten Simulationsexperimente (bei verschiedenen Parameterwerten) bestätigen, daß M_2 ein instabil wirkendes Modell ist, was bedeutet, daß dieses Modell das entsprechende lexikalische System nicht adäquat beschreibt, denn dieses System kann nicht in allen möglichen Fällen instabil wirken.

Abbildung 16 zeigt, daß eine Verminderung der Frequenz ΔF = −50 eine Verkleinerung der Länge hervorruft, bis L(t) den Wert Null annimmt. Ein solches Modellverhalten kann wohl in empirischen Untersuchungen nicht bestätigt werden, d.h., das Modell ist nicht für die Beschreibung der Wirklichkeit (des entsprechenden Systems) geeignet. Das kann auch in Abbildung 17 (Frequenzzunahme um 500) bestätigt werden.

3.4. Indirekte Abhängigkeiten

Köhler untersucht auch – ausgehend von den Gleichungen (2)–(5) – indirekte Abhängigkeiten; z.B. folgt für Köhler aus

$$PL = a_1 L^{b_1} \qquad \text{und} \qquad L = a_4 F^{b_4}$$

die Abhängigkeit

$$PL = a_5 F^{b_5} .$$

Ein solches Vorgehen – obwohl formal völlig korrekt – kann nicht zu exakten Ergebnissen führen, da bisher noch nicht bewiesen wurde, welche der jeweils verwendeten Größen als abhängige und welche als unabhängige Variable anzusetzen ist. In Kategorien der Systemtheorie ist ein System nicht möglich, in dem die abhängige Variable die unabhängige Variable steuert, wenn keine Rückkopplung besteht. In allen von Köhler

untersuchten Abhängigkeiten wurde eine solche Rückkopplung nicht modelliert (in drei Fällen wurde die Rückkopplung auch nicht im Blockschaltbild berücksichtigt).

Das Vorgehen von Köhler muß im Endeffekt dazu führen, daß eine Variable gleichzeitig abhängige und unabhängige Variable ist, was außer seiner formalen Inadäquatheit auch keinen systemtheoretischen Erkenntniswert hat. Diesen Fall zeigt Abbildung 19.

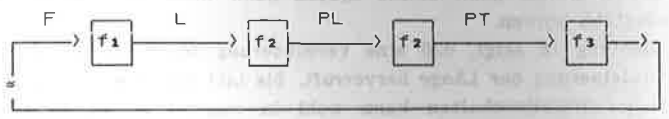


Abb. 19. Indirekte Zusammenhänge nach Köhler

In der Systemtheorie würde man diese indirekten Abhängigkeiten nach dem in Abb. 20 dargestellten Schema untersuchen. Erstens müßten die Abhängigkeiten jeder der vier untersuchten Eigenschaften (F(t), L(t), PL(t), PT(t)) von der Zeit bekannt sein. Wollte man dann die Abhängigkeit zwischen PL und F untersuchen, so müßte man neben F als Eingangsgröße auch zusätzlich L und eventuell auch PT berücksichtigen. Dies gilt aber nur, wenn vorausgesetzt wird, daß PL eine von den anderen Variablen abhängige Variable ist. Falls sich das als falsch erweist, muß man ein entsprechendes anderes Modell bauen, wobei jedoch dann, wenn man indirekte Abhängigkeiten untersuchen will, stets mehrere Eingangsgrößen zu berücksichtigen hat.

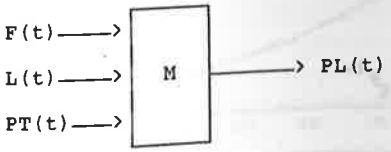


Abb. 20. Systemtheoretische Modellierung der Abhängigkeit zwischen PL(t) und F(t), L(t), PT(t)

Solche Untersuchungen setzen eine völlig andere empirische Datenerfassung voraus (im Vergleich zu den Untersuchungen Köhlers), da für jede lexikalische Einheit gleichzeitig empirische Werte für alle vier untersuchten Eigenschaften (und das außerdem in Abhängigkeit von der Zeit) berücksichtigt werden müssen.

Aufgrund der Kompliziertheit des lexikalischen Systems, dessen Subsysteme (z.B. Lexik der Fachsprachen) sich mit unterschiedlicher Dynamik entwickeln, wäre eine getrennte Untersuchung dieser Subsysteme angebracht. Die Anwendung der Methoden der Systemtheorie erlaubt es dann, aus diesen Submodellen ein komplexes Modell der Dynamik der Lexik natürlicher Sprachen zu erarbeiten.

4. Zusammenfassung

Die Notwendigkeit einer exakten Beschreibung und Interpretation des in empirischen Untersuchungen ermittelten Datenmaterials gilt nicht nur für Naturwissenschaften, sondern für alle wissenschaftlichen Untersuchungen allgemein. Auch in sprachwissenschaftlichen Untersuchungen wurden schon vor mehreren Jahrzehnten Proben einer mathematischen Beschreibung bestimmter Aspekte des jeweiligen empirischen Materials vorgenommen (z.B. durch die Ableitung bestimmter Parameter und deren Berechnungsverfahren). Später begann man damit, die Abhängigkeiten zwischen bestimmten sprachlichen Eigenschaften mit mathematischen Funktionen zu beschreiben, die jedoch nicht aus theoretischen Überlegungen bzw. aus theoretischen Modellen abgeleitet wurden. Dies war erst in den letzten Jahren möglich, wo man deduktiv einzelne Sprachgesetze gefunden hat. Einige dieser Gesetze (z.B. das Menzerathsche Gesetz) wurden in empirischen Untersuchungen mehrerer Sprachen bestätigt. Hier handelt es sich aber um Sprachgesetze, die isoliert betrachtet wurden, d.h. nicht in ihren gegenseitigen Abhängigkeiten. Erst Köhler (1986) hat den Versuch unternommen, unter Anwendung von bestimmten systemtheoretischen Elementen ein Modell aufzubauen, welches die Abhängigkeit zwischen mehreren lexikalischen Eigenschaften beschreiben will. Diese Arbeit brachte somit einen wesentlichen Erkenntnisfortschritt, wobei jedoch die verwendeten Methoden eine exakte Lösung des gestellten Problems unmöglich machten. Besonders auf diesen Aspekt wollten wir in unserem Artikel hinweisen. Eine exakte Modellierung von einzelnen und auch mehreren srpachlichen Erscheinungen bedarf einer exakten und umfassenden Anwendung system-

theoretischer Methoden und Verfahren. Die Grundvoraussetzung für deren Anwendung besteht jedoch darin, daß die zu untersuchenden Erscheinungen in Abhängigkeit von der Zeit gesehen werden, d.h. als diachronische Erscheinungen. Diese Bedingung war jedoch in den Untersuchungen von Köhler nicht erfüllt. In der Systemtheorie ist eine synchrone Betrachtung der zu untersuchenden Erscheinungen stets ein Spezialfall einer vorhergehenden diachronischen Betrachtung und nicht umgekehrt. Dies ist wohl auch leicht verständlich, da nicht jedes synchrone Sprachgesetz gleichzeitig im diachronischen Sinne Gesetzescharakter tragen muß.

Die Anwendung systemtheoretischer Methoden in sprachwissenschaftlichen Untersuchungen wird zur Zeit vor allem dadurch erschwert, daß keine ausreichenden diachronischen empirischen Daten vorliegen und deren Gewinnung – abgesehen von dem enormen notwendigen Arbeitsaufwand – oft nur sehr schwer und in bestimmten Subsystemen der Sprache möglich ist. Diese Arbeiten müssen jedoch aufgenommen werden, da davon der weitere Fortschritt der synergetischen Modellierung natürlicher Sprachen abhängt.

Anhand einiger Beispiele, die lediglich zur reinen Demonstration des Aufbaus und der Anwednung eines systemtheoretischen Lexikmodells in Simulationsuntersuchungen dienen sollten, konnte gezeigt werden, daß man gewisse Rückschlüsse über die Richtung der Abhängigkeiten zwischen den untersuchten Variablen gewinnen kann. Außerdem wurde deutlich, daß auch sprachliche Untersuchungen experimentelle Untersuchungen sein können, wie wir es aus den Natur- und technischen Wissenschaften schon kennen. Dies eröffnet völlig neue Dimensionen sprachwissenschaftlicher Forschungen und der Anwendung von Methoden, die in den Natur- und technischen Wissenschaften ausgearbeitet wurden. Sprachwissenschaftliche Untersuchungen, die in diese Richtung gehen, leisten einen großen Beitrag dazu, die künstlich zwischen Sprachwissenschaft (als Geisteswissenschaft) und den Natur- und technischen Wissenschaften gezogene Grenze zu überwinden.

Literatur

Altmann, G. (1983), Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung.* Göttingen, Herodot, 54-90.

Isermann, R. (1971), Experimentelle Analyse der Dynamik von Regelstrekken. Mannheim, Hochschulfachschulbücher-Verlag.

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

Philippow, E. (Hrsg.) (1977), Taschenbuch Elektronik. Band 2. Grundlagen der Informationstechnik. Berlin, VEB Verlag Technik.

Reinisch, K. (1974), Kybernetische Grundlagen und Beschreibung kontinuierlicher Systeme. Berlin, Verlag Technik.

Thoma, M. (1971), Theorie linearer Regelungssysteme. Braunschweig, Vieweg.

Hammerl, R. (ed.), Glottometrika 10, 1988.

Transformierte sprachtypologische Profilvektoren

Peter Anreiter, Innsbruck

1. Ein sprachtypologischer Profilvektor kann als systematische, geordnete Zusammenfassung von artgleichen typologischen Daten, die durch verschiedene mathematische Operationen gewonnen wurden, definiert werden. Solche Vektoren sind zwar an sich absolute Größen, d.h. sie sind nicht referierbar auf genormte Einheitsvektoren, erlangen aber ihre wissenschaftliche Verwertbarkeit und Aussagekraft erst durch den Vergleich mit anderen strukturgleichen Vektoren anderer Sprachen bzw. Sprachsysteme. Mit anderen Worten: Ein Vektor[1)] über einer bestimmten Menge von sprachlichen Daten einer bestimmten Sprache charakterisiert und beschreibt einen gewissen Teilbereich eben dieser Sprache exakt für sich, besitzt aber, wenn er isoliert betrachtet wird, nur geringen Erkenntniswert; diesen erhält er erst durch weitere Vergleiche. Wichtig ist ferner der Umstand, daß solche Profilvektoren keine Mengen (im mengentheoretischen Sinne) sind; der Unterschied zwischen beiden "Ansammlungen" besteht darin, daß ein Vektor eine vorher festgelegte, streng geregelte Abfolge und Anordnung der Elemente aufweist und daß gleiche Elemente mehrfach aufscheinen können (wie es gerade bei den Assoziativitätsvektoren über der Phonemmenge einer Sprache der Fall ist) und als solche vermerkt werden müssen.

Diese Profilvektoren sind wertvolle Instrumente für sprachtypologische Untersuchungen; ihre Bedeutung wird leider vielfach noch unterschätzt.

2. Beobachtungen haben ergeben, daß Profilvektoren in der Regel aus Bruchzahlen zusammengesetzt sind, wobei die Variation sich auf die Zähler erstreckt bei gleichzeitiger Konstanz des Nenners. Sei also ein beliebiger Nenner mit N symbolisiert und die einzelnen Zähler mit Z_1, Z_2, ..., Z_n bezeichnet, so ergibt sich als allgemeine Form eines beliebigen Profilvektors:

$$PV = \left[\frac{Z_1}{N}, \frac{Z_2}{N}, \ldots, \frac{Z_n}{N}\right]$$

bzw. $$\frac{1}{N}\left[Z_1, Z_2, \ldots, Z_n\right],$$

wobei der Vollständigkeit halber bemerkt werden muß, daß ohne weiteres ein $Z_i = Z_j = Z_k$ usf. sein kann oder auch nicht.

3. Transformation sprachtypologischer Profilvektoren in analytisch-geometrische Vektoren

Man kann nun den Versuch wagen, sprachtypologische Profilvektoren in geometrische Vektoren zu transformieren und diese auf analytisch-geometrische Weise darzustellen. Dies geschieht am besten dadurch, daß man den Nenner als Maßzahl übernimmt und die einzelnen Zähler aufsummiert. Die Zählersumme repräsentiert dabei die erste Komponente, der Nenner die zweite Komponente des so entstehenden Vektors:

$$\underset{\text{typol}}{PV} \xrightarrow{\text{transform.}} \underset{\text{geom}}{V}$$

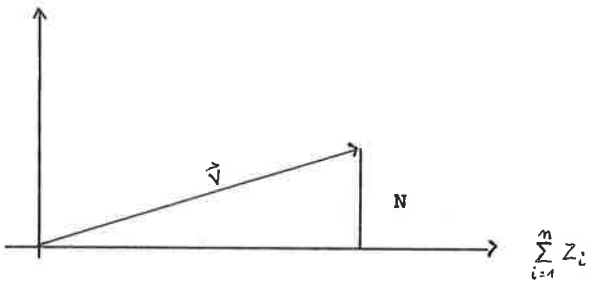
oder:

$$PV \xrightarrow{\text{transform.}} \vec{V}$$

mithin:

$$\frac{1}{N}\left[Z_1, Z_2, \ldots, Z_n\right] \xrightarrow{\text{transform.}} \begin{pmatrix} \sum_{i=1}^{n} Z_i \\ N \end{pmatrix}$$

Der Vorteil dieser Transformation beruht nun auf der Möglichkeit, den Ausgangsvektor auch graphisch innerhalb eines zugrundegelegten Koordinatensystems darstellen zu können, wodurch eine gewisse Anschaulichkeit erzielt wird:



4. Goniometrische und zyklometrische Interpretation vektorieller Daten

Bekannlich schließt jeder Vektor mit der Abzisse einen bestimmten Winkel δ ein, der für die Weiterführung unseres Problems von hoher Bedeutung ist. Man kann nämlich die beiden Vektorenkomponenten (die gewisse sprachtypologische Daten repräsentieren) in die Abhängigkeit des Winkels δ bringen, und zwar nach den bekannten Gleichungen

$$\tan \delta = \frac{N}{\sum_{i=1}^{n} Z_i} \Longrightarrow \delta = \arctan \frac{N}{\sum_{i=1}^{n} Z_i} .$$

Den in der Mathematik nicht definierten Fall, nämlich

$$\sum_{i=1}^{n} Z_i = 0,$$

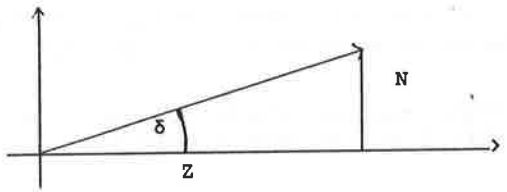
müssen wir ausschließen.

Auch der Fall, daß tan δ-->∞, also δ = 90°, muß als rein hypothetische Trivialität exkludiert werden.

Ferner ist darauf zu verweisen, daß wir natürlich auch mit anderen Winkelfunktionen operieren könnten, und zwar dann, wenn der Betrag des Vektors |V| und eine Kathete bekannt wären, also etwa

$$\sin \delta = \frac{N}{|V|} \Longrightarrow \delta = \arcsin \frac{N}{|V|} \quad \text{für } |V| \neq 0.$$

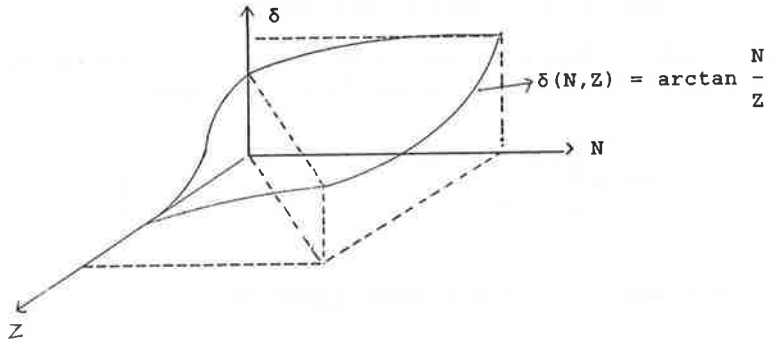
Graphisch ließen sich diese Beziehungen nun folgendermaßen ausdrücken:



Die Darstellungsweise des geometrischen Vektors im Koordinatensystem des zweidimensionalen Raumes ist korrekt, wenn wir für die Zählersumme einerseits und die Maßzahl des Nenners andererseits konkrete Zahlenwerte ermittelt haben. Wenn wir aber die zyklometrische Funktion

$$\delta = \arctan \frac{N}{Z}$$

ganz allgemein betrachten, verschiebt sich unser Problem gewissermaßen auf eine höhere Ebene. Denn der eingeschlossene Winkel δ variiert nicht nur durch die Variation von N bei konstantem Z, sondern genauso durch die Veränderung von Z bei gleichbleibendem N. Wenn also Z und N als Variable betrachtet werden, deren Werte sich auch gegenläufig verändern können, ist δ eine Funktion von N und Z. Bei der graphischen Darstellung dieser Relationen müssen wir uns demnach im dreidimensionalen Raum bewegen, wobei der Funktionsgraph durch sogenannte Schichtebenen (die in der folgenden Zeichnung der Anschaulichkeit halber nicht dargestellt sind) erzeugt wird:



5. Konkrete Beispiele

Aus der Fülle der theoretisch möglichen Profilvektoren seien nun drei charakteristische und für unsere Zwecke verwertbare herausgenommen:

5.1. Der "phonologische Profilvektor" über der Menge der distinktiven Merkmale

Es sei hier nicht beabsichtigt, eine Definition der distinktiven Merkmale zu geben; wichtig für uns ist es zu wissen, daß ein Phonem selbst als Menge der distinktiven Merkmale aufgefaßt (und somit in Matrizenform und/oder Mengengraphen dargestellt) werden kann. Wenn auch in diesem Rahmen auf eine "linguistische Verbaldefinition" verzichtet wird, so soll doch die Möglichkeit einer "mathematischen Abgrenzung" der distinktiven Merkmale hervorgehoben werden: Sei p ein beliebiges Phonem einer Sprache S, sei ferner $M_x(p)$ ein Phonem p, das durch das Merkmal M_x charakterisiert ist, sei ferner $\{M_x(p)\}$ die Menge aller Phoneme mit dem Merkmal M_x und sei endlich $\mu\{M_x(p)\}$ die Mächtigkeit dieser Menge, so läßt sich unter dem Gesichtspunkt, daß das Verhältnis zwischen der Anzahl der Phoneme mit dem Merkmal M_x und dem distinktiven Merkmal selbst konstant ist, die Ausprägung eines distinktiven Merkmals DM_x wie folgt darstellen:

$$DM_x = \frac{\mu\{M_x(p)\}}{|P|},$$

wobei $|P|$ die Anzahl der Elemente des zugrundeliegenden Phonemsystems der zu untersuchenden Sprache S symbolisieren soll.

Der gesamte Profilvektor (in welchem nun die Reihenfolge der einzelnen Elemente genormt sein muß) hätte nun folgendes Aussehen:

$$DM = \left[\frac{\mu\{M_1(p)\}}{|P|}, \frac{\mu\{M_2(p)\}}{|P|}, \ldots, \frac{\mu\{M_n(p)\}}{|P|}\right]$$

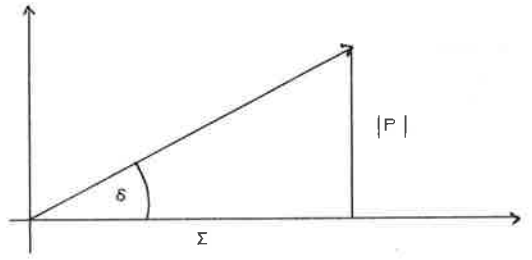
oder nach Herausheben des gemeinsamen Quotienten:

$$DM = \frac{1}{|P|}\left[\mu\{M_1(p)\}, \mu\{M_2(p)\}, \ldots, \mu\{M_n(p)\}\right].$$

Dieser Vektor ist wissenschaftlich nur dann sinnvoll verwertbar, wenn – wie schon oben erwähnt – die Reihenfolge der Einzelkomponenten verbindlich festgelegt wird, wenn also die Anordnung der distinktiven Merkmale unabhängig von der Art und Mächtigkeit des jeweiligen Phonemsystems normiert ist; wäre dies nicht der Fall, könnte man keine sinnvollen Vergleiche zwischen den Profilvektoren verschiedener Sprachen erstellen. Bei der Transformation dieses Vektors in einen analytisch-geometrischen des zweidimensionalen Raumes ist natürlich, da das Wesen der Transformation in der Summation der Zähler besteht und weiters bezüglich der Addition die Strukturgesetze der Kommutativität und Assoziativität[2] gelten, die Beachtung dieser strengen Anordnung hinfällig geworden: Die Summe der Zähler eines phonologischen Profilvektors einer bestimmten Sprache ergibt eine gewisse natürliche Zahl, der Nenner ist durch den Umfang des Phonomsystems von vornherein schon festgelegt; diese beiden Zahlen bilden nun die Komponenten des transformierten Vektors:

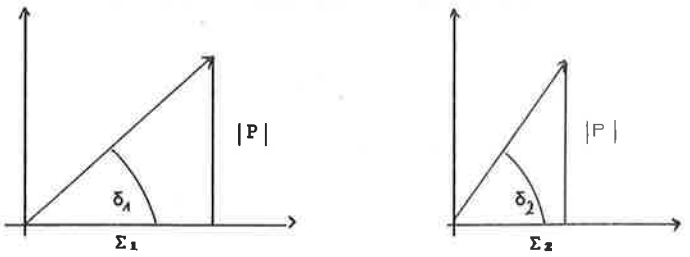
$$DM \xrightarrow{\text{transform.}} \overrightarrow{DM}\begin{pmatrix} \sum_{x=1}^{n} \mu\{M_x(p)\} \\ |P| \end{pmatrix}.$$

Im kartesischen Koordinatensystem läßt sich dieser transformierte Vektor auf folgende Weise darstellen:



Umfangreiche Stichproben in natürlichen und künstlich erstellten Phonemsystemen haben nun ergeben, daß sich der Winkel, den der Vektor mit

der Abszisse einschließt, zwischen 70 und 80 Grad bewegt. Dies gilt allerdings nur unter der Bedingung, daß wir das von Altmann-Lehfeldt (1973,75) aufgestellte Inventar von 10 distinktiven Merkmalen zugrunde legen, und zwar {voc, cons, ant, cor, nas, cont, vioced, high, low, back}. Hierbei wird genau dem Merkmalsystem gefolgt, das die beiden Gelehrten zur Klassifikation ihrer Phonemmenge {i,e,a,o,u,l,r,p,b,m,t,d,n,s,z,k,g,ŋ,w,y} benützen. Erst wenn wir diese Merkmalmenge für weitere Untersuchungen beibehalten, wenn wir es "normieren", erhält der so wichtige Winkel δ wissenschaftliche Relevanz und Aussagekraft, und zwar hinsichtlich seiner relativen Funktion auf andere Winkel anderer Profilvektoren. Es ist nun zu beachten, daß sich die Größe des Winkels nicht direkt proportional zum Umfang des zugrundeliegenden Phonemsystems verhalten *muß*. Die Winkelgröße ist sowohl von der Anzahl der Phoneme des Gesamtsystems als auch von der artikulatorischen Gestaltung eben dieser Phoneme, die nach dem genannten Merkmalschema untersucht werden, abhängig. Und dabei gilt weiters der Satz, daß bei konstanter Phonem*anzahl* der Winkel δ umso kleiner wird, je größer der numerische Wert von $\Sigma\mu\{M_x(p)\}$ ist:



$(\Sigma_1 > \Sigma_2 \dashrightarrow \delta_1 < \delta_2$, wenn $|P| = const.)$

Für die sprachtypologische Betrachtung von Phonemsystemen ist also nicht die *Länge* des Profilvektors, sondern die *Winkelgröße* die maßgebliche Variable. Sie läßt sich, wie erwähnt, zyklometrisch als

$$\delta = \operatorname{arc\,cot} \frac{\sum_{x=1}^{n} \mu\{M_x(p)\}}{|P|}$$

berechnen.

Da nun δ eine Funktion über dieser Summe Σ einerseits und über |P| andererseits ist, besitzt ein angegebener Winkel für sich allein noch keine Aussagekraft. Denn der Winkel kann bei konstantem |P| und variabler Summe Σ einen gewissen Wert annehmen, der a priori identisch sein kann dem Wert, der bei konstanter Summe Σ und variierendem |P| entsteht. Das heißt mit anderen Worten: Man kann aus einem vorgegebenen Winkel δ nicht ad hoc auf die Mächtigkeit der Menge der distinktiven Merkmale schließen – genauso wenig wie auf den Umfang des Phonemsystems. Eine Verringerung der einen Menge kann durch entsprechende Vergrößerung der anderen Menge ausgeglichen werden. Daher ist es sinnvoll, bei konkreten Aufgabenstellungen neben dem Winkel δ auch z.B. den Umfang des Phonemsystems anzugeben.

Bei allen goniometrischen Operationen verdient schließlich die Betrachtung von Rand-, Extrem- und Trivialfällen großes Interesse. Nehmen wir z.B. als Gedankenexperiment folgendes an: Es existiere zu einer Sprache ein Phonemsystem, das mittels unseres normierten Inventars von distinktiven Merkmalen kaum beschreibbar ist. Das würde bedeuten, daß unsere Summe Σ sehr klein wäre. Dadurch steigt aber – wie wir schon angedeutet haben – die Winkelgröße an. Im Extremfall würde δ nun den Wert von 90 Grad annehmen, wofür ein Tangenswert nicht definiert ist. Es führt also letztlich die Annahme, daß es zu einem gegebenen Phonemsystem keine Merkmalsmatrix gibt, die dieses Phonemsystem beschreiben könnte, zu einem naturgemäß rein hypothetischen Ergebnis:

$$\Sigma \dashrightarrow 0 \Longrightarrow \delta \dashrightarrow 90° \text{ (für } |P| = const.)$$

Der zweite Trivialfall, den wir ins Auge fassen wollen, wäre dann gegeben, wenn wir zwar eine Merkmalsmatrix hätten, jedoch kein Phonemsystem, das es zu beschreiben gilt. Dann sinkt die Winkelgröße auf 0 Grad ab.

5.2. Der "Assoziativitätsvektor" AV über einer gegebenen Phonemmenge P

Unter *Phonemassoziativität* verstehe man das gegenseitige Verhalten von Elementen eines gegebenen endlichen Phonemsystems innerhalb ihrer Distribution. Sei x ein beliebiges Phonem eines Phonemsystems P, seien ferner Ax die Menge aller Phoneme, die dem x unmittelbar vorausgehen können, und xB die Menge aller Phoneme, die dem x unmittelbar nachfolgen können, so bezeichne man |Ax| als das Maß der *Attraktivität* von

x und |xB| als das Maß der *Aggressivität* von x. Auf Grund der aus rein statistischen Erwägungen gerechtfertigten Annahme, daß die Wahrscheinlichkeit, daß eine bestimmte Phonemkombination xy auftaucht, a priori gleich groß ist wie diejenige, daß dies nicht der Fall ist, und aufbauend auf der Erkenntnis, daß die Wahrscheinlichkeit, daß ein Phonem x exakt |Ax| Vorgänger (bzw. |xB| Nachfolger) besitzt, gleich

$$P(X=|AX|) = \binom{n}{|AX|} p^{|Ax|} q^{n-|Ax|}$$

bzw.
$$P(X=|xB|) = \binom{n}{|xB|} p^{|xB|} q^{n-|xB|}$$

ist, ist es angebracht, mit Altmann-Lehfeldt (1973,98) um den Erwartungswert der betreffenden Zufallsvariablen einen 95% starken Konfidenzstreifen, dessen jeweilige Breite sich natürlich nach der Mächtigkeit der zugrunde liegenden Phonemmenge richtet, zu legen, wobei uns nun die beiden Randzahlen k_1 und k_2 des Intervalls Aufschlüsse über die Maße und Submaße der Assoziativität geben. Kommt nämlich die Maßzahl von |Ax| bzw |xB| zwischen 0 und k_1 zu liegen, bezeichnet man das Phonem x als *nichtattraktiv* (NAT) bzw. *nichtaggressiv* (NAG); Zahlen im Intervall $\langle k_2, P\rangle$ beziehen sich auf *attraktive* (AT) bzw. *aggressive* (AG) Phoneme. Die Mehrzahl der Fälle liegt nach statistischen Beobachtungen zwischen k_1 und k_2 und wird als *semiattraktiv* (SAT) bzw. *semiaggressiv* (SAG) bezeichnet.

Da wir es nun mit sechs "Assoziativitätsmerkmalen" zu tun haben, lassen sich durch wechselseitige Intersektionen neun "Assoziativitätsmengen" konstruieren, nämlich

(1) AG ∩ AT
(2) AG ∩ SAT
(3) AG ∩ NAT
(4) SAG ∩ AT
(5) SAG ∩ SAT
(6) SAG ∩ NAT
(7) NAG ∩ AT
(8) NAG ∩ SAT
(9) NAG ∩ NAT.

Für die Bildung von Profilvektoren ist es aber nicht so sehr von Wichtigkeit, *welche* Elemente in den einzelnen Durchschnittsmengen stehen, sondern wie groß ihre *Anzahl* ist. Wir legen mithin unser Augenmerk auf die Mächtigkeit der einzelnen Intersektionen, da ja nur diese arithmetisch/statistisch (und somit allgemein typologisch) verwertbar ist. Der Assoziativitätsvektor ist – wie übrigens die anderen Profilvektoren auch – eine geordnete Zusammenfassung von Bruchzahlen, wobei der Nenner identisch ist mit der Mächtigkeit der zugrunde liegenden Phonemmenge und die einzelnen Zähler die Mächtigkeit der Intersektion der Einzelassoziativitäten repräsentieren (wobei nun μ{X} bedeutet: die Mächtigkeit von {X}).

$$AV = \left[\frac{\mu\{AG \cap AT\}}{|P|}, \frac{\mu\{AG \cap SAT\}}{|P|}, \frac{\mu\{AG \cap NAT\}}{|P|}, \ldots, \frac{\mu\{NAG \cap NAT\}}{|P|}\right]$$

Der Einfachheit halber substituieren wir μ{AG ∩ AT} durch z_1, μ{AG∩SAT} durch z_2 usw. bis μ{NAG ∩ NAT} durch z_9. Dann erhält ein allgemeiner Assoziativitätsvektor folgendes Aussehen:

$$AV = \left[\frac{z_1}{|P|}, \frac{z_2}{|P|}, \ldots, \frac{z_9}{|P|}\right]$$

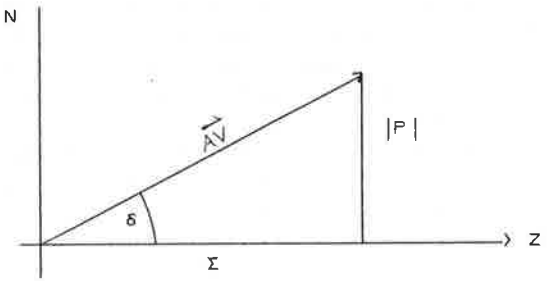
bzw.
$$AV = \frac{1}{|P|}\left[z_1, z_2, \ldots, z_9\right].$$

Der transformierte Vektor besitzt nun als erste Komponente die Summe der einzelnen z-Werte und als zweite Komponente die Maßzahl von P:

$$AV \xrightarrow{\text{transform.}} \vec{AV}$$

$$\frac{1}{|P|}\left[z_1, z_2, \ldots, z_9\right] \xrightarrow{\text{transform.}} \begin{pmatrix} \sum_{x=1}^{9} z_i \\ |P| \end{pmatrix}$$

Somit:



Der durch die Komponenten aufgespannte Assoziativitätsvektor schließt nun mit der z-Achse einen Winkel ein, der "Assoziativitätswinkel" genannt werden soll. Dieser wird berechnet als

$$\delta = \arctan \frac{|P|}{\Sigma z}$$

5.3. Zyklometrisch-goniometrische Interpretation von Phonemhäufigkeiten

Bekanntlich gibt es viele Methoden, um Vokal- oder Konsonantenhäufigkeiten zu ermitteln; erinnert sei nur etwa an die Indizes von Nikonov, Isačenko, Krámský u.a. G. Altmann hat ferner gezeigt, daß sich alle Indizes aus einer einzigen Formel generieren lassen (vgl. Altmann 1971, 189-197). In diesem Abschnitt soll gezeigt werden, daß man das Phänomen der Vokalhaltigkeit einer Sprache auch goniometrisch interpretieren kann.

Gegeben sei ein Textkorpus von hinreichend großem Umfang, damit entsprechende Rückschlüsse gezogen werden können. Es soll die Vokalhaltigkeit dieses Textes ermittelt werden. Dabei werden als Betrachtungseinheiten die einzelnen Wörter herangezogen. Der Index der Vokalhaltigkeit (VH) werde nun durch den Bruch

$$VH_x = \frac{\text{Frequenz } f_x \text{ eines bestimmten Vokals } x}{\text{Summe aller Wörter}}$$

oder kürzer: $VH_x = \frac{f_x}{\Sigma W}$

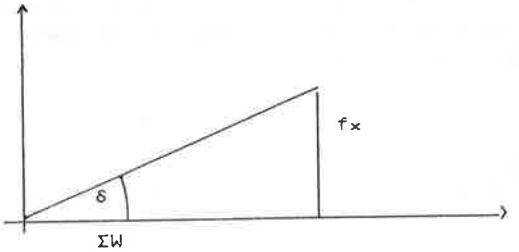
repräsentiert.

Dieser Index wirft nun zwei Probleme auf: (1) Was ist ein "Wort"? (2) Können Einheiten verschiedener Ebenen überhaupt in eine derartige Relation gesetzt werden?

Ad (1): Die Diskussion um eine befriedigende Erklärung, was ein "Wort" ist, hat in der Linguistik noch zu keinem befriedigenden Ergebnis geführt. Wir werden uns sicherlich schwer tun, diesen Index auf polysynthetische oder gewisse agglutinierende Sprachen anzuwenden. Wenn wir aber als Untersuchungskorpora Texte indogermanischer Sprachen heranziehen, in welchen sich ein "Wort" als "...kleinste, durch Wortakzent und Grenzsignale ... theoretisch isolierbare Lautsegmente, die auf ... orthographisch-graphemischer Ebene durch Leerstellen im Schriftbild isoliert werden" (vgl. Bußmann 1983, 585) definieren läßt, können wir auf eine theoretische Diskussion über mögliche andere Aspekte des "Wortes" verzichten.

Ad (2): Mir will scheinen, daß die prinzipielle Frage der Vergleichbarkeit von Einheiten verschiedener grammatischer Ebenen sekundär ist gegenüber der Frage nach Sinnhaftigkeit und praktischer Verwertbarkeit. Sollte sich herausstellen, daß sich der Index in der Praxix nicht bewährt, wird er zu korrigieren sein. Dies müssen aber erst weitere Untersuchungen zeigen. Vorerst wird der Versuch unternommen, mit ihm zu operieren.

Die durch diesen Index errechneten Werte können graphisch nun recht gut veranschaulicht werden, indem man auf der Abszisse des Koordinatensystems die Summe aller Wörter aufträgt, auf der Ordinate die dazugehörige Frequenz fx des zu untersuchenden Vokals x:



Der dabei eingeschlossene Winkel δ wird nun als alleinige Größe bezüglich der Vokalhaltigkeit betrachtet. Der Umfang der jeweiligen Vokale verhält

sich dabei direkt proportional zur Winkelgröße. Auf die beiden Extremfälle, daß nämlich δ einerseits = 0 Grad ist und andererseits in einem begrenzten Korpus ein bestimmter Vokal x unendlich oft vorkommt, soll nicht näher eingegangen werden. Der Winkel selbst läßt sich berechnen als

$$\delta = \arctan \frac{f_x}{\Sigma W}$$

Über die Menge der untersuchten Vokale, also über $V = \{x_1, x_2, \ldots, x_n\}$, läßt sich nun ein sprachtypologischer Profilvektor erstellen:

$$VH_x = \left[\frac{f_1}{\Sigma W}, \frac{f_2}{\Sigma W}, \frac{f_3}{\Sigma W}, \ldots, \frac{f_n}{\Sigma W}\right]$$

bzw. $$VH_x = \frac{1}{\Sigma W}\left[f_1, f_2, f_3, \ldots, f_n\right] .$$

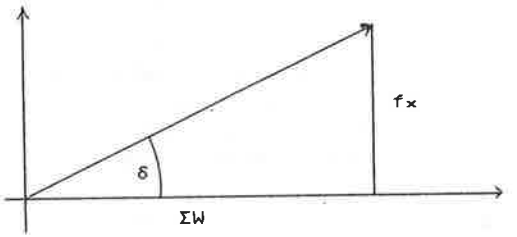
Dieser Vektor läßt sich nun in einen analytisch-geometrischen Vektor transformieren:

$$VH_x \xrightarrow{\text{transform.}} \overrightarrow{VH}_x \begin{pmatrix} f_x \\ \\ \Sigma W \end{pmatrix} .$$

Bevor wir diesen Vektor graphisch illustrieren und zyklometrisch interpretieren, ist es notwendig, die Komponenten zu vertauschen, um so die direkte Proportionalität zwischen Winkelgröße und Summe der zu betrachtenden Vokale zu erzeugen:

$$\overrightarrow{VH}_x \begin{pmatrix} f_x \\ \\ \Sigma W \end{pmatrix} \longrightarrow \overrightarrow{VH}_x \begin{pmatrix} \Sigma W \\ \\ f_x \end{pmatrix} .$$

Dadurch ergibt sich folgendes Bild:



Anmerkungen

1) Ein linguistischer Profilvektor ist u.a. gekennzeichnet durch eine endliche oder abzählbar unendliche Menge von Komponenten. Die Kerngrößen von Vektoren der ebenen bzw. dreidimensionalen Geometrie sind ebenfalls die Einzelkomponenten; die Bezeichnung "linguistischer Profilvektor" dürfte also aus dem Benennungsinstrumentarium der Mathematik stammen. Auf eine exakte mathematische Definition eines "Vektors" soll jedoch in diesem Rahmen verzichtet werden.

2) Kommutativgesetz bezüglich der Addition: x+y = y+x; Assoziativgesetz bezüglich der Addition: x+(y+z) = (x+y)+z.

Literatur

Altmann, G. (1971), Introduction to Quantitative Phonology. Bochum, Brockmeyer.

Altmann, G., Lehfeldt W. (1973), Allgemeine Sprachtypologie. München, Fink.

Bußmann, H. (1983), Lexikon der Sprachwissenschaft. Stuttgart.

Subjektive Ähnlichkeit deutscher Laute

Frank W. a Campo, Bochum/Köln
Slavko Geršić, Köln
Carl L. Naumann, Aachen
Gabriel Altmann, Bochum

1. Einführung

In einer früheren Arbeit haben Geršić, Naumann und Altmann (1985) die Problematik der subjektiven Lautähnlichkeit analysiert und vertraten die Ansicht, daß die objektive, auf linguistischer und physikalischer Begriffsbildung basierende Laut/Phonemähnlichkeit linguistisch dann als adäquat zu betrachten ist, wenn sie mit der subjektiven Ähnlichkeit übereinstimmt. Ein Vergleich der beiden Ähnlichkeiten wurde, soweit uns bekannt, noch in keiner Sprache durchgeführt. Um dies für das Deutsche möglich zu machen, muß die subjektive Lautähnlichkeit auf verschiedene Weisen gemessen und schrittweise mit den bekannten objektiven Messungen (vgl. Grimes, Agard 1959; Peterson, Harary 1961; Tolstaja 1968; Altmann 1969; Geršić 1971; Altmann, Lehfeldt 1972, 1980; Afendras, Tzannes, Trepanier 1973; Lindner 1975, 1980; Batóg, Steffen-Batogowa 1980; Lehfeldt 1980; Grotjahn 1980) verglichen werden.

Objektive Lauteigenschaften und daraus resultierende Ähnlichkeitsbewertungen ergeben sich aus diversen theoretischen Begriffsbildungen. Eigenschaften wie Formanten, Artikulationsart und -stelle, Distribution, Häufigkeit, distinktive Merkmale, Organverschiebungen u.a. spielen dabei eine entscheidende Rolle; werden sie einmal definiert, so kann man ihre Ausprägungen ohne Sprecherurteile ("objektiv") ermitteln. Die Frage, ob sie theoretisch relevant sind, d.h. ob man mit ihrer Hilfe eine phonetische/phonologische Theorie aufbauen kann, kann man positiv nur unter zwei Bedingungen beantworten:

(a) Sie müssen mit den Sprecherurteilen übereinstimmen, da subjektive Sprecherurteile gerade die Realität darstellen, an der wir unsere Theorie überprüfen müssen. Eine vollständige Übereinstimmung subjektiver und objektiver Lautähnlichkeit wird wohl nie erreicht, daher muß man zumindest nach einer Approximation trachten. Diese läßt sich leider nur durch trial-and-error bewerkstelligen, indem man die subjektive Ähnlichkeit durch verschiedene Experimentanordnungen und die objektive

durch verschiedene Ähnlichkeitsmerkmale ermittelt. In diesem Beitrag wird nur eine Experimentanordnung präsentiert.

(b) Auch wenn beide Ähnlichkeitsarten nahe Übereinstimmung aufweisen, haben sie eine theoretische Relevanz erst dann, wenn man sie für die Aufstellung von Gesetzen verwenden kann. Gesetze sind das Gerüst einer Theorie; ohne sie hat eine Theorie nur einen niedrigen gnoseologischen Status (Prototheorie), und für ihre Begriffsbildung fehlt der Nachweis, daß sie (theoretisch) fruchtbar ist. Dies ist die größte Schwäche z.B. moderner "Syntax-Theorien", die lediglich einen beschreibenden, jedoch keinen erklärenden Status haben.

2. Die Ermittlung subjektiver Lautähnlichkeiten

55 Versuchspersonen erhielten nach der Methode von Fillenbaum und Rapoport (1971) den folgenden Text und die phonetisch transkribierten Laute:

> *Sie haben eine Liste deutscher Laute bekommen (s. unten), die Ihnen auf Wunsch auch akustisch vorgeführt werden können. Suchen Sie die zwei Laute heraus, die Ihnen am ähnlichsten vorkommen, schreiben Sie sie nebeneinander oben auf das Ihnen ausgehändigte Blatt Papier, und verbinden Sie die Lautsymbole mit einem Strich. Sie können nun darunter weitere Laute schreiben, die den ersten ähnlich sind, und sie jeweils mit einem vertikalen Strich verbinden. Sie können aber auch, bevor Sie an dem ersten Bäumchen weiter arbeiten, aus der Lautliste weitere Lautpaare heraussuchen und neue Bäumchen anfangen, denen Sie jeweils wieder ähnliche Laute zuordnen. Wenn Sie Ähnlichkeiten zwischen den Lauten zweier Bäume feststellen, verbinden Sie diese Laute miteinander. Fahren Sie mit der Zuordnung der Laute zu den Bäumchen und der Verbindung der Bäumchen fort, bis Sie alle Laute untergebracht und alle Bäumchen miteinander verbunden haben.*

a	ɑ	b	c	d	e	ɛ	ø	œ	ə
f	ɟ	g	h	i	ɪ	j	k	l	m
n	ŋ	o	ɔ	p	r	ʀ	s	ɭ	ʒ
t	u	ʊ	v	x	ç	y	ʏ	z	ʔ

Als Grundlage für diese Messung wurden 40 deutsche Laute gewählt, wie sie unten aufgeführt sind. Einige von ihnen sind Varianten (c,ç,j), andere sind sprecherabhängig (r–R, a–ɑ). Ihre Einbeziehung verursacht jedoch keine Verzerrung des Resultats.

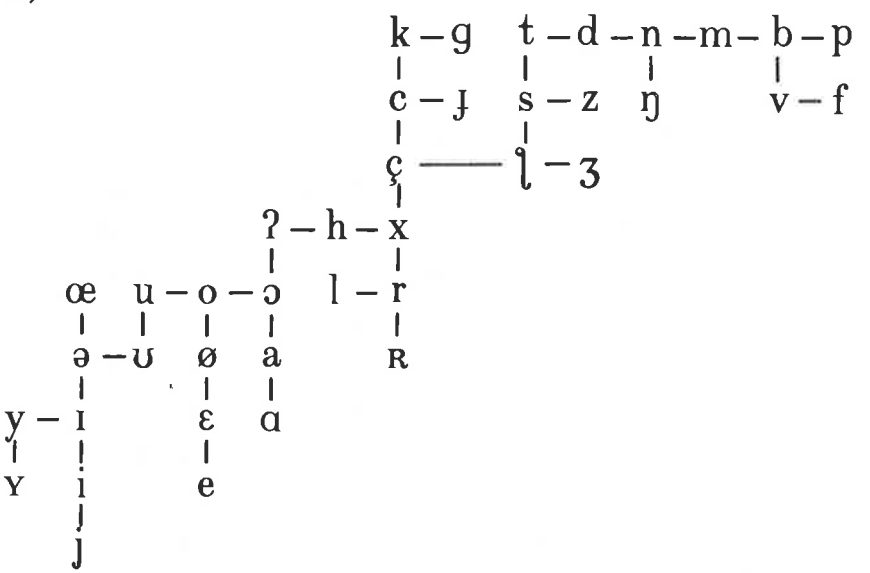
Die Vpn waren einige Studenten der Phonetik, Studenten anderer Fächer und Nichtstudenten mit unterschiedlicher Ausbildung. Den meisten mußten die Laute von einem Tonbandgerät vorgespielt werden, und der Versuchsleiter stand bei Nachfragen zur Verfügung. Nicht alle Vpn haben den letzten Satz der Anleitung befolgt, so daß mehrere nichtzusammenhängende Bäume entstanden. Zwei typische Fälle sind in der Abbildung 1a,b zu sehen. In den so entstandenen Graphen wurde die Distanz $d_{ij}(k)$ zwischen zwei Lauten i und j für die Vpn k als die minimale Anzahl der Kanten des sie verbindenden Weges gemessen (vgl. Batóg, Steffen-Batogowa 1980). Die Distanz zwischen den Lauten, die durch keinen Weg verbunden wurden, wurde im gegebenen Fall nicht in Betracht gezogen. In der Tat ist eine derartige Distanz unendlich, sie würde aber zu keinem vernünftigen Resultat führen, und daher wurde sie eliminiert.

Die endgültige Distanz zwischen den Lauten i und j wurde folgendermaßen ermittelt:

$$d_{ij} = \frac{1}{n} \sum_{k=1}^{n} d_{ij}(k), \qquad \begin{array}{l} i,j = 1,2,\ldots,40 \text{ (Laute)} \\ k = 1,2,\ldots,55 \text{ (Vpn)} \end{array}$$

d.h. als die durchschnittliche Distanz über alle n Vpn, die die Distanz zwischen i und j als endlich (d.h. durch einen zusammenhängenden Graph) dargestellt haben. Nicht immer war n = 55; d.h. n ≤ 55. Die so entstandenen Distanzen sind in der Tabelle 1 aufgeführt. Die Distanzmatrix ist symmetrisch.

a)



b)

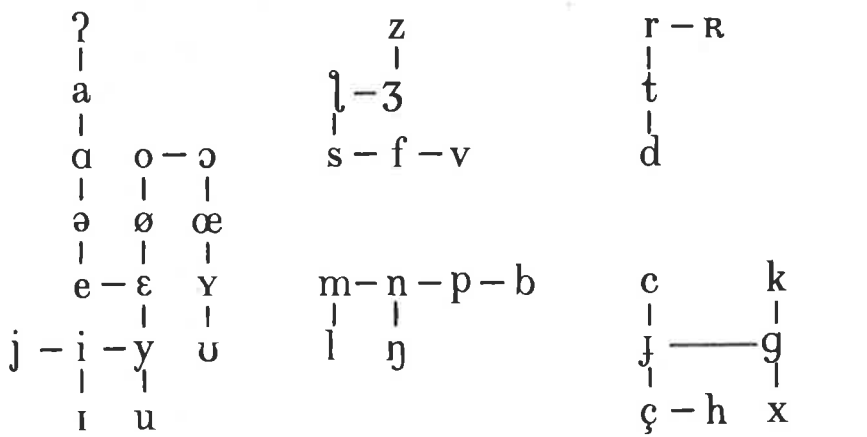


Abb. 1. Zwei typische Ähnlichkeitszuordnungen

Tabelle 1.
Subjektive Distanzen zwischen deutschen Lauten

	m	n	ŋ	r	ʀ	l	p	b
m	0.00	1.05	2.07	6.00	6.53	3.68	3.81	3.53
n	1.05	0.00	1.27	5.71	6.12	3.40	3.91	3.63
ŋ	2.07	1.27	0.00	6.24	6.53	4.24	4.03	3.69
r	6.00	5.71	6.24	0.00	1.22	1.98	5.93	6.07
ʀ	6.53	6.12	6.53	1.22	0.00	2.34	6.13	6.27
l	3.68	3.40	4.24	1.98	2.34	0.00	5.50	5.44
p	3.81	3.91	4.03	5.93	6.13	5.50	0.00	1.25
b	3.53	3.63	3.69	6.07	6.27	5.44	1.25	0.00
t	5.07	4.79	4.83	5.83	6.00	6.24	2.33	2.73
d	4.83	4.55	4.45	5.39	5.71	5.59	2.73	2.22
k	4.91	4.34	3.97	5.65	6.06	5.88	3.90	3.97
g	4.03	3.45	2.76	6.29	6.71	6.00	4.20	3.63
c	5.21	4.64	4.24	5.94	6.35	6.06	4.22	4.25
ʝ	4.67	4.09	3.45	6.12	6.53	6.18	4.68	4.17
ʔ	8.05	7.58	7.21	8.00	8.00	7.67	5.36	5.96
f	5.64	5.71	6.29	5.29	5.24	5.41	4.00	3.80
v	5.71	5.79	6.50	5.76	5.71	5.76	4.21	3.95
s	6.75	6.25	6.75	5.27	4.73	5.76	5.82	5.71
z	6.88	6.38	6.88	5.45	5.27	6.10	6.29	6.18
ʃ	6.87	6.40	6.80	4.59	4.41	5.14	6.13	6.06
ʒ	7.20	6.87	7.00	4.95	4.86	5.67	6.31	6.25
ç	7.54	7.00	7.46	3.63	3.48	4.29	6.73	7.07
x	7.53	7.07	7.07	2.79	2.63	3.87	6.38	6.47
h	8.13	7.67	7.67	3.88	3.74	4.38	6.44	6.65

	t	d	k	g	c	ʝ	ʔ	f
m	5.07	4.83	4.91	4.03	5.21	4.67	8.05	5.64
n	4.79	4.55	4.34	3.45	4.64	4.09	7.58	5.71
ŋ	4.83	4.45	3.97	2.76	4.24	3.45	7.21	6.29
r	5.83	5.39	5.65	6.29	5.94	6.12	8.00	5.29
ʀ	6.00	5.71	6.06	6.71	6.35	6.53	8.00	5.24
l	6.24	5.59	5.88	6.00	6.06	6.18	7.67	5.41
p	2.33	2.73	3.90	4.20	4.22	4.68	5.36	4.00
b	2.73	2.22	3.97	3.63	4.25	4.17	5.96	3.80
t	0.00	1.67	3.36	4.10	3.71	4.60	4.85	4.58
d	1.67	0.00	3.46	3.32	3.83	3.88	5.54	4.84
k	3.36	3.46	0.00	1.87	1.26	2.32	3.26	6.44
g	4.10	3.32	1.87	0.00	2.44	1.24	4.72	6.28
c	3.71	3.83	1.26	2.44	0.00	2.72	3.70	6.22
ʝ	4.60	3.88	2.32	1.24	2.72	0.00	5.13	6.50
ʔ	4.85	5.54	3.26	4.72	3.70	5.13	0.00	8.43
f	4.58	4.84	6.44	6.28	6.22	6.50	8.43	0.00
v	4.89	5.26	6.83	6.67	6.61	6.89	8.81	1.13
s	4.21	4.58	6.10	6.15	6.00	6.65	7.50	3.54
z	4.95	5.11	6.50	6.55	6.40	7.05	7.95	4.10
ʃ	4.78	5.00	5.95	6.05	5.79	6.53	7.24	3.50
ʒ	5.17	5.28	5.95	5.95	5.79	6.32	7.52	4.03
ç	5.87	5.71	4.79	5.67	4.63	5.89	6.25	4.56
x	5.41	5.44	4.20	5.14	4.40	5.57	5.64	4.68
h	5.47	5.61	4.14	5.14	4.19	5.50	4.59	5.85

	v	s	z	ʃ	ʒ	ç	x	h
m	5.71	6.75	6.88	6.87	7.20	7.54	7.53	8.13
n	5.79	6.25	6.38	6.40	6.87	7.00	7.07	7.67
ŋ	6.50	6.75	6.88	6.80	7.00	7.46	7.07	7.67
r	5.76	5.27	5.45	4.59	4.95	3.63	2.79	3.88
ʀ	5.71	4.73	5.27	4.41	4.86	3.48	2.63	3.74
l	5.76	5.76	6.10	5.14	5.67	4.29	3.87	4.38
p	4.21	5.82	6.29	6.13	6.31	6.73	6.38	6.44
b	3.95	5.71	6.18	6.06	6.25	7.07	6.47	6.65
t	4.89	4.21	4.95	4.78	5.17	5.87	5.41	5.47
d	5.26	4.58	5.11	5.00	5.28	5.71	5.44	5.61
k	6.83	6.10	6.50	5.95	5.95	4.79	4.20	4.14
g	6.67	6.15	6.55	6.05	5.95	5.67	5.14	5.14
c	6.61	6.00	6.40	5.79	5.79	4.63	4.40	4.19
ɟ	6.89	6.65	7.05	6.53	6.32	5.89	5.57	5.50
ʔ	8.81	7.50	7.95	7.24	7.52	6.25	5.64	4.59
f	1.13	3.54	4.10	3.50	4.03	4.56	4.68	5.85
v	0.00	3.74	4.26	3.84	4.37	4.85	5.24	6.35
s	3.74	0.00	1.31	1.62	2.35	3.19	3.76	4.70
z	4.26	1.31	0.00	2.02	2.25	3.67	4.24	5.20
ʃ	3.84	1.62	2.02	0.00	1.47	2.18	2.91	4.00
ʒ	4.37	2.35	2.25	1.47	0.00	2.87	3.40	4.42
ç	4.85	3.19	3.67	2.18	2.87	0.00	1.48	2.31
x	5.24	3.76	4.24	2.91	3.40	1.48	0.00	1.73
h	6.35	4.70	5.20	4.00	4.42	2.31	1.73	0.00

	i	ɪ	j	y	ʏ	e	ɛ	ə
m	9.30	9.70	9.18	11.60	12.00	12.30	13.00	12.10
n	8.90	9.30	8.91	11.20	11.60	11.90	12.60	11.70
ŋ	9.60	10.00	9.64	11.90	12.30	12.60	13.30	12.40
r	7.07	7.64	5.33	9.64	10.14	9.86	10.71	9.93
ʀ	6.93	7.50	5.22	9.50	10.00	9.71	10.57	9.79
l	7.00	7.29	5.18	9.54	9.64	10.00	10.85	9.43
p	8.50	8.90	8.15	10.80	11.20	11.50	12.20	11.30
b	8.60	9.00	8.00	10.90	11.30	11.60	12.30	11.40
t	8.30	8.70	6.93	10.60	11.00	11.30	12.00	11.10
d	9.30	9.70	6.64	11.60	12.00	12.30	13.00	12.10
k	8.17	8.50	5.75	10.25	10.58	11.50	12.08	11.33
g	7.75	8.08	5.50	9.83	10.17	11.08	11.67	10.92
c	8.00	8.33	5.63	10.08	10.42	11.33	11.92	11.17
ɟ	8.08	8.42	5.88	10.17	10.50	11.42	12.00	11.25
ʔ	7.33	7.31	6.42	8.79	8.65	5.80	6.16	5.44
f	7.65	8.35	5.83	9.94	10.12	10.65	11.47	11.25
v	7.65	8.35	5.91	9.94	10.12	10.65	11.47	11.19
s	6.33	6.89	4.68	9.17	9.28	9.61	10.22	10.12
z	6.72	7.28	5.00	9.56	9.67	10.00	10.61	10.47
ʃ	5.50	6.06	3.69	8.33	8.44	8.78	9.39	9.29
ʒ	5.72	6.28	3.38	8.56	8.67	9.00	9.61	9.59
ç	4.43	5.19	2.59	7.09	7.43	7.70	8.35	8.48
x	5.24	5.75	3.52	7.65	7.90	8.48	9.00	8.80
h	5.36	5.90	4.14	7.71	7.90	8.41	8.95	8.76

	ø	œ	a	ɑ	o	ɔ	u	ʊ
m	12.30	13.10	12.50	12.50	12.10	11.70	11.90	12.50
n	11.90	12.70	12.10	12.10	11.70	11.30	11.50	12.10
ŋ	12.60	13.40	12.80	12.80	12.40	12.00	12.20	12.80
ɾ	10.86	11.00	10.13	10.47	11.21	10.93	11.07	11.50
ʁ	10.71	10.86	10.07	10.40	11.07	10.79	10.93	11.36
l	10.21	10.36	10.36	10.64	11.23	11.00	10.92	11.46
p	11.50	12.30	11.70	11.70	11.30	10.90	11.10	11.70
b	11.60	12.40	11.80	11.80	11.40	11.00	11.20	11.80
t	11.30	12.10	11.50	11.50	11.10	10.70	10.90	11.50
d	12.30	13.10	12.50	12.50	12.10	11.70	11.90	12.50
k	11.33	11.83	12.00	12.09	11.33	11.00	10.92	11.58
g	10.92	11.42	11.45	11.55	10.92	10.58	10.50	11.17
c	11.17	11.67	11.73	11.82	11.17	10.83	10.75	11.42
ɟ	11.25	11.75	11.73	11.82	11.25	10.92	10.83	11.50
ʔ	6.88	7.42	6.38	6.88	7.40	7.04	8.42	8.96
f	11.31	11.69	11.33	11.40	11.06	11.19	10.00	10.63
v	11.25	11.63	11.20	11.27	11.00	11.13	9.94	10.56
s	10.24	10.76	10.63	10.88	11.00	10.71	10.41	10.94
z	10.59	11.12	10.88	11.13	11.35	11.06	10.76	11.29
ʃ	9.41	9.94	9.69	9.94	10.18	9.88	9.59	10.12
ʒ	9.71	10.24	9.94	10.19	10.47	10.18	9.88	10.41
ç	8.67	8.95	9.11	9.47	9.35	9.15	9.05	9.45
x	8.90	9.25	9.32	9.58	9.40	9.20	9.20	9.70
h	8.81	9.10	9.00	9.20	9.14	9.00	9.00	9.52

	i	ɾ	j	y	ʏ	e	ɛ	ə	ø	œ	a	ɑ	o	ɔ	u	ʊ
i	0.00	1.30	1.86	2.94	3.18	3.72	4.66	4.52	4.77	4.82	5.90	6.00	5.60	5.95	4.93	5.70
ɾ	1.30	0.00	2.64	3.24	3.15	4.31	4.89	4.43	4.78	4.62	6.02	6.07	5.88	5.91	5.23	5.75
j	1.86	2.64	0.00	4.63	4.93	5.26	6.24	6.11	6.74	6.69	7.12	7.27	7.24	7.53	6.37	7.06
y	2.94	3.24	4.63	0.00	1.40	4.70	5.20	5.05	4.20	4.31	6.02	6.14	4.81	5.43	3.62	4.41
ʏ	3.18	3.15	4.93	1.40	0.00	5.04	5.58	5.20	4.33	4.18	6.37	6.46	5.19	5.51	4.16	4.22
e	3.72	4.31	5.26	4.70	5.04	0.00	1.53	2.02	3.25	3.27	3.53	3.62	4.21	4.65	5.09	5.82
ɛ	4.66	4.89	6.24	5.20	5.58	1.53	0.00	1.86	2.81	2.75	3.02	3.23	3.94	4.13	5.20	5.93
ə	4.52	4.43	6.11	5.05	5.20	2.02	1.86	0.00	2.87	2.57	3.34	3.47	4.04	4.08	5.22	5.83
ø	4.77	4.78	6.74	4.20	4.33	3.25	2.81	2.87	0.00	1.71	3.94	4.10	2.92	3.16	4.04	4.78
œ	4.82	4.62	6.69	4.31	4.18	3.27	2.75	2.57	1.71	0.00	3.84	3.88	3.10	3.41	4.39	4.80
a	5.90	6.02	7.12	6.02	6.37	3.53	3.02	3.34	3.94	3.84	0.00	1.05	3.56	3.39	5.18	5.65
ɑ	6.00	6.07	7.27	6.14	6.46	3.62	3.23	3.47	4.10	3.88	1.05	0.00	3.62	3.39	5.14	5.50
o	5.60	5.88	7.24	4.81	5.19	4.21	3.94	4.04	2.92	3.10	3.56	3.62	0.00	1.39	3.00	3.00
ɔ	5.95	5.91	7.53	5.43	5.51	4.65	4.13	4.08	3.16	3.41	3.39	3.39	1.39	0.00	3.31	3.28
u	4.93	5.23	6.37	3.62	4.16	5.09	5.20	5.22	4.04	4.39	5.18	5.14	2.50	3.31	0.00	1.48
ʊ	5.70	5.75	7.06	4.41	4.22	5.82	5.93	5.83	4.78	4.80	5.65	5.50	3.00	3.28	1.48	0.00

3. Graphentheoretische Darstellung

Die von den Vpn ermittelten Graphen sind alle unterschiedlich, manche sind nicht zusammenhängend, stellenweise sieht man, daß sich einige Vpn von der optischen Ähnlichkeit zweier Symbole leiten ließen.

Aus den berechneten Distanzen lassen sich jedoch sowohl der Gesamtbaum (Minimalbaum), als auch die einzelnen Zusammenhangskomponenten leicht wiederherstellen. Ein derartiger Minimalbaum erfaßt dann die Wege der maximalen Lautähnlichkeit, wie sie für die Sprachgemeinschaft (anhand der gegebenen Stichprobe) gelten. Zu diesem Zweck benutzen wir den Algorithmus von Prim (1957; vgl. auch Bock 1974):

(1) Man wähle einen beliebigen Laut i.
(2) Man finde zu i den nächstliegenden Laut und verbinde sie mit einer Kante. Dies ist der Anfangsbaum.
(3) Zu diesem Baum adjungiere man den nächstgelegenen, im Graph noch nicht erfaßten Laut und die betreffende Kante.
(4) Man wiederhole Schritt (3) solange, bis alle Laute im Baum erfaßt sind.

Es ist gleichgültig, mit welchem Laut man anfängt. Dieser Algorithmus ist sehr leicht programmierbar. Den resultierenden Baum findet man in Abb. 2. Zahlreiche andere Konstruktionsmethoden findet man bei Bock (1974).

In Abb. 2 zeichneten wir die Kantenlängen frei und gaben nur die Abstände an. Man bekommt aus dem Computer das Bild nicht so wie oben dargestellt, aber die Nachbarschaften und die angegebenen Distanzen entsprechen den Berechnungen. Aus der Abbildung kann man "eine gewisse" Übereinstimmung mit der objektiven Klassifikation der Laute erkennen. Am interessantesten ist die Tatsache, daß [j] dem [i] näher steht als den Frikativen. Dies unterstützt eine Klassifikation des j als Glide (Wurzel 1970) und steht im Gegensatz zu der Auffassung, j sei als stimmhaftes Gegenstück zu ç konsonantisch (Wängler 1961; Meinhold, Stock 1980).

Ein ebenso übersichtliches Bild kann man auch durch eine hierarchische Klassifikation erhalten. In Abb. 3 findet man die Klassifikation mit der complete-linkage Methode. Hier wurde [j] direkt zu den Vokalen zugeordnet, sonst erhalten wir im Grunde das gleiche Bild. Die hierarchische Klassifikation zeigt, daß die Vpn insgesamt recht exakt nach verschiedenen Eigenschaften gleichzeitig urteilen. Die Tatsache, daß [?] von den velaren Verschlußlauten stark getrennt ist, kann an dem Aggregationskriterium liegen oder daran, daß der Sprecher/Hörer diesem Laut eine

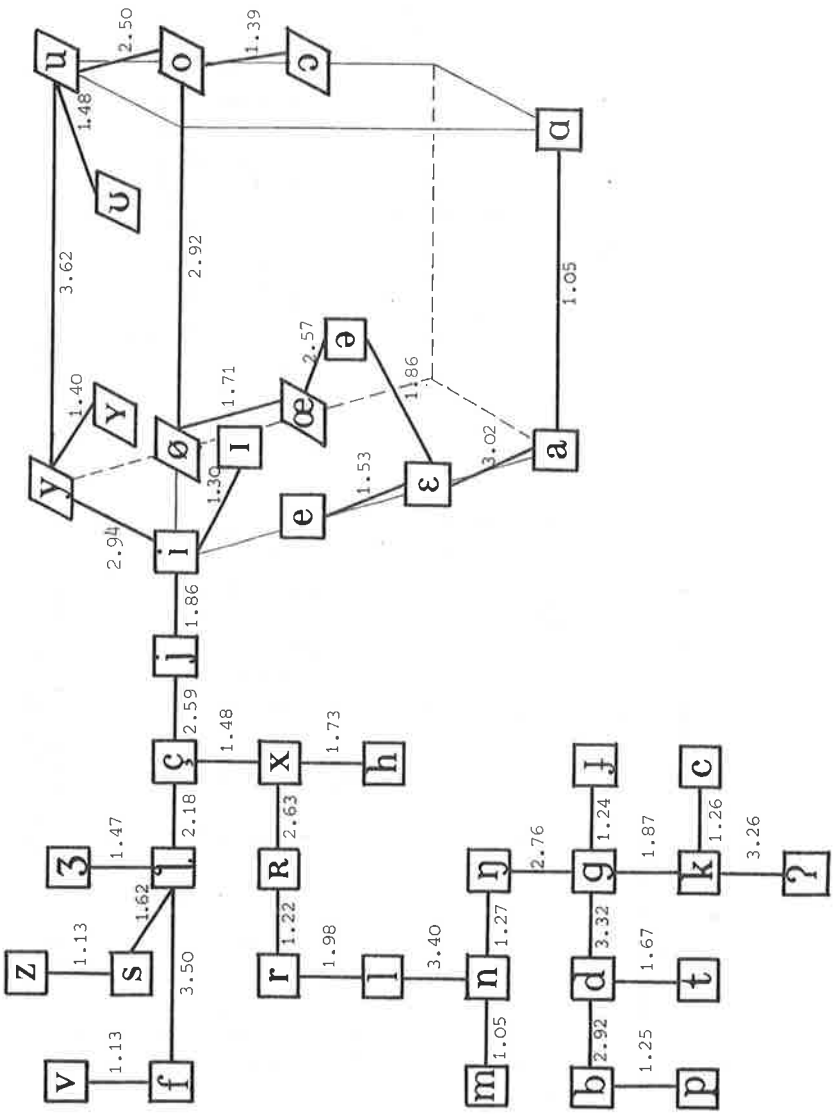


Abb. 2. Minimalbaum für subjektive Distanzen deutscher Laute

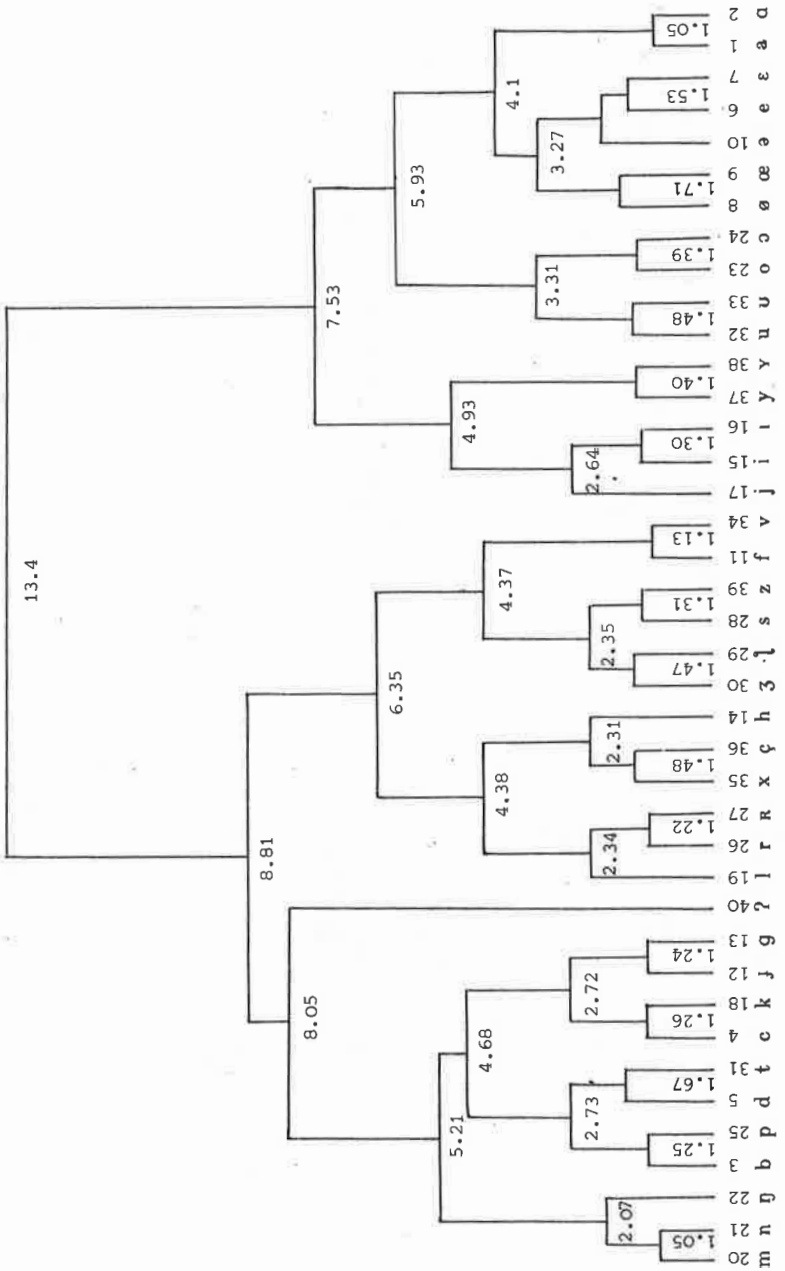


Abb. 3. Klassifikation deutschr Laute nach subjektiver Ähnlichkeit

besondere Stelle zuordnet. Im Graph sieht man aber seine eindeutige Zuordnung zu den Verschlußlauten.

4. Darstellung im euklidischen Raum

Es liegt nahe zu untersuchen, ob sich die Distanzen als (fehlerhafte) Messungen von Abständen zwischen Punkten in einem euklidischen Raum interpretieren lassen. Es ist also die Aufgabe

$$\operatorname{argmin}_{X \in M} \{f(Y - \Delta(X))\}$$

zu lösen, wobei

M	die Menge der reellwertigen (lxm)-Matrizen ist (l = Anzahl der untersuchten Laute, m = Anzahl der Dimensionen des Raumes),
$Y = (y_{ij})$	die experimentell ermittelte Abstandsmatrix ($1 \le i \le l$, $1 \le j \le l$),
$\Delta(X) = (\delta_{ij})$	die Matrix der euklidischen Abstände zwischen den Zeilenvektoren von X ($1 \le i \le l$, $1 \le j \le l$) und
f(.)	die euklidische Norm.

Wir entscheiden uns für m = 2, also eine Darstellung in der Zeichenebene. In einem höher dimensionierten Raum hätte zwar der verbleibende Darstellungsfehler vermindert werden können, doch wäre das Ergebnis unanschaulicher geworden. (Beliebig klein kann der Fehler durch Erhöhung von m nicht gemacht werden, da im Datenkorpus die Dreiecksgleichung nicht durchgängig erfüllt ist.)

Die Analyse führten wir dreimal aus: Je für Vokale und Konsonanten getrennt und für alle Laute zusammen. /j/ zählten wir zu den Vokalen (s.o.). Ausgangspunkt der nach dem Gradientenverfahren durchgeführten Iterationen war jeweils eine Matrix $X^{(0)}$, in der die Punkte gleichmäßig auf einem Kreis mit Mittelpunkt (0,0) und Radius 10 angeordnet waren (zum Gradientenverfahren s. Rumelhart et al. 1986 unter dem Stichwort "δ-rule"). Die berechneten Koordinaten findet man in Tabelle 2a,b,c. Die entsprechenden graphischen Darstellungen zeigt Abbildung 4.

In untenstehender Tabelle sind einige Fehlergrößen für die errechneten Approximationen und die Korrelationen zwischen den empirisch ermittelten Distanzmatrizen und den sich aus den Approximationen ergebenden wiedergegeben. In Anbetracht der schon erwähnten Tatsache, daß die empirisch ermittelten Distanzen nicht der Dreiecksungleichung genü-

Tabelle 2
Berechnete Koordinaten dt. Laute
(a) Vokale

i	:	-3.24, -0.10	ø	:	0.92, 0.17
ɪ	:	-3.26, -0.64	œ	:	0.61, 0.70
j	:	-5.09, 0.02	ɑ	:	2.68, 2.36
y	:	-1.73, -2.36	ɑ	:	2.63, 2.67
ʏ	:	-1.84, -2.87	o	:	2.59, -0.85
e	:	-0.97, 2.34	ɔ	:	3.18, -0.41
ɛ	:	0.28, 2.78	u	:	1.37, -2.77
ə	:	-0.13, 2.34	ʊ	:	1.99, -3.38

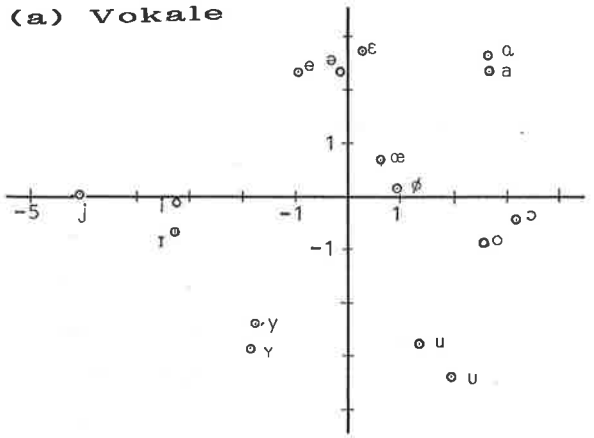
(b) Konsonanten

m	:	-3.00, 3.29	c	:	2.46, 2.14
n	:	-2.54, 3.06	ɟ	:	1.23, 3.82
ŋ	:	-1.83, 3.90	ʔ	:	5.03, 2.37
r	:	-3.47, -1.50	f	:	1.66, -2.45
ʀ	:	-3.44, -2.11	v	:	2.30, -3.11
l	:	-3.74, -0.10	s	:	0.60, -3.43
p	:	-0.43, 2.09	z	:	0.51, -4.26
b	:	-0.49, 2.44	ɭ	:	-0.15, -3.29
t	:	1.55, 0.75	ʒ	:	-0.39, -3.91
d	:	0.57, 1.15	ç	:	-1.26, -3.00
k	:	1.97, 2.44	x	:	-1.31, -2.20
g	:	0.99, 3.32	h	:	3.18, -1.42

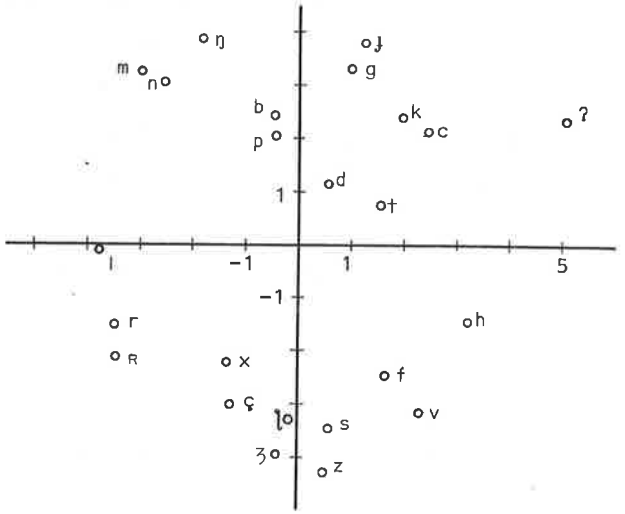
(c) Alle Laute

i	:	2.94, -0.60	ʀ	:	-2.87, -2.72
ɪ	:	3.38, -0.81	l	:	-3.40, -0.99
j	:	1.31, -1.33	p	:	-5.78, -0.07
y	:	5.53, -1.78	b	:	-6.02, -0.21
ʏ	:	5.84, -2.26	t	:	-5.25, -0.92
e	:	5.18, 2.92	d	:	-6.07, -1.29
ɛ	:	6.09, 3.54	k	:	-4.36, 2.03
ə	:	5.14, 3.52	g	:	-4.29, 2.61
ø	:	6.22, 2.19	c	:	-3.99, 2.16
œ	:	6.89, 2.47	ɟ	:	-4.46, 3.13
a	:	5.28, 4.73	ʔ	:	-0.55, 3.96
ɑ	:	5.51, 4.84	f	:	-4.03, -4.12
o	:	7.03, 1.25	v	:	-3.86, -4.59
ɔ	:	6.64, 1.51	s	:	-2.04, -4.30
u	:	6.82, -0.53	z	:	-2.00, -4.96
ʊ	:	7.55, -1.04	ɭ	:	-1.25, -3.51
m	:	-6.92, 1.33	ʒ	:	-1.10, -4.01
n	:	-6.28, 1.64	ç	:	-0.65, -2.11
ŋ	:	-6.66, 2.22	x	:	-1.22, -1.46
r	:	-3.25, -2.31	h	:	-1.07, -0.14

(a) Vokale



(b) Konsonanten



(c) Alle Laute

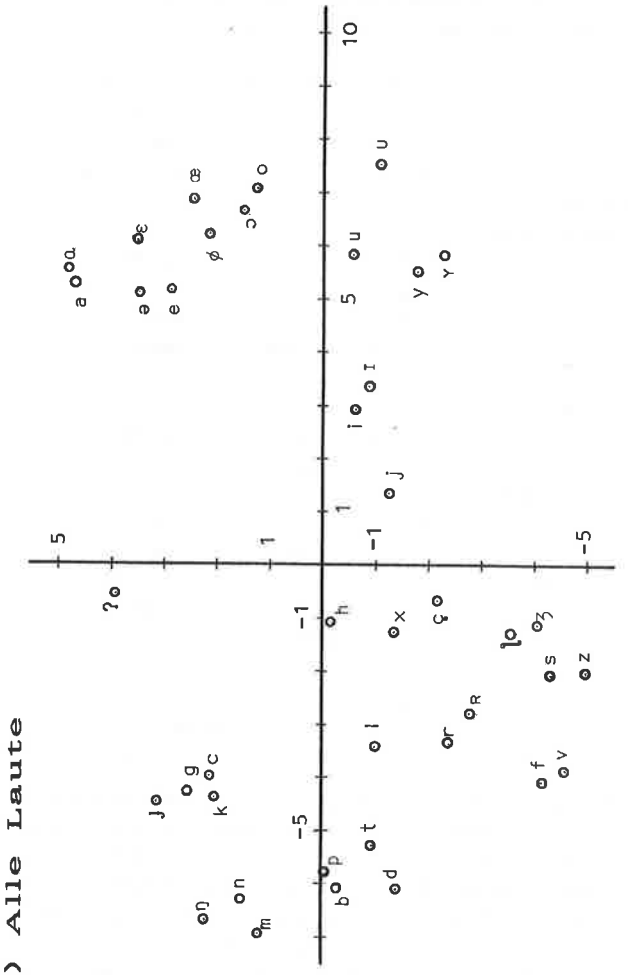


Abb.4. Graphische Darstellung der berechneten Abstände

gen, scheinen die sich aus den approximierten Koordinaten ergebenden Distanzen mit den empirischen zufriedenstellend übereinzustimmen. Dies zeigen auch die Korrelationskoeffizienten: Die Nullhypothese kann in allen Fällen auf hohem Signifikanzniveau abgelehnt werden.

	Vokale	Konsonanten	Gesamt
$f(Y-\Delta(X))$	9.15	28.12	44.95
Mittlere quadratische Abweichung zwischen Y und X	0.0357	0.0488	0.0281
Größter relativer Fehler	0.70 (a:ɑ)	1.64 (x:h)	0.78 (p:b)
Mittlerer relativer Fehler	0.143	0.043	0.167
Streuung des relativen Fehlers	0.143	0.170	0.164
Korelation zwischen Y und ΔX	0.9726	0.8460	0.9635

5. Vergleich mit objektiven Distanzen

An dieser Stelle werden wir überprüfen, ob die subjektiven Lautdistanzen mit denen übereinstimmen, die man aufgrund von Lautzerlegung in dichotome distinktive Merkmale erhält. Eine dichotome Skala ist die stärkste Reduktion eines Begriffsfeldes, bei der viel Information verloren geht. Auch wenn die Dichotomisierung für beschreibende Zwecke (z.B. Regeldarstellung) praktisch ist, ist es fraglich, ob man sie für theoriebildende Zwecke (d.h. für die Formulierung von Gesetzeshypothesen) verwenden kann. Bis jetzt wurde dies weder von den strukturalistischen noch von den generativistischen Phonologen erwiesen. Beim Vergleich sind wir folgendermaßen verfahren:

Als Grundlage der dichotomisierten Darstellung haben wir die Analyse von Wurzel (1970) mit kleinen Ergänzungen genommen. In Tabelle 3 findet man die distinktiven Merkmale deutscher Laute. Die Ähnlichkeiten zwischen den Lauten wurden mit Hilfe von 11 Ähnlichkeitsmaßen berechnet, um die Möglichkeit einer Verzerrung durch ein spezielles Maß auszuschließen. Die Maße findet man in der ersten Spalte von Tabelle 5. Die verwendeten Symbole folgen aus Tabelle 4. So ist A die Zahl der Merkmale, für die zwei Laute gemeinsam 1 haben, B die Anzahl der Merkmale, für die X eine 1 und Y eine 2 hat, usw.

So ist beispielsweise die Ähnlichkeit zwischen [m] und [r] aus Tabelle 3 nach der Formel I (Tab.6) wie folgt:

$$S_{I(m,r)} = \frac{A + D}{N} = \frac{8 + 3}{14} = \frac{11}{14} = 0.7857.$$

Aus den Ähnlichkeiten wurden Distanzen mit der Formel

$$D = 1 - S$$

gebildet und diese Distanzen wurden (für jedes Maß separat) mit den subjektiven Distanzen verglichen. Die Übereinstimmung wurde wieder mit dem Korrelationskoeffizienten überprüft. Die Werte sind in Tabelle 5 für Vokale allein (Spalte 3), Konsonanten allein (Spalte 4) und alle Laute zusammmen (Spalte 5) enthalten. Die benutzten Ähnlichkeitsmaße findet man z.B. in Bock (1974). Die Nullhypothese ist wiederum in allen Fällen abzulehnen.

Da die Realität der Sprache durch den Sprecher/Hörer gegeben wird, kann man der Ermittlung der subjektiven Distanzen keine substantiellen, sondern nur verfahrenstechnische (strategische) Fehlerquellen zuschreiben, wie z.B. zu wenige Versuchspersonen, die den Test durchgeführt haben, oder problematische Versuchsanordnung oder problematische Distanzermittlung. Die Übereinstimmung der Distanzen aus dichotomen Merkmalen weist darauf hin, daß die dichotomische Begriffsbildung hinreichend gut ist. Als Aufgabe für die Zukunft bleibt also der Vergleich mit anderen phonetischen/phonologischen Begriffsbildungen und Überprüfungen in anderen Sprachen.

Tabelle 3

a) Distinktive Merkmale der Konsonanten

	m	n	ŋ	r	ʀ	l	p	b	t	d	k	g	c	ɟ	ʔ	f	v	s	z	ʃ	ʒ	ç	x	h
konsonantisch	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1
silbisch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
nasal	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
obstruent	1	1	1	1	1	1	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1
niedrig	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	2
hoch	1	1	2	1	1	1	1	1	1	1	2	2	2	2	1	1	1	1	1	2	2	2	2	1
hinter	1	1	2	1	2	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	2	1
rund	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
anterior	2	2	1	2	1	2	2	2	2	2	1	1	1	1	1	2	2	2	2	1	1	1	1	1
koronal	1	2	1	2	1	2	1	1	2	2	1	1	1	1	1	1	2	2	2	2	1	1	1	1
dauernd	1	1	1	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
frikativ	1	1	1	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	1
stimmhaft	2	2	2	2	2	2	1	2	1	2	1	2	1	2	1	1	2	1	2	1	1	1	1	1
lateral	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
gespannt	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø

b) Distinktive Merkmale der Vokale

	i	ɪ	j	y	ʏ	e	ɛ	ə	ø	œ	a	ɑ	o	ɔ	u	ʊ
konsonantisch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
silbisch	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
nasal	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
obstruent	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
niedrig	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1
hoch	2	2	2	2	2	1	1	1	1	1	1	1	1	1	2	2
hinter	1	1	1	1	1	1	1	2	1	1	2	2	2	2	2	2
rund	1	1	1	2	2	1	1	1	2	2	1	1	2	2	2	2
anterior	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
koronal	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dauernd	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
frikativ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
stimmhaft	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
lateral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
gespannt	2	1	ø	2	1	2	1	1	2	1	2	1	2	1	2	1

Tabelle 4
Koinzidenztabelle für die Berechnung der Ähnlichkeiten/Distanzen

Merkmals-ausprägung / Laut X	Laut Y: 1	Laut Y: 2	
1	A	B	A+B
2	C	D	C+D
	A+C	B+D	N

Tabelle 5
Vergleich von subjektiven Distanzen und Merkmalsdistanzen

Ähnlichkeitsmaß		Vokale	Konsonanten	Alle Laute
I.	$\frac{A+D}{N}$	0.7089	0.5797	0.7225
II.	$\frac{A+D}{N+B+C}$	0.7096	0.5877	0.7225
III.	$1 - \frac{B+C}{2N - (B+C)}$	0.7074	0.5685	0.7130
IV.	$\left\lvert\frac{A+D-(B+C)}{N}\right\rvert$	0.7091	0.5659	0.6704
V.	$\frac{2D}{2D+B+C}$	0.7014	0.4845	0.6460
VI.	$\frac{1}{2}\left(\frac{D}{B+D} + \frac{D}{C+D}\right)$	0.6683	0.4712	0.6494
VII.	$\frac{D}{\sqrt{(B+D)(C+D)}}$	0.6895	0.4784	0.6483
VIII.	$\frac{D}{D+2(B+C)}$	0.6861	0.5355	0.6705
IX.	$\frac{AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$	0.7045	0.5246	0.6827
X.	$\frac{(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)}$	0.6870	0.5072	0.5947
XI.	$\frac{AD-BC}{AD+BC}$	0.5599	0.4406	0.6227

Literatur

Afendras, E.A., Tzannes, N.S., Trepanier, J.G. (1973) Distance, variation and change in phonology: statistical aspects. *Folia Linguistica* 6, 1–27.

Altmann, G. (1969) Differences between phonemes. *Phonetica* 19, 118–132.

Altmann,G., Lehfeldt, W. (1972) Typologie der phonologischen Distributionsprofile. *Beiträge zur Linguistik- und Informationsverarbeitung* 22, 8–32.

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

Batóg, T., Steffen-Batogowa, M. (1980) A distance function in phonetics. *Lingua Posnaniensis* 23, 47–58.

Bock, H.H. (1974) Automatische Klassifikation. Göttingen, Vandenhoeck & Ruprecht.

Fillenbaum, S., Rapoport, A. (1971) Structures in the subjective lexicon. New York, Academic Press.

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

Geršić, S., Naumann, C.L., Altmann, G. (1985) Subjektive Lautähnlichkeit. *Beiträge zur Phonetik und Linguistik* 50, 101–120.

Grimes, J.E., Agard, F.B. (1959) Linguistic divergence in Romance. *Language* 35, 598–604.

Grotjahn, R. (1980) Zur Quantifizierung der Schwierigkeit des Sprechbewegungsablaufs. In: Grotjahn,R., Hopkins,E.(Hrsg.), *Empirical research on language teaching and language acquisition.* Bochum, Brockmeyer 1980, 199–231.

Lehfeldt, W. (1980) Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs. *Glottometrika* 2, 44–61

Lindner, G. (1975) Der Sprechbewegunsablauf. Eine phonetische Studie des Deutschen. Berlin, Akademie-Verlag.

Lindner, G. (1980) Lautfolgestrukturen im Deutschen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 33, 468–477.

Meinhold, G., Stock, E. (1980) Phonologie der deutschen Gegenwartssprache. Leipzig, VEB Bibliographisches Institut.

Peterson, G.H., Harary, F. (1961) Foundations of phonemic theory. In: Jakobson, R.(Hrsg.) *Structure of language and its mathematical aspects.* Providence, Rhode Island 1961, 139–165.

Prim, R.C. (1957) Shortest connection matrix network and some generalizations. *Bell System Technical Journal* 36, 1389–1401.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986) Learning internal representation by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Hrsg.), *Parallel distributed processing.* Cambridge (Mass.)–London 1986, Vol. I, Kap. 8.

Tolstaja, S.M. (1983) Fonologičeskoe rasstojanie i sočetaemost soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija* Nr. 3, 66–81.

Wängler, H–H. (1961²) Atlas deutscher Sprachlaute. Berlin, Akademie-Verlag.

Wurzel, W.U. (1970) Studien zur deutschen Lautstruktur. Berlin, Akademie-Verlag.

Hammerl, R. (ed.) Glottometrika 10, 1988.

Diversification in the modern Hebrew verbal system

Judith Junger, Rijswijk/Bochum

0. Introduction

Linguistic theories strive to find laws and regularities in the apparently chaotic mass of phenomena called 'language'. Their principles are, however, based on differing methodological approaches. Thus the Transformational Generative stream looks for an idealized native speaker's competence, i.e. the set of regularities which a speaker knows unconsciously and which form the underlying system of his/her language. In such approaches the concrete descriptions are based primarily on the syntactic, and to some extent on the semantic and pragmatic features of the language (cf. Chomsky 1968, 1982). Functionalist theories on the other hand, such as the Prague School grammarians Mathesius (1946), Sgall (1983) or Daneš and Vachek (1964) or Dik's Functional Grammar (1978) choose as a methodological approach the communicative task of language, and in view of this function look at speaker's performance, and the rules and regularities by which the language complies to its role.

A yet different approach is the one advocated by the followers of systems theory, as illustrated in Köhler (1986) and in Köhler & Altmann (1986). This theory advocates a view according to which language is a system that has its own regularities, and these are not rules but mechanisms. The laws are language independent, i.e. equal for all languages, yet with different parameters for all the boundary conditions. Most important, they have a self regulatory character. The principles on which this self regulatory character of language is based should hold over all its domains and should lend themselves to mathematical modelling. One such a descriptive principle was formulated by Zipf (1949), namely, the principle of least effort. According to this principle, the unifications and diversifications in language (which can be described by mathematical models) are due to the speaker's and hearer's aim to alleviate his/her physical or mental effort. Unification and diversification, however, are not merely random fluctuations but processes with both a creative and a

self regulatory character. Their results can be modelled by mathematical functions.

This hypothesis has been successfully applied to various domains of the lexicon, in different languages (cf. Beöthy, Altmann 1984a,b; Altmann 1985a,b; Altmann, Best, Kind 1987; Köhler 1986; Rothe 1981).

The aim of this paper is to examine a case of diversification in Modern Hebrew, more concretely, in a sub-section of Modern Hebrew: its verbal system, in a descriptive way.

1. The verbal system of Hebrew

Semitic languages differ in several polts from the Indo-European ones. The aspect relevant to this discussion is the morphological system. Semitic languages have a system whereby morphology, syntax and semantics are very strongly intertwined. The greater part of the vocabulary is formed by combining a consonantal root with a morpho-phonemic pattern. The root indicates a certain semantic field, and the patterns, the concrete form. The patterns are combined with the root and yield the concrete form (i.e. words). While nominal and adjectival patterns have only a morphological value, verbal patterns also carry syntactic values, expressing transitivity, causativity, reflexivity, etc. This is illustrated below, in (1) verbal and nominal forms derived from a root in Hebrew, in (2) a similar example for Arabic.

(1) ***Hebrew***

a. Verbal forms from the root k.t.b.[1] 'write':

katav 'wrote': 3p.s.m. past tense
ktov! 'write': imperative
kotevet 'write': 3p.s.f. *benoni* (in Modern Hebrew present tense)
yixtevu 'will write': 3p.pl.m. future tense
nixtav 'was written': 3p.s.m.
hixtiv 'made write, dictated': 3p.s.m.

1 There is a phonological change of /b/ to /v/ which is subject to rules regarding word initial or syllable initial position.

b. Nominal forms from the root k.t.b. 'write':

katav 'correspondent, journalist' (m.)
katava 'newspaper report' (f.)
mixtav 'letter' (m.)
ktovet 'address' (f.)
maxteva 'desk' (f.)
ktiv 'spelling' (m.)

(2) ***Arabic***

a. Verbal forms derived from the root k.t.b. 'write':

katab 'wrote': 3p.s.m.
yiktib 'will write': 3p.s.m.
ktib! 'write': imperative
kaatib 'had written': 3p.s.m.

b. Nominal forms from the root k.t.b. 'write':

kaatib 'clerk'
kitaab 'book'
maktab 'office, desk'
maktaba 'library'
maktuub 'written'.

In Hebrew grammars the verbal forms are called ***binyanim***, and the nominal forms are called ***mišqalim***; the terminology was introduced by medieval grammarians, such as Ibn Ginah, Ibn Ezra and others.

The verbal forms conjugate further for tense, number, gender and person.

Consider the following list of all seven *binyanim* in MH, and their syntactic functions (in descending order of frequency):

(3)	***binyan***	***functions***
	PAAL (B1)	active transitive or intransitive
	NIFAL (B2)	'passive' of PAAL
		active intransitive
		reciprocal
		inchoative

PIEL (B3)	active transitive
	causative
PUAL (B4)	'passive' of PIEL
HIFIL (B5)	active transitive
	causative
	inchoative
HUFAL (B6)	'passive' of HIFIL
HITPAEL (B7)	active intransitive
	reflexive
	reciprocal
	inchoative.

The following sentences (from Berman 1978) illustrate the syntactic processes reflected in the *binyanim* forms of verbs with a shared root:

(i) ***Activity/passivity***

active:

(4a) dan imen (B3) et hacevet.
Dan trained GM the-team
'Dan trained the team'.

passive:

(4b) hacevet uman (B4) 'al yedei dan.
the-team was-trained by Dan
'The team was trained by Dan'.

middle:

(4c) hacevet hitamen (B7).
the-team trained (= practiced)
'The team trained'.

active:

(4d) 'hageŝem hirtiv (B5) et harehov.
the-rain made-wet GM the-street
'The rain made the street wet'.

passive:

(4e) harehov hurtav (B6).
the-street was-made-wet
'The street was made wet'.

middle:

(4f) hakvisa nirteva (B2) bageŝem.
the-laundry got-wet (f.) in-the-rain
'The laundry got (became) wet from the rain'.

(ii) ***Causativity***

active:

(5a) haprahim gadlu (B1) bagina.
the-flowers grew in-the-garden
'The flowers grew in the garden'.

causative:

(5b) dan gidel (B3) et haprahim bagina.
Dan grew (= raised) GM the-flowers in-the-garden
'Dan grew the flowers in the garden'.

active:

(5c) dan yaraŝ (B1) harbe kesef (mehazkena)
Dan inherited much money (from the old lady)
'Dan inherited a lot of money (from the old lady)'.

causative:

(5d) hazkena horiŝa (B5) harbe kesef le-dan
the-old-lady left (= bequeathed) much money to-Dan
'The old lady bequeathed much money to Dan'.

(iii) ***Reflexivity***

active:

(6a) dan lavaš (B1) et hame'il.
Dan wore (= put on) GM the-coat
'Dan put on the coat'.

reflexive:

(6b) dan hilbiš (B5) et 'acmo = dan hitlabeš (B7)
Dan made-wear GM himself = Dan dressed
'Dan dressed'.

(iv) ***Incoativeness state*** (with adjective)

(7a) habasar kar (B1)
the-meat cold
'The meat is cold'.

inchoative:

(7b) habasar hitkarer (B7)
the-meat got-cold
'The meat got cold'.

(v) ***Reciprocity***

(8a) raiti (B1) et rina šavu'a še'avar.
saw-I GM Rina week last
'I saw Rina last week'.

(8b) rina veani mitraot (B7) pa'am bešavu'a.
Rina and-I see-reciprocal once a-week
'Rina and I see each other once a week'.

(vi) ***Ingression***

active:

(9a) dan 'amad (B1) 'al hašulhan.
Dan stood on the-table
'Dan stood on the table'.

ingressive:

(9b) dan ne'amad (B2).
Dan stood up (= got up)
'Dan stood up'.

Before going into the question of how the laws of unification and diversification are manifested in the Modern Hebrew verbal system, here is some statistical data on the occurrence of the roots in the seven *binyanim*, divided into groups of the various configurations in which they occur. The following list gives only the total number of roots in one to seven *binyanim*, i.e. the distribution of the variable completeness/deficiency of the system. A detailed distribution is listed in the Appendix.

Number of binyanim in which an individual root occurs x	Number of roots occurring in x binyanim
1	263
2	574
3	691
4	320
5	257
6	185
7	165
Total	2452

2. Unification and diversification in the overal verbal system of Modern Hebrew

In this section we will look at the law of unification and diversification on the level of the overall verbal system of Modern Hebrew. As shown in the previous section the organizing principle of the verbal system are the *binyanim*, the patterns which are the morphological device for deriving the concrete words from the roots, and which at the same time express semantic and syntactic features of the verbs. The balance between diversification and unification can be expected to manifest itself in the frequency of occurrence of the various *binyanim*. As we saw in section 1, apart from the morphological difference, the *binyanim* differ from each other in the

a. valency of the verbs in the various *binyanim*;
b. the number of functions of the *binyanim*.

For these two features, diversification would manifest itself in high valency (as suggested in Altmann 1987) as opposed to unification which is manifest in low valency; in addition, diversification would be manifest in a multi-functionality of the *binyanim*, whereas unification in a uni-functionality of the *binyanim*. Concretely, this divides the seven *binyanim* into the following groups:

(i) ***Valency***:

1 valency: PAAL, PUAL, NIFAL, HIFIL, HUFAL, HITPAEL
2 valencies: PAAL, PIEL, HIFIL
3 valencies: HIFIL.

The PAAL and the HIFIL occur in two valency classes; therefore it is necessary to check the roots individually in order to decide into which class they belong.

(ii) Multifunctionality:

- Total consistency: only 1 function:

PUAL and HUFAL (always passive)

- Partial consistency: 2 possible functions:

PAAL (transitive, intransitive)
PIEL (active transitive, causative)

- Divergent: 3 or 4 possible functions:

HIFIL (causative, inchoative, active-transitive)
HITPAEL (active-intransitive, reflexive, reciprocal)

- Very divergent: 5 possible functions:

NIFAL (passive, active-intransitive, inchoative, reflexive, reciprocal).

The root counts gave the results as presented in Tables 1 to 4.

Table 1

The valency distribution of the roots(in all the *binyanim*)

Valency x	Number of roots f_x
1	4634
2	3206
3	17

Table 2

The relation between the variables valency and number of functions expressed in terms of the number of *binyanim*

		Number of functions				
		1	2	3	4	5
Valency	1	2			1	1
	2		2			
	3			1		

Table 3

The frequency distribution of the *binyanim* as function of their valency

Valency	*binyan*	Number of roots	
1	PAAL	607	
	NIFAL	942	
	PUAL	1068	
	HUFAL	723	
	HITPAEL	1294	
	HIFIL	9	
			(4643)
2	PAAL	644	
	PIEL	1517	
	HIFIL	1036	
			(3197)
3	HIFIL	17	

The root distribution as a relation to number of functions per *binyan* is summarized in Table 4.

Table 4

The frequency distribution of the *binyanim* in terms of the number of their semantic–syntactic functions

Number of functions	*binyan*	Number of roots
1	PUAL (B4)	1068
	HUFAL (B6)	723
2	PIEL (B3)	1517
	PAAL (B1)	1076
3	HIFIL (B5)	1062
4	HITPAEL (B7)	1294
5	NIFAL (B2)	942

The data of Tables 1 to 4 can be summed up qualitatively as presented in Table 5 or quantitatively as presented in Table 6.

Table 5

The membership of the *binyanim* in the valency and function classes

Valency	The maximum number of functions 1	2	3	4	5
1	PUAL HUFAL			HITPAEL	NIFAL
2		PAAL PIEL			
3			HIFIL		

Table 6

The number of roots in the valency and function classes

Valency	The maximum number of functions 1	2	3	4	5
1	1665			1294	942
2		2768			
3			1062		

Note that if these numbers are added up they amount to 7731, whereas the basis for the *binyanim* distribution were 2542 roots only. The reason is that these numbers were reached by adding up all the roots that occur in the PAAL, NIFAL etc. But this means that many roots are counted more than once: for example there are 79 roots occurring in the PAAL and the NIFAL patterns; these 79 roots are counted once when all the roots occurring in the PAAL are counted, and another time when all the roots occurring in the NIFAL are counted. Or the 165 roots occurring in all 7 patterns are counted seven times, for the sum of each pattern. These counts are nevertheless useful because they give an indication of the relative size of the valency and function classes.

The tendencies of unification and diversification in language function as follows: both the speaker and the hearer strive towards minimal effort. Ideally for the speaker the situation of minimal effort would be

achieved by total unification, that is when one word would express all meanings, as this minimizes the effort of memorizing and also the effort of encoding the message in words. Such a situation is, of course, impossible, as it would lead to total unclarity when the message has to be decoded. For the hearer, on the other hand, the situation of least effort would be achieved by total diversification: when for every meaning there would be a separate word. This minimizes the effort of decoding. In other words, the speaker tends to a ratio of 'word:meaning' of 1:all, whereas the hearer strives to a ratio of 'word:meaning' of 1:1. In reality in all the natural languages a compromise is reached between these two tendencies. In a so called 'normal' compromise situation the ratio 'word:number of meanings' should be of a 'bell shape' as in Figure 1 below.

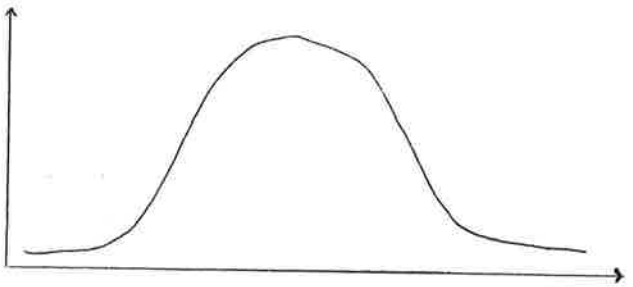


Figure 1. The normal ratio 'word:number of meanings'

In reality, this compromise situation is seldom achieved because of the active role played by the speaker: in most cases the curve is skew. When checking these two tendencies of unification and diversification in the Modern Hebrew verbal system we will look at root distribution with regard to the features of valency and uni/multi-functionality. In the Hebrew verbal system both of these parameters are expressed throught the *binyanim*.

On a general level a situation of total unification means that all the roots occur in 1 *binyan* only, and the situation of total diversification means that all the roots occur in all 7 *binyanim*. This situation is in fact quite well approximated in Contemporary Hebrew. The distribution of the roots in the *binyanim* is expressed in Figure 2.

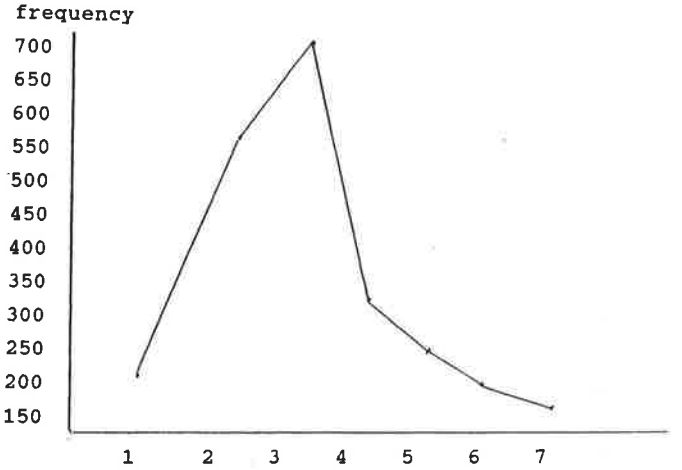


Figure 2. The distribution of the roots in the *binyanim*

The distribution of the roots according to their maximal valency is expressed in Figure 3 and the distribution of the roots according to their number of functions in Figure 4.

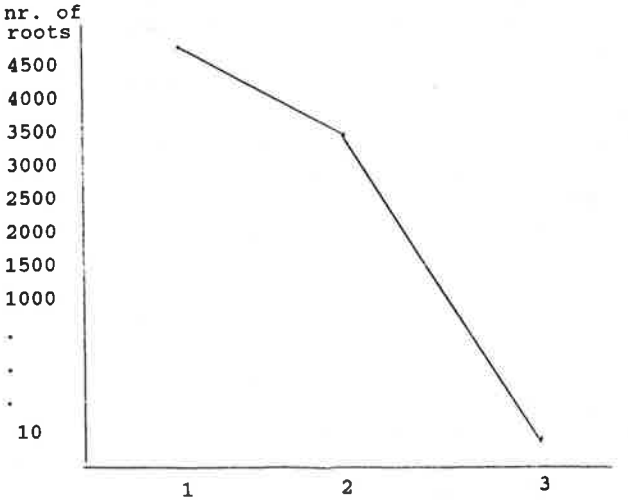


Figure 3. The distribution of roots according to their maximal valency

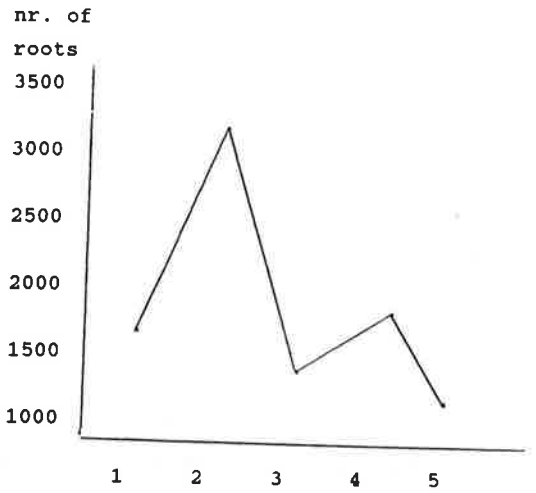


Figure 4. The distribution of roots according to their maximal number of functions

These figures allow us to state the following facts:

(1) The modus of the frequencies is in 3 *binyanim*, i.e. most roots can occur in three classes.

(2) The highest frequency of roots is with valency 1, the lowest with valency 3.

(3) The highest frequency of roots is in *binyanim* with 2 functions, the lowest in *binyanim* with 5 functions.

In other words, there is a clear tendency towards low valency and a low number of functions. In terms of speaker-hearer competition this means the following:

- low valency means that the verbs occur in a syntagmatically simple environment, which in turn means simplicity in encoding. Simplicity of encoding results in a low effort for the speaker.
- low number of functions means a 1:1 ratio of 'function:meaning', which in the verbal system is the equivalent of 1:1 ratio 'lexical meaning:word in the nominal system'. This ratio implies ease of decoding, that is low effort for the hearer.

Hence, in fact both tendencies 'low effort for the hearer' and 'low effort for the speaker' are present in the *binyanim* system. The compromise between them, however, is not represented in the bell-shape diagram, because they do not overlap. They are manifested in two different aspects of the verbal system: for the speaker low valency and for the hearer occurrence in *binyanim* with a low number of functions.

3. Unification and diversification for each *binyan*

In the previous section the diversification was checked on the level of the whole verbal system. In this section it will be checked for each verbal pattern separately, by comparing the root distribution diagram to the normal bell shape.

Let us look now at the distribution of the roots for each *binyan* individually according to its combinations with other *binyanim* as presented in Tables 7 to 13.

Table 7

Distribution of the roots in the PAAL

Number of joint *binyanim*	Number of roots
1	85
2	152
3	193
4	262
5	227
6	167
7	165 (1251)

Note that in this table, and also in all the following tables for each separate *binyan* the numbers are absolute. That is, here every root has been counted only once, so these numbers have an absolute value. Only the tables based on adding-up tables for individual patterns have overlaps and represent merely relative size and no absolute numbers. Table 7 is to be read as follows: there are 85 roots that can occur only in PAAL; there are 152 roots that can occur in PAAL and in one other *binyan;*

there are 193 roots that can occur in PAAL and in two other *binyanim*, etc.

Table 8

Distribution of the roots in the NIFAL

Number of joint *binyanim*	Number of roots
1	8
2	83
3	121
4	218
5	194
6	153
7	165 (942)

Table 9

Distribution of the roots in the PIEL

Number of joint *binyanim*	Number of roots
1	101
2	154
3	518
4	193
5	207
6	179
7	165 (1517)

Table 10

Distribution of the roots in the PUAL

Number of joint *binyanim*	Number of roots
1	0
2	24
3	466
4	140
5	126
6	147
7	165 (1068)

Table 11

Distribution of the roots in the HIFIL

Number of joint *binyanim*	Number of roots
1	31
2	133
3	148
4	210
5	193
6	182
7	165 (1062)

Table 12

Distribution of the roots in the HUFAL

Number of joint *binyanim*	Number of roots
1	0
2	84
3	92
4	119
5	133
6	130
7	165 (723)

Table 13

Distribution of the roots in the HITPAEL

Number of joint *binyanim*	Number of roots
1	38
2	114
3	494
4	138
5	187
6	158
7	165 (1294)

In order to see how the tendencies of unification and diversification are manifested in each *binyan* there is a diagram representation below. For each *binyan* the features of valency and number of functions, i.e. the features representing speaker/hearer effort will be discussed for the largest and the smallest group of roots, since it is expected that the two extreme points are the ones which give interesting information not only on the situation in MH at present but also on the direction of development. Remember that low valency is low speaker effort and low number of functions for the pattern is low hearer effort.

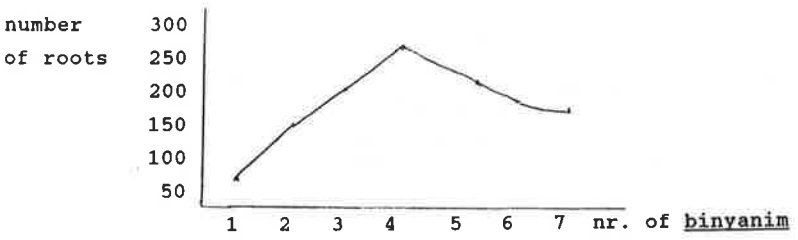


Figure 5. The PAAL pattern

The curve for the PAAL is fairly close to the normal bell shape. It is not entirely symmetric, but there is a balance between the root distribution in 2 and in 6, 7 *binyanim*. The highest number of roots occurs in 4 *binyanim* (262), with the following valency:

PAAL transitive (valency 2): 114
PAAL intransitive (valency 1): 148.

Regarding functionality, the PAAL is fairly constant, having two possible functions (transitive and intransitive). In terms of speaker-hearer effort, the situation is quite balanced: on the one hand, there are more roots with valency 1 than with valency 2 (i.e. low speaker effort), but not significantly, and the low hearer effort of the low number of functions is also present. Hence, although the diagram is not exactly of

a bell shape, the balance between the tendencies of unification and diversification is fairly good.

The lowest number of roots in PAAL occurs when PAAL is the only pattern, and it is for all 85 roots intransitive. This means higher priority to low speaker effort.

Looking at the two extremes we see that there is a slight weighting in favour for the speaker (low valency). One can presume that in earlier historical stages the valency was higher, but the situation is changing now towards a decrease of the speaker's comfort (which probably implies automatic increase of hearer's comfort).

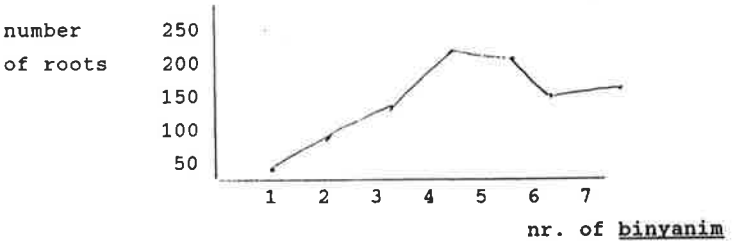


Figure 6. The NIFAL pattern

The curve for the NIFAL pattern has a bell-like shape, but slanted. The highest number of roots occurs in 4 *binyanim*, (218 roots) but there is no symmetry between the tails of the bell, that is, between the number of roots in 1 and 2 vs. 6 and 7 *binyanim*.

The roots in the NIFAL in the group of 4 *binyanim* have the following class affiliation:

intransitive : 43
passive: 182 total valency 1: 225 roots.
inchoative: 5

Here there is a clear preference of the speaker's low effort (expressed in low valency) vs. the hearer's effort (since the functionality is not so simple any more: there are 3 possible functions).

The lowest number of roots in the NIFAL pattern occur when the NIFAL is the only *binyan* in which the root occurs. There are 8 roots only in NIFAL, all of them intransitive (which is logically the only possibility, as all the other functions express a contrast and have to be derived so that the root in question occurs in more patterns).

Summing up, the distribution of the roots in the NIFAL pattern shows preference for low effort for the speaker.

The curves of both the PAAL and the NIFAL are not of the normal bell shape, but are not very different from it. The following three curves, for the PIEL, the PUAL and the HITPAEL patterns show a large deviance from the bell shape:

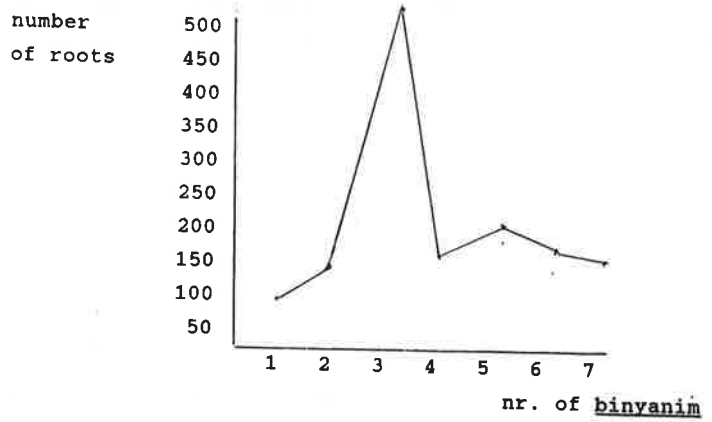


Figure 7. The PIEL pattern

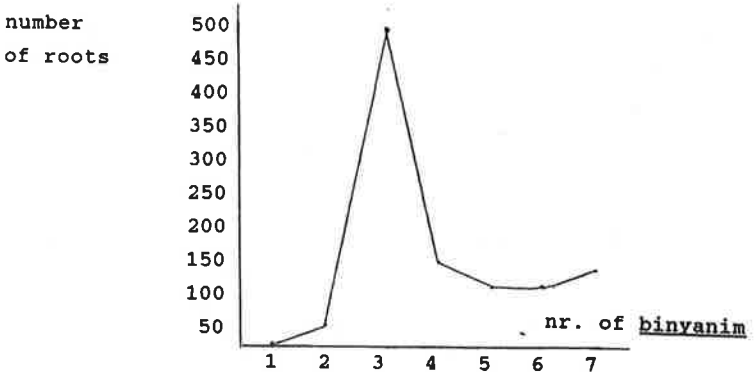


Figure 8. The PUAL pattern

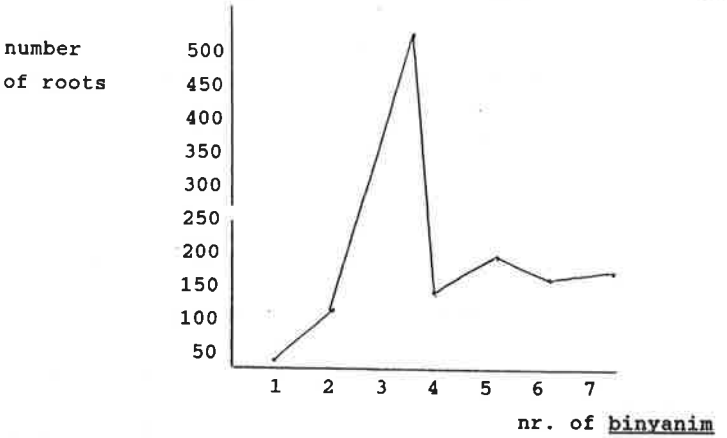


Figure 9. The HITPAEL pattern

The patterns PIEL, PUAL and HITPAEL clearly form a separate block in the system. Their diagrams are very similar. In the total verbal system the roots occurring in these three patterns form the largest configurational sub-system: 432 roots (see Junger 1988, ch.4). In all three graphs there is a high peak in occurrence of the combination of 3 *binyanim*, each time the combination PIEL PUAL HITPAEL. The syntactic relations expressed by the PIEL, PUAL and HITPAEL patterns are active transitive - passive - middle (intransitive). Not only is the number of roots highest in this group, but it also includes most of the roots which were recently added to Modern Hebrew. These are easy to recognize as they have four radicals (vs. the traditional three radicals)[2]. In terms of speaker-hearer effort the picture is as follows:

2 Here a clear tendency in the development of the Hebrew syntax can be seen: a simplification in the *binyanim* system. The *binyanim* system also traditionally expressed features like causativity, reflexivity and reciprocity; these tasks are disappearing, and the syntactic relations expressed by the *binyanim* are the basic ones of transitivity and active-passive contrast. This tendency, shown here in the diagrams, is supported also by independent studies (Berman 1979; Bolozky 1978). They show that there is an increasing tendency in Modern Hebrew (certainly as regards the spoken language) to express causativity by means of the periphrastic form *garam* 'cause', and to express reflexivity and reciprocity, instead of with HITPAEL pattern, rather by means of a verb in the active transitive form (PAAL or PIEL) and the reflexive or reciprocal pronouns.

Binyan	*Functions*		*Valency*	
PIEL	transitive:	358	2	
	causative:	69	2	total valency 2: 427
	intransitive:	5	1	
PUAL	'passive':	432	1	
HITPAEL	intransitive:	347	1	
	inchoative:	69	1	
	reflexive:	14	1	
	reciprocal:	2	1	total valency 1: 432.

In terms of speaker-hearer effort the following tendencies are found:

1. There is a ratio 2:1 of valency 1:valency 2, i.e. preference of speaker's ease.

2. Except for the PUAL which has only one function, both the PIEL and the HITPAEL have 3 and 4 functions respectively, which is contrary to the hearer's ease of decoding.

Hence in the case of the PIEL, PUAL and HITPAEL patterns there seems to be a clear preference of the spreaker's minimal effort.

Both the HIFIL and the HUFAL patterns have a curve which deviates from the norm. Furthermore, although they form a pair of active-passive equivalents, their curves are entirely different. Whereas the curves of the PAAL and NIFAL showed some similarity, and those of the PIEL and PUAL great similarity, those of the HIFIL and HUFAL are entirely different.

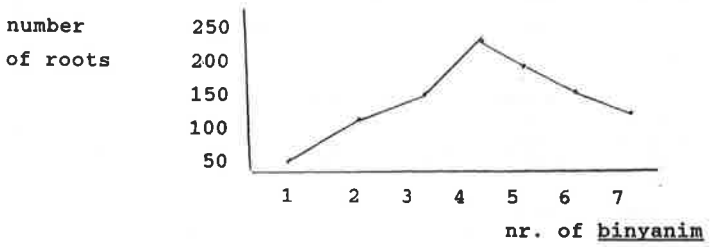


Figure 10. The HIFIL pattern

The HIFIL pattern, like the PAAL and the NIFAL has the highest frequency of occurrence in a combination of four *binyanim*. Its distribution is as follows:

causative : 198 roots
active-transitive: 1 root.

In all cases it has valency 2. This is a fairly balanced situation with regard to speaker vs. hearer effort, with a similar degree of multi-funtionality and valency.

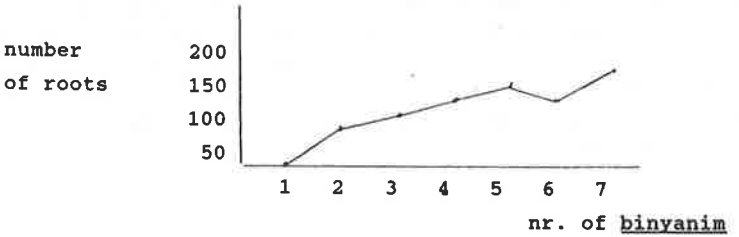


Figure 11. The HUFAL pattern

The highest distribution of the roots in the HUFAL is 7 *binyanim*. But it has only 1 function: 'passive', and its valency is always 1. Here too, therefore, there is no clear difference between speaker or hearer minimal effort.

4. Conclusion

In this article, an attempt has been made to investigate the tendencies of unification and diversification in the verbal system of Modern Hebrew. These features were analyzed in the context of Zipf's forces whereby in every language there is a conflict between the principles of least effort of, on the one side, the speaker and, on the other, the hearer. These are

expressed in the tendencies of unification and diversification. Unification is the expression of many (in the extreme case all the) meanings by one word, whereas diversification is the expression of every meaning by a seperate word. The former results in maximal ease of encoding messages, and hence least effort for the speaker, whereas the latter results in maximal ease of decoding, and hence least effort for the addressee. In the verbal system of Modern Hebrew both the syntactic and semantic features are expressed by means of the *binyanim*. It is claimed that the tendencies of unification and diversification in the *binyanim* system regard not one specific aspect, but rather two different features: valency of each *binyan* and the number of syntactic-semantic functions each *binyan* can have. That is the principle of least effwort is expressed in two different aspects of the verbal system: least effort for the speaker is manifested in low valency (simple syntagmatic environment) whereas least effort for the hearer is expressed in a *binyan* having only one or maximally two possible functions. In section 2 and 3 respectively the interaction between these two features is described for the whole of the system and for each pattern separately. It has been demonstrated that indeed in most cases there is a balanced compromise situation between the two principles.

Appendix

The root distribution in the *binyanim* system (Junger 1987)

(i) Roots occuring in one *binyan* (263 roots)

PAAL (85)
NIFAL (8)
PIEL (101)
HITPAEL (38)
HIFIL (31)
PUAL (0)
HUFAL (0)

(ii) Roots occurring in two *binyanim* (574 roots)

PIEL PUAL (22)
PIEL HITPAEL (97)
PAAL HIFIL (32)
PAAL NIFAL (79)
HIFIL HUFAL (84)
PAAL PIEL (25)
PAAL HITPAEL (14)
PIEL HIFIL (9)
NIFAL HIFIL (3)
HIFIL HITPAEL (3)
PUAL HIFIL (2)
PAAL PUAL (2)
NIFAL PIEL (1)

(iii) Roots occurring in three *binyanim* (691 roots)

PIEL PUAL HITPAEL (432)
PAAL HIFIL HUFAL (50)
PAAL NIFAL HIFIL (46)
PAAL NIFAL PIEL (24)
NIFAL HIFIL HUFAL (21)
PAAL PIEL PUAL (19)
PAAL PIEL HITPAEL (19)
PAAL NIFAL HITPAEL (16)
PAAL HIFIL HITPAEL (11)
HIFIL HUFAL HITPAEL (9)
PIEL HIFIL HUFAL (7)
PIEL PUAL HIFIL (7)
PAAL NIFAL HUFAL (5)
NIFAL PIEL PUAL (5)
PIEL HIFIL HITPAEL (5)
PAAL NIFAL PUAL (2)
PAAL PUAL HITPAEL (1)
PAAL PUAL HIFIL (1)
NIFAL PIEL HITPAEL (1)
NIFAL PIEL HIFIL (1)

(iv) Roots occurring in four *binyanim* (320 roots)

PAAL NIFAL HIFIL HUFAL (79)
PAAL NIFAL PIEL PUAL (50)
PAAL PIEL PUAL HITPAEL (30)
PAAL NIFAL HIFIL HITPAEL (23)
PAAL NIFAL PIEL HIFIL (20)
PIEL PUAL HIFIL HITPAEL (19)
PAAL HIFIL HUFAL HITPAEL (17)
PAAL NIFAL PIEL HITPAEL (15)
PAAL PIEL HIFIL HITPAEL (13)
NIFAL PIEL PUAL HITPAEL (13)
PAAL PIEL PUAL HIFIL (10)
PIEL PUAL HIFIL HUFAL (8)
NIFAL PIEL HIFIL HUFAL (6)
NIFAL PIEL PUAL HIFIL (4)
PIEL HIFIL HUFAL HITPAEL (3)
PAAL NIFAL PUAL HIFIL (2)
NIFAL HIFIL HUFAL HITPAEL (2)
PAAL NIFAL PUAL HITPAEL (1)
PAAL NIFAL PUAL HUFAL (1)
PAAL PUAL HIFIL HUFAL (1)
NIFAL PIEL HIFIL HITPAEL (1)
NIFAL PUAL HIFIL HUFAL (1)
PIEL HIFIL HUFAL HITPAEL (1)

(v) Roots ocurring in five *binyanim* (257 roots)

PAAL NIFAL PIEL PUAL HITPAEL (61)
PAAL NIFAL HITPAEL HIFIL HUFAL (46)
PAAL NIFAL PIEL HIFIL HUFAL (30)
PAAL NIFAL PIEL PUAL HIFIL (21)
PAAL PIEL PUAL HIFIL HITPAEL (17)
PAAL PIEL HITPAEL HIFIL HUFAL (17)

PAAL NIFAL PIEL HITPAEL HIFIL (17)
PIEL PUAL HITPAEL HIFIL HUFAL (15)
PAAL PIEL PUAL HIFIL HUFAL (12)
NIFAL PIEL PUAL HIFIL HITPAEL (7)
NIFAL PIEL PUAL HIFIL HUFAL (5)
NIFAL PIEL HIFIL HUFAL HITPAEL (2)
PAAL NIFAL PUAL HIFIL HUFAL (2)
PAAL NIFAL PIEL HUFAL HITPAEL (1)
PAAL NIFAL PUAL HIFIL HITPAEL (1)
PAAL PIEL PUAL HUFAL HITPAEL (1)
PAAL PUAL HIFIL HUFAL HITPAEL (1)
NIFAL PIEL PUAL HUFAL HITPAEL (1)

(vi) Roots occurring in six *binyanim* (185 roots)

PAAL NIFAL PIEL PUAL HITPAEL HIFIL (56)
PAAL NIFAL PIEL HITPAEL HIFIL HUFAL (39)
PAAL PIEL PUAL HITPAEL HIFIL HUFAL (33)
PAAL NIFAL PIEL PUAL HIFIL HUFAL (28)
NIFAL PIEL PUAL HITPAEL HIFIL HUFAL (19)
PAAL NIFAL PUAL HITPAEL HIFIL HUFAL (7)
PAAL NIFAL PIEL PUAL HITPAEL HUFAL (4)

(vii) Roots occurring in all seven *binyanim* (165 roots)

PAAL NIFAL PIEL PUAL HIFIL HUFAL HITPAEL (165)

In total there are 2452 roots.

References

Altmann, G. (1985a) Semantische Diversifikation. *Folia Linguistica 19, 177–200.*

Altmann, G. (1985b) Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139–155.*

Altmann, G., Best, K.–H., Kind, B. (1987) Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8, 130–139.*

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

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

Berman, R. (1975) The morphological realization of syntactic processes in the *binyanim* system. *Hebrew Computational Linguistics 9, 25–39.*

Berman, R. (1978) Modern Hebrew Structure. Tel Aviv, University Publishing Projects.

Berman, R. (1979) Lexical decomposition and lexical unity in the expression of derived verbal categories in Modern Hebrew. *Afroasiatic Linguistics 6, 117–142.*

Bolotzky, Sh. (1978) Word formation strategies in the Hebrew verb system: denominative verbs. *Afroasiatic Linguistics 5, 11–136.*

Chomsky, N. (1965) Aspects of the theory of syntax. Cambridge, Mass., MIT Press.

Chomsky, N. (1982) Lectures on government and binding. Dordrecht, Foris.

Daneš, F., Vachek, J. (1964) Prague studies in structural grammar today. *Travaux Linguistique de Prague 1.*

Dik, S.C. (1978) Functional grammar. 3rd ed. Dordrecht, Foris.

Junger, J. (1985a) Morphological causatives in Modern Hebrew. In: A.M.Bolkestein, J.L.Mackenzie, C.de Groot (eds.), *Predicates and terms in functional grammar.* Dordrecht, Foris, 235–257.

Junger, J. (1985b) Valentie reductie in het Modern Hebreeuws. *Tijdschrift voor Taal– en Tekst Wetenschap 2, 141–160.*

Junger, J. (1988) Predicate formation in the verbal system of Modern Hebrew. Dordrecht, Foris.

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

Köhler, R., Altmann, G. (1986) Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft 5, 253–265.*

Mathesius, V. (1946) On some problems of the systematic analysis of grammar. *Travaux du Circle Linguistique de Prague 6, 95–106.*

Rothe, U. (1986) Die Semantik des textuellen et. Frankfurt, Lang.

Sgall, P. (1983) Semantics and pragmatics from a Praguian point of view. *Travaux Linguistiques de Prague 10,277–281.*

Zipf, G.K. (1949) Human behaviour and the principle of least effort. Cambridge, Mass., Addison–Wesley.

Hammerl, R. (Ed.), Glottometrika 10, 1988.

Hypotheses about compounds*

Gabriel Altmann, Bochum

1. Introduction

There is no way to form a sharp boundary between a "pure compound" and a free combination of two words that would apply to all languages. A compound represents a strong bonding of two words or stems, but it is always an artificial criterion that determines how great this strength must be in order to designate an entity as a compound.

In spite of this difficulty we proceed on the assumption that in a given language it is in principle always possible to single out compound-like entities, although in individual cases decision problems may arise.

Compounds can be classified according to different properties, e.g. according to the word classes of the components (N+N, S+N, A+A etc.), according to the number of the components, according to the manner/type of bonding, etc.

All these problems are surely important both for the investigation of grammar and for the teaching of languages, because they provide information about the *status quo* of a language, but they are merely results of the dynamics of composition, which is the only point important for a theory of language. The exterior appearance of a compound may be different in different languages, but the dynamics of composition, which in our opinion abides by laws, must be the same for all languages.

The basic motivation for forming compounds is naturally the need to express a concept not having as yet a sound form, i.e. the need to express oneself or a state of affairs. This "Bühlerian" requirements can be served by several means, according to the grammatical processes permitted in the given language and according to the productivity of the process for the creation of a compound of the required kind. Theoretically, the speaker has several possibilities to express a state of affairs; they all differ in their degree of compactness or descriptivity. The least compact, i.e. the most descriptive way to express a concept is paraphrasing, e.g. with a relative clause. If the speaker chooses a more compact form (compound, derivate, reduplication, etc.) e.g. by deciding to use a compound, then on one hand he partially satisfies his need for minimizing the production effort (cf. Köhler 1986), but on the other hand he loses a number of degrees of freedom associated with free paraphrasing.

* This study was written as part of the project "Language Synergetics" sponsored by the STIFTUNG VOLKSWAGENWERK.

From the instant when the speaker makes this decision, he seems to be motivated only semantically when building a new compound, and he seems to choose the stems so as to express the concept as exactly as possible. His freedom of choice is nevertheless immense, since he can use formal, functional, processual, etc. similarities, metaphors, contiguities, etc. in order to express the concept by means of a compound. Thus the German "Kindergarten" is no "garden", "Baumschule" (tree nursery) is no "school", and "Katzenjammer" (hangover) has nothing to do with cats.

Though in most compounds we can detect some semantic affinity between the meaning of the joint stems and that of the compound itself, nevertheless we have the impression that from the semantic point of view there are no rules of compounding, the control is very weak (if at all present), the arbitrariness of the process is so strong that we can speak of chaos.

The correctness of this assumption can easily be corroborated if one compares synonymous compounds in different languages: one finds that the meanings of the components are different, e.g. the compounds "railway", "chemin de fer" and "Eisenbahn" have all at least one different component. This is the reason why historical semantics has attained only descriptive results as yet. Now since in the semantics of the construction of individual compounds there are no invariants, the theory of language is not interested in this kind of research.

There are of course other factors which – under the given boundary conditions – control the compounding process, at least stochastically. From the synergetic point of view we consider them to be order parameters (cf. Haken 1978), such as are present in Köhler's (1986) self-regulating system. The processes put into operation by these factors are lawlike, i.e. they hold for all languages, they can be derived deductively, they can be imbedded into a system of equivalent statements and the results can be tested statistically (cf. Bunge 1967).

Below we shall examine the impact of Köhler's order parameters *polylexy, length, frequency* and *polytexty* on the process of compounding and the "regularity" they call into existence. We shall for the moment consider merely the qualitative aspect of the problem without proposing mathematical models.

2. Hypotheses concerning meaning

Though the meanings of the participant stems of a compound need not be components of the meaning of the compound, cf. "Katzenjammer" (hangover), and though there are no rules as to how compounds in particular cases must be semantically formed, the meaning is not irrelevant for the compounding process. However, only those aspects of meaning are relevant for the theory of language that hold for the language as a whole and at the same time for all languages. Four hypotheses concerning meaning can be formulated:

A. The need for minimizing memory effort (cf. Köhler 1986) causes both the speaker and the hearer to construct the majority of compounds in such a way that the meanings of all their morphological components (participant stems) are at the same time components of the meaning of the compound, e.g. "Dampfschiff", "Baumaschine" etc. But in "Kindergarten" only the meaning of "Kind" is part of the meaning of the compound. "Garten" is used in a metaphorical sense. The lexicon of a language surely includes fewer compounds of the latter kind and still fewer of the sort where no participant stem has a meaning which is a component of the meaning of the compound, as is the case with "hangover". This is because the smaller the semantic correspondence, the greater the memory effort. So in all languages having compounds the proportions of compounds according to the measure of semantic correspondence are fixed. We can formulate the hypothesis as follows:

> The number of compounds in a language (having compounds) decreases proportionally to the measure of semantic correspondence of the components with the compound; or, inversely, the number of compounds in a language is the greater, the greater the semantic correspondence of the compound with its components.

The derivation of a mathematical model is here quite simple; the problem is how to test such a model. With compounds consisting of two components our variable "semantic correspondence" can take only the values 0, 1 or 2, so that we obtain an empirical frequency distribution with only 3 classes. The mathematical model can consequently only have one parameter in order to be testable (e.g. with the aid of a chi-square test). In case of a model with more parameters we must dispense with a statistical test. The situation improves somewhat with longer compounds, which have automatically more frequency classes. Unfortunately there are few long compounds in languages, so that the testing of this hypothesis will be somewhat difficult. There are three ways to alleviate the testing problem: (i) Using only the distribution of the longest compounds, for which there are enough frequency classes; (ii) deriving a two dimensional distribution with the variables "semantic correspondence" and "length" (= number of components); in this case we gain more degrees of freedom; (iii) refining the measure of correspondence so that it has more outcomes than just "yes" and "no" for each comparison.

B. The more meanings a word has, i.e. the greater its polylexy, the greater its chance of being used in a compound. If a word has a tendency to undergo some kind of diversification with each of its meanings (cf. Altmann in this volume) then its participation in the forming of compounds will increase. Thus the following hypothesis is plausible:

> The greater the polylexy of a word the more compounds there are of which it is a component (cf. Rothe in this volume).

C. It is not feasible to assume that a two-stem compound with zero semantic correspondences (as shown in A) tends to incorporate further stems (though this is not impossible). It is more probable with compounds having a higher degree of correspondence. If this assumption is correct then the following hypothesis is plausible:

> The longer a compound, the greater its semantic correspondence with its components.

Practically this means that the average correspondence (i.e. the mean of the frequency distribution in A) monotonically increases with the increase of the length of the compound. The respective mathematical model will be testable only in languages having very long compounds (e.g. German).

D. Compounding is a specification, a narrowing of the (extension of the) meaning. Therefore a compound has in general fewer meanings (lesser polylexy) than its head component and fewer meanings than all its components on the average. This assumption leads to the following hypothesis:

> The longer a compound the fewer meanings it has (on the average).

The model for this hypothesis is a monotonically decreasing curve, most probably identical with that of Menzerath's law (cf. Altmann 1980).

4. Hypotheses concerning length

Since length is related to polylexy in a lawlike manner, it must exert an influence on compounding, too.

A. Since the increase of polylexy leads to shortening of words it is plausible to assume that

> the shorter a word, the more frequently it occurs in compounds.

Similarly, as with other hypotheses, this need not hold for the language as a whole, but possibly merely for a certain word class. (The condition that the language has compounds at all must of course be fulfilled.)

B. The length of the compound affects the length of the components - if Menzerath's law holds (cf. Altmann 1980) - so that we can formulate the hypothesis:

> The longer the compound, the shorter its components.

The length of the compound is expressed as the number of its components (stems or words), the length of the components is measured in terms of the number of syllables, morphemes or phonemes. (The condition that word length in the language is variable must be fulfilled.)

C. The majority of compounds in a language consist of two components, since first those means are exploited that cause the least effort. Only later compounds consisting of more components are built. Therefore it is plausible to assume that

> the number of compounds decreases with their increasing length.

The resulting frequency distribution is monotonically decreasing. The testing of the respective model will be problematic, since in most languages there are scarcely compounds consisting of more than four components.

5. Hypotheses concerning frequency

In Köhler's circuit (Regelkreis), frequency is coupled with polylexy and length, so that it necessarily must influence the compounding.

The more frequently a word occurs, the more meanings it has (depending on the word class), the more frequently it occurs in the neighbourhood of different words, and the greater its tendency to produce compounds. Therefore we can formulate the hypothesis:

> The more frequent a word, the more compounds it produces.

This hypothesis holds only in general. In technical language where the need for specification (cf. Köhler 1986) is very high, frequency is not coupled with polylexy in the same way as in colloquial language. Nevertheless, for testing this hypothesis both technical and general dictionaries as well as long texts can be used, but the parameters of the model may be quite different.

6. Hypotheses concerning cotextuality (polytexty)

Cotextuality, meaning the measure of occurrence in different environments, is coupled with the above properties in Köhler's circuit and allows us to formulate two hypotheses.

A. As a consequence of its relation to the above properties we conclude that

> the greater the cotextuality of a word, the more compounds it produces

since the greater the number of environments or texts in which it occurs the greater its chances to build compounds.

B. If the length of a compound increases, its meaning will be more specific, the polylexy decreases. This leads automatically to the hypothesis that

> the longer a compound the smaller its cotextuality.

7. A hypothesis concerning age

The age of the word has for the time being no place in Köhler's circuit, though Zipf introduced this property into his dynamic view (1949). We can assume that the longer a word exists in a language the more possibilities it has to form compounds, even if this possibility is not equal for all word classes. Yet at least for some word classes it holds that

the older a word the more compounds it produces.

There are two problems associated with this hypothesis. First, the variable "age" cannot easily be measured. The first occurrence in the written language need not be identical with the birthday of the word. For the majority of words not even the century of their emergence can be ascertained. Second, any hypothesis holds only if the *ceteris paribus* condition is fulfilled. As far as age is concerned the *ceteris paribus* condition means that the frequency, the polytexty, the length and the polylexy of the words examined must be similar and that these properties have not changed in time or have changed equally. Otherwise such words as e.g. "computer" would falsify any hypothesis concerning the age of words.

Though the deriving of a mathematical model for this hypothesis is easy, its testability is very low.

8. Conclusion

All these hypothese are candidates for laws, yet in different degree, depending on their deducibility, testability and systematization.

The derivation should not be difficult, since in our assumptions mostly a single independent variable appears. The several hypotheses can be combined to form models with more than one variable.

The testing of the models is more difficult, since for some hypotheses data are not easily obtained and for all of them many languages must be taken into account, which is difficult even for a team of investigators.

The embedding of these hypotheses in a system of equivalent statements is not so problematic, since most of them are associated with the circuit of Köhler and are simple consequences of self-regulation in language.

References

ALTMANN, G. (1980), Prolegomena to Menzerath's law. *Glottometrika* 2, 1–10.

ALTMANN, G. (1988), Diversification processes of the word. In: Köhler, R. (ed.), Studies in Language Synergetics (to appear).

HAKEN, H. (1978), Synergetics. Berlin, Springer.

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

ROTHE, U. (1988), Bedeutungsmenge und Komposition. In: Köhler, R. (ed.), Studies in Language Synergetics (to appear).

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

Hammerl, R. (ed.), Glottometrika 10, 1988.

Kritische Aspekte zum Martinschen Gesetz

Stefan Schierholz, Göttingen

1. Einleitung

Ausgehend von den Untersuchungen Robert Martins, der den Abstraktheitsgrad von Substantiven der französischen Sprache ermittelt hat (Martin 1974), haben Altmann, Kind (1983) ein mathematisches Modell konzipiert, das den Zusammenhang der von Martin erfaßten Daten theoretisch beschreibt. Dieses Modell kann jedoch in seiner jetzt vorliegenden Form nur Gültigkeit für das Französische beanspruchen. Ob der von Altmann, Kind dargestellte Prozeß der Klassenbildung und die Klassenzugehörigkeit eines Begriffs (vgl. Altmann, Kind 1983,1) in gleicher Weise in anderen Sprachen gelten, muß angezweifelt werden.

Diese Zweifel werden durch die von Hammerl (1987) durchgeführte Untersuchung bestätigt, der zusätzlich die Daten der Abstraktheitsbestimmung im Polnischen (Sambor 1982) berücksichtigt hat und zu dem Ergebnis kommt, daß das von Altmann, Kind entwickelte Modell "für die Beschreibung der Daten der polnischen Sprache nicht besonders geeignet" (Hammerl 1987,116) sei.

Außerdem weist Hammerl darauf hin, daß für einen exakten Vergleich der Daten aus verschiedenen Sprachen die Konzeption der benutzten Wörterbücher und das methodische Vorgehen für die Ermittlung der Begriffsketten gleich sein müßte (vgl. Hammerl 1987, 115).

Die erste Forderung scheint vorerst nicht erfüllbar zu sein. Das ist umso bedenklicher, als man feststellen muß, daß die Erhebung der Daten bei der von Martin verwendeten Vorgehensweise entscheidend von der Konzeption des verwendeten Wörterbuchs abhängt (vgl. Schierholz 1988).

Die zweite Forderung ist nur bedingt umsetzbar, weil Martin seine Vorgehensweise unzureichend erläutert hat und zudem einige unverständliche Abgrenzungen getroffen hat.

In der vorliegenden Arbeit werden Daten der deutschen Sprache zur Bestimmung des Abstraktheitsgrades von Substantiven dargestellt. Dabei

wird besonders darauf geachtet, daß klare definitorische Abgrenzungen eine intersubjektive Überprüfbarkeit der Ergebnisse gewährleisten und den Anforderungen der quantitativen Linguistik Rechnung getragen wird (vgl. Altmann 1972, 6 f.).

Die ermittelten Ergebnisse werden an der von Altmann, Kind entwickelten Formel zu den französischen Daten getestet. Abschließend folgt eine Diskussion, die sich weniger mit mathematischen, dafür mehr mit linguistischen Aspekten befassen wird.

2. Die Bestimmung der Abstraktheitsebenen in der deutschen Sprache

Um in dieser Untersuchung zu einer Optimierung des Objektivitätskriteriums zu gelangen, muß Martins Vorgehensweise einigen Modifikationen unterworfen werden.

Die Vervollständigung der Ketten nach eigener Kompetenz (vgl. Altmann, Kind 1983, 2 f.) wird grundsätzlich abgelehnt, weil die Länge der Ketten damit vor allem vom Wortschatz des Untersuchenden abhängen würde. Somit würde der Untersuchungsgegenstand nicht nur das Wörterbuch, sondern auch der Untersuchende selbst sein. Ebenso wäre eine intersubjektive Überprüfbarkeit der Ergebnisse nicht gewährleistet, weil eine andere Person die Ketten wenigstens teilweise auf eine andere Weise komplettieren würde, so daß man allzu intuitive Daten erhielte. Dies ist jedoch mit den Anforderungen der quantitativen Linguistik, die mit meßbaren, möglichst objektiven Ergebnissen aufwarten will, nicht zu vereinbaren.

Martins Abgrenzung der metonymischen Erklärungen kann ebenfalls nicht übernommen werden (vgl. Altmann, Kind 1983, 3), weil keine eindeutige Definition zur Kennzeichnung metonymischer Erklärungen vorliegt. Dies gilt für das Französische wie für die deutsche Sprache.

2.1 Die Ermittlung der Stichprobe

Die Daten zur Bildung der Ketten liefert ein einsprachiges Wörterbuch der deutschen Sprache (Wahrig 1981). Zunächst werden alle Lemmata mit großgeschriebenen Initialen durchnummeriert. Man kommt auf insgesamt 7421 Wörter, von denen 1483 Wörter in die Stichprobe aufgenommen wer-

den (vgl. Schierholz 1982, 17). Die Stichprobe wird in zwei Gruppen zu 742 bzw. 741 Wörtern aufgeteilt, um eine gegenseitige Überprüfung der Ergebnisse zu ermöglichen.

2.2 Die Bildung der Begriffsketten

Die Bedeutungsangaben zu den jeweiligen Lemmata sind im Wörterbuch kursiv gedruckt. Es werden nur die Erklärungswörter, die sich in dem nachfolgenden Substitutionstest unmittelbar auf das jeweilige Lemma beziehen lassen, berücksichtigt.

(Ein/eine/der/die/das) "L" ist (ein/eine/der/die/das) "E".

"L" ist das jeweilige Lemma, "E" das Erklärungswort, bei manchen Wörtern kann der Artikel fehlen.

Beispiel:[1)]

ALLTAG: 1. Tag, der kein Sonntag od. Feiertag ist
2. gleichförmiger Tagesablauf.

Folgende Sätze können mit dem Substitutionstest gebildet werden:

(a) Ein Alttag ist ein Tag.
(b) #Ein Alltag ist ein Sonntag.#
(c) #Ein Alltag ist ein Feiertag.#
(d) Ein Alltag ist ein Tagesablauf.

Die Aussagen (a) und (d) ergeben einen Sinn. "Tag" und "Tagesablauf" können als Erklärungswörter für das Wort "Alltag" markiert werden.[2)] Bei Pluraliatantum wird das Erklärungswort im Singular notiert.

Beispiel:

LEUTE: Personen.

Das Erklärungswort für "Leute" lautet "Person".

2.2.1 Die Wahl der Erklärungswörter

(1) Grundsätzlich wird das erste Erklärungswort einer Bedeutungserklärung, das nach dem vorgestellten Einsetztest ermittelt werden kann, verwendet.

Beispiel:

ABLEGER: Pflanzenteil, Senker,
Zweigunternehmen, Teil.

Dies gilt auch bei Doppelerklärungswörtern (z.B. "Stück Holz") und bei mit Konjunktionen verbundenen Erklärungswörtern ("und", "oder").

Beispiel:

KLOTZ: großes Stück Holz.

(2) Erklärungswörter, die in runden Klammern stehen oder als Synonyme (im Wörterbuch Sy vorangestellt) gekennzeichnet sind, werden berücksichtigt.

Beispiel:

BAD: (Behälter mit) Wasser zum Baden

AAS: ...; Sy Kadaver.

(3) Erklärungswörter nach Gleichheitszeichen und nach Hinweisen auf Begriffserweiterungen [im Wörterbuch mit "– a.)" und "–" gekennzeichnet] werden ebenfalls gezählt.

Beispiel:

ARENA: ... = Manege

ALLEGORIE: ...; – a. Sinnbild.

2.2.2 Der Abbruch der Ketten

(1) Sollte man unter Berücksichtigung der oben aufgestellten Regeln kein Erklärungswort zu einem Substantiv ermitteln können, so ist die Kette an dieser Stelle abzubrechen.

Die Kettenbildung ist ebenfalls beendet, wenn eine Erklärung zirkulär verläuft.

Zirkularität liegt vor, wenn zwei Substantive in einer Kette unmittelbar aufeinander folgen und sich gegenseitig erklären.

Beispiel:

Abteilung - Raum - Teil - Stück - (Teil).

Das letzte Wort "Teil" ist zu streichen, "Stück" bildet das Ende der Kette.

(2) Substantive oder Hinweise in Winkelklammern und orthographische Varianten der Substantive werden nicht als Erklärungswörter berücksichtigt.

Beispiel:

GEMAHL: <...; nicht als Bezeichnung für den eigenen Ehemann verwendet> Ehemann;...

ALLIANZ: ...oV Alliance.

(3) Substantivierungen von Adjektiven oder Verben werden nicht als Erklärungswörter markiert, wenn ihnen "alles", "etwas" oder ein bestimmter bzw. unbestimmter Artikel vorangestellt ist.

Beispiel:

ANFANG: etwas Erstes, Ursprüngliches; Sy Beginn

AUFBRUCH: das Aufbrechen, das Weggehen, Abreise.

(4) Nichtsubstantivische Erklärungen zu einem Lemma wie "jemand, der..." oder "etwas, was ..." werden nicht berücksichtigt.

Beispiel:

ANGEBER: 1 jmd., der einen anderen angibt (anzeigt)
2 jmd., der (beim Spiel) angibt
3 jmd., der angibt, sich wichtig tut.

Treten die Fälle (2), (3) oder (4) auf, so ist die Kette ebenfalls abzubrechen.

2.2.3 Beispiele zur Kettenbildung

Die folgenden Beispiele sollen die Anwendung der oben aufgestellten definitorischen Abgrenzungen verdeutlichen.

Aggression - Angriff - Beginn - Anfang - (Beginn).

Amphibium - Amphibie - Tier - Lebewesen - Organismus - (Lebewesen).

Die Ketten sind jeweils wegen ihrer Zirkularität abzubrechen. Im ersten Fall gibt es vier, im zweiten fünf Ebenen. "Aggression" und "Amphibium" gehören in Ebene i = 1, "Beginn" und "Tier" in Ebene i = 3.

2.3 Die Darstellung der Ergebnisse

Die Ergebnisse der Bestimmung der Abstrakheitsebenen sind in Tabelle 1 eingetragen. Bei der Auszählung haben sich insgesamt 16 Ebenen ergeben. Die Anzahl der Ebenen liegt also wesentlich höher als im Französischen (vgl. Martin 1974, 70, Tab. 2). Die Ursache ist eine Kette, die sehr häufig vorkommt:

Verbindung - Zustand - Beschaffenheit - Natur - Welt - - Gesamtheit - Ganzes - Einheit - Zusammenhang - (Verbindung).

Wenn ein beliebiges Wort in den unteren Ebenen eines dieser neun Wörter als Erklärungswort hat, so muß anschließend die ganze Wortkette

durchlaufen werden, bis es zu der zirkulären Definition kommt. In jedem Fall sind zu dem Auftreten eines dieser Wörter immer acht Ebenen dazuzuzählen.

Die Auftellung der Stichprobenmenge in zwei fast gleichgroße Gruppen hat sehr homogene Ergebnisse erbracht. Die größte Differenz erscheint auf Ebene i = 5 mit 10 Wörtern. Die Abnahme der Werte von Ebene i zu Ebene i+1 ist jedoch in beiden Zählungen sehr gleichförmig. Im Vergleich zu den Ergebnissen Robert Martins erfolgt die Abnahme aber

Tabelle 1

Anzahl der Wörter auf einzelnen Abstraktheitsebenen

i	N_i	N_i-N_{i+1}	A_{1i}	A_{2i}
1	1483	444	742	741
2	1039	596	586	578
3	443	251	267	273
4	192	93	128	133
5	99	50	65	75
6	49	19	37	38
7	30	10	24	24
8	20	5	18	17
9	15	3	13	13
10	12	3	11	10
11	9	-	9	8
12	9	1	8	8
13	8	2	7	7
14	6	-	4	3
15	6	4	4	2
16	2	2	2	-

i = *Abstraktheitsebene*

A_{1i} = *Anzahl der Wörter aus der ersten Gruppe der Stichprobe*

A_{2i} = *Anzahl der Wörter aus der zweiten Grupppe der Stichprobe*

N_i = $A_{1i} + A_{2i}$

N_i-N_{i+1} = *Anzahl der Wörter, wenn jedes Wort nur einmal pro Ebene auftritt.*

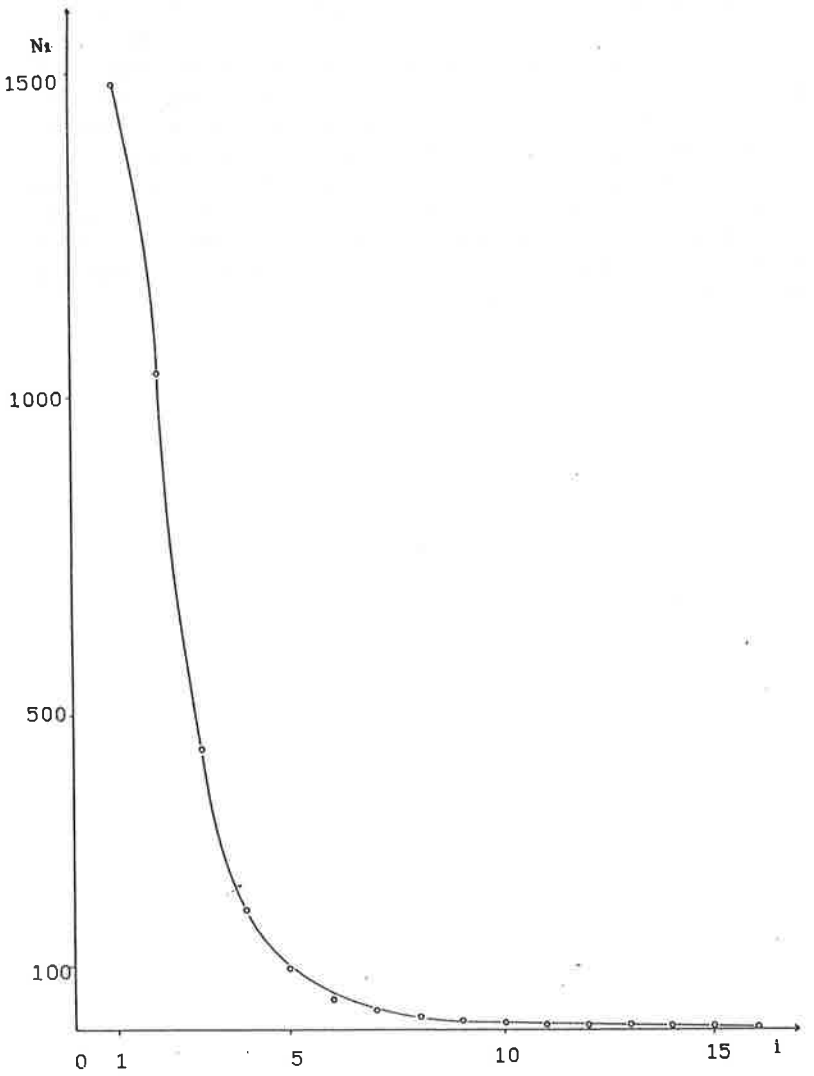


Abb. 1. Graphische Darstellung der Wortanzahl pro Abstraktheitsebene (i = Abstraktheitsebene, Ni = Anzahl der Wörter).

wesentlich langsamer. In Abbildung 1 sind die Ergebnisse der Tabelle 1 graphisch dargestellt. Die Kurve verläuft zwischen den Werten von i = 1 bis i = 2 konvex und von i = 2 bis i = 16 konkav. Theoretisch ist der Wert N_1 also zu klein. Dieser Sachverhalt wird durch Abbildung 2 noch deutlicher hervorgebracht. Das Histogramm zeigt in der Ebene i = 1 weniger Wörter als in der Ebene i = 2. Die Werte sind durch die Subtraktion der Werte N_{i+1} von N_i errechnet, aber es ist natürlich nicht möglich, daß es weniger Lemmata N_1 als Erklärungswörter N_2 gibt, wenn pro Stichwort nur ein Erklärungswort gezählt wird. Zwar würde eine größere Stichprobenmenge andere Ergebnisse liefern; man kann jedoch nicht sagen, daß die Stichprobenmenge mit zwanzig Prozent zu klein sei, sondern die Anzahl der Erklärungswörter N_2 ist zu groß.

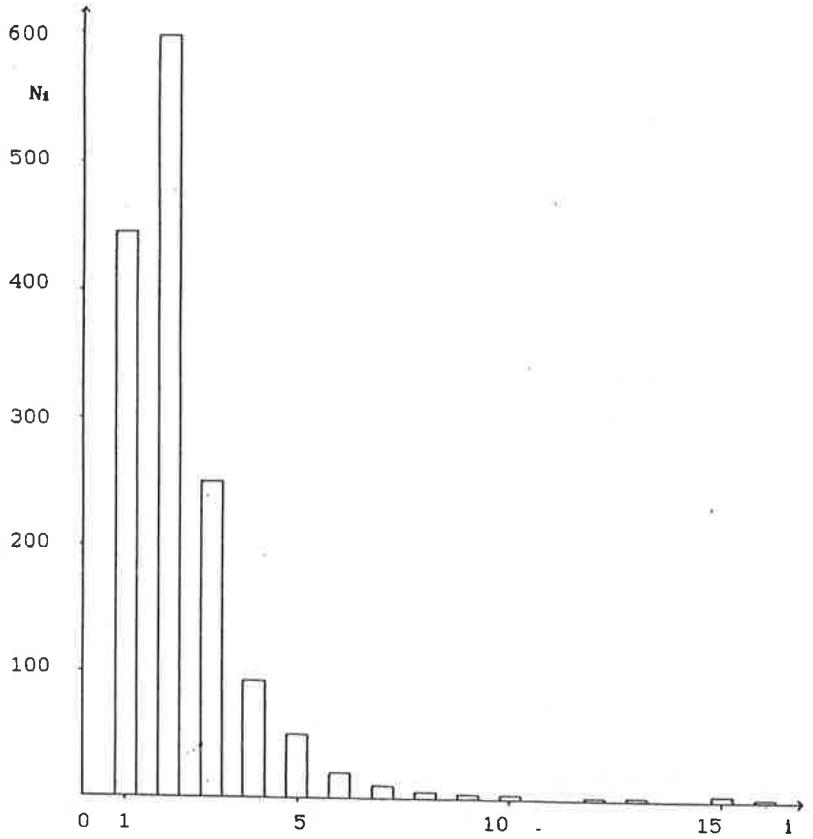


Abb. 2. Histogramm der verschiedenen Wörter nach Abstraktheitsebenen (i = Abstraktheitsebene, N_i = Anzahl der Wörter).

2.4 Die mathematische Beschreibung

In der Überprüfung der Ergebnisse wird zu zeigen sein, ob die Anordnung der Substantive der von Altmann, Kind (1983, 4 ff.) entwickelten Funktionsgleichung über die Klassifikationstendenz der Sprache entspricht.

Altmann, Kind sind von der Annahme ausgegangen, daß sich die Zahl der Oberbegriffe in der Ebene i + 1 proportional zu der in Ebene i vermindert und gleichzeitig die Anzahl der Wörter auf einer Ebene proportional zu der Höhe der Ebene ist.

Aus der Zusammensetzung dieser beiden Annahmen entwickeln Altmann, Kind (1983, 4) die allgemeine Formel:

$$N_i = N_1 i!a^{i-1} \quad . \qquad (1)$$

In dieser Gleichung sind N_1 und der Proportionalitätskoeffizient a die Konstanten; N_1 wird geschätzt als die Anzahl der Wörter in der ersten Ebene. Für die Berechnung der Konstante a bestehen mehrere Möglichkeiten, von denen zunächst die erste angewendet werden soll (vgl. Altmann, Kind 1983, 5):

$$a^* = \frac{N_2}{2N_1}. \qquad (2)$$

Setzt man aus der Stichprobe die Werte N_1 und N_2 ein, so erhält man:

$$a^* = \frac{1039}{2(1483)} = 0.3503034 \quad .$$

Nimmt man $N_1^* = 1483$ und $a^* = 0.3503034$ und setzt diese in die Formel (1) ein, so lauten die Ergebnisse wie in Tabelle 2 angegeben.

Tabelle 2

Beobachtete und berechnete Anzahl der Wörter auf einzelnen Abstraktheitsebenen nach Formel (1)

i	Ni	
	beobachtet	berechnet
1	1483	1483.0000
2	1039	1038.9998
3	443	1091.8949
4	192	1529.9787
5	99	2679.7750
.	.	.
.	.	.
.	.	.

Eine Berechnung weiterer Werte kann man sich ersparen. Die Werte steigen mit zunehmendem i an, statt zu fallen. Die große Abweichung der Werte ist mit einem zu hohen Wert für a zu erklären. Darum soll eine andere Schätzung von a nach folgender Formel durchgeführt werden (Altmann, Kind 1983, 6):

$$\ln a^* = \frac{\sum_i (i-1)\ln N_i - \ln N_1 \sum_i (i-1) - \sum_i (i-1)\ln(i!)}{\sum_i (i-1)^2}. \quad (3)$$

Die Berechnung der einzelnen Größen ergibt:

$$\Sigma(i-1)\ln N_i = 288.0208$$

$$\Sigma(i-1) = \frac{16(17)}{2} - 16 = 120$$

$$\ln N_1(120) = 876.2187$$

$$\Sigma(i-1)\ln(i!) = 1603.735282$$

$$\sum_{i=1}^{16} (i-1)^2 = \sum_{i=1}^{15} i^2 = \frac{15\,(16)\,(31)}{6} = 1240.$$

Beim Einsetzen in die Formel (3) erhält man:

$$\ln a^* = \frac{288.0208 - 876.2187 - 1603.735282}{1240} = -1.767688$$

$$a^* = 0.1707.$$

Mit diesem neuen Wert für a ergibt das Einsetzen in die Formel (1) folgendes Ergebnis (vgl. Tabelle 3):

Tabelle 3

Beobachtete und berechnete Anzahl der Wörter auf einzelnen Abstraktheitsebenen

i	Ni	
	beobachtet	berechnet
1	1483	1483.0000
2	1039	506.2962
3	443	259.2735
4	192	177.0310
5	99	151.0880
6	49	154.7444
7	30	184.6157
8	20	251.1372
.	.	.
.	.	.
.	.	.

Auch hier kann die Berechnung abgebrochen werden, weil die errechneten Werte ab N_5 wieder ansteigen. Wegen der großen Diskrepanzen zwischen den beobachteten und errechneten Werten sollen weitere Berechnungen ausbleiben.

3. Die Diskussion der Ergebnisse

Als Ursache für die nicht erfolgte Bestätigung des von Altmann, Kind formulierten mathematischen Zusammenhangs können die Fehlerquellen in folgenden Bereichen gesucht weren.

(1) Martins Gesetz der Abstraktheitsebenen muß modifiziert werden.
(2) Die Art der Datenerhebung, die im Vergleich zum Französischen einigen Veränderungen unterworfen war, ist fehlerhaft gewesen.
(3) Das Modell gilt nicht für die deutsche Sprache.
(4) Das benutzte Datenmaterial ist für derartige Untersuchungen ungeeignet.

3.1 Martins Gesetz der Abstraktheitsebenen

In der Untersuchung Hammerls zum Martingesetz wird darauf hingewiesen, daß das mathematische Modell von Altmann, Kind als ein Spezialfall der Formel

$$N_{i+1} = \frac{c}{(i+1)^a} N_i^b$$

anzusehen ist (Hammerl 1987, 119). Der Anlaß zu der Überarbeitung sind die Daten zur polnischen Sprache, bei denen eine große Anzahl an Ebenen auftritt, und die Abnahme der Wortanzahl von N_1 zu N_2 wesentlich langsamer als im Französischen erfolgt (vgl. Tab. 4).

Da diese Tendenz auch bei den Daten zur deutschen Sprache zu beobachten ist, könnte man daraus folgern, daß die von Altmann, Kind entwickelte Formel nur für das Französiche gilt und modifiziert werden muß, wenn auch andere Sprachen damit beschrieben werden sollen.

Tabelle 4

Zahl der Wörter auf den einzelnen Abstraktheitsebenen nach Martin (1974, 70), Sambor (Hammerl 1987, 115) und Schierholz

Ebene i	Zahl der Wörter Ni		
	Französisch	Polnisch	Deutsch
1	1723	1000	1483
2	348	618	1039
3	108	217	443
4	39	110	192
5	13	44	99
6	3	16	49
7		9	30
8		3	20
9		1	15
10			12
11			9
12			9
13			8
14			6
15			6
16			2

Andererseits dienen die theoretischen Formulierungen von Altmann, Kind, die die Relation zwischen zunehmender Abstraktheitsebene und abnehmender Wortanzahl beschreiben, dazu, die Richtigkeit von Analysen zu prüfen. Bei einem negativen Ergebnis sind die Fehler in den Datenmengen oder Datenerhebungen zu suchen. In diesem Fall wird das auch dadurch unterstützt, daß die Schätzung des Proportionalitätskoeffizienten a mit zwei verschiedenen Methoden zu sehr großen Differenzen geführt hat (0.1707 und 0.3503034).

Bei der Beschreibung der Daten der französischen Sprache ergeben sich für a wesentlich homogenere Werte (0.100987; 0.084856; 0.100329) (vgl. Altmann, Kind 1983, 6 ff.). Aus diesem Grunde kann die Ursache für die gescheiterten Rechenversuche auch bei den Kriterien zur Datengewinnung oder in dem verwendeten Datensatz vermutet werden .

3.2 Die Datenerhebung

Die in der Untersuchung zur deutschen Sprache modifizierten definitorischen Abgrenzungen sind oben erläutert und begründet worden (vgl. Kap. 2.2). Davon haben sich in der Untersuchung zur polnischen Sprache vor allem die folgenden Regelfestlegungen unterschieden:[3)]

- Ist kein Erklärungswort des zu erklärenden Wortes vorhanden, wird die Kette nach eigener Kompetenz vervollständigt.
- Es ist zu beachten, daß das Wort in der Ebene i+1 immer ein genus proximum des Wortes in Ebene i ist.
- Die Erklärungswörter innerhalb einer Kette müssen für alle Wörter auf den darunter befindlichen Ebenen als Bedeutungserklärung fungieren können.

Durch diese Regeln kann sich vor allem die Länge der gebildeten Ketten erheblich verändern. Dazu einige Beispiele aus der Untersuchung zur deutschen Sprache:

(a) *Nachsicht – Geduld – Fähigkeit –.*

Diese Kette wird abgebrochen, weil "Fähigkeit" im benutzten Wörterbuch nicht lemmatisiert ist. Eine Vervollständigung nach eigener Kompetenz könnte wie folgt aussehen:

(b) ... – *Fähigkeit – Eigenschaft – Merkmal – (Eigenschaft).*

Die Substantive können alle Substantive in den unter ihnen liegenden Ebenen erklären. Das Ende der Kette kommt durch Zirkularität zustande.

(c) *Kraftwerk – Anlage – Tätigkeit – Handeln.*

Die Kette wird abgebrochen, weil "Handeln" nicht als Substantiv im Wörterbuch steht. Die vier Wörter würden in die Ebenen i = 1 bis i = 4 eingeordnet. Unter den oben geschilderten Modifikationen müßte man die Kette verändern, weil ein "Kraftwerk" zwar eine "Anlage", aber keine "Tätigkeit" oder ein "Handeln" ist. Eine Umbildung könnte folgendermaßen geschehen:

(d) *Kraftwerk – Anlage – Bau – Gebäude – Bauwerk – (Bau).*

Diese Beispiele zeigen, daß sich die Ketten teilweise erheblich umgestalten lassen. Durch Beispiel (c) wird klar, daß die oberen Ebenen (etwa ab i=10) wahrscheinlich verschwinden, weil man keine Erklärungswörter finden wird, die alle Wörter in den 14 oder 15 darunter liegenden Ebenen erklären können. In den mittleren Ebenen (etwa i=3 bis i=9) wird die Anzahl der Wörter abnehmen, wenn die Ketten entsprechend Beispiel (c) nach dem Wort "Anlage" abgebrochen und auf eine Vervollständigung nach eigener Kompetenz verzichtet werden würde, denn durch letzteres werden die Ketten wie in Beispiel (b) verlängert. Dies darf jedoch nicht überwertet werden, weil die Zahl der verschiedenen Wörter kaum zunehmen könnte. Das Wort "Eigenschaft" (Ebene i=4) würde auch in anderen Ketten auf dieser Ebene existieren, so daß die Zunahme an neuen Wörtern sehr gering wäre.

Jedoch ist ein anderes Problem der Kettenbildung an diesem Beispiel deutlich erkennbar. Statt der in Beispiel (b) vorgeschlagenen Vervollständigung der Kette kann man nach dem Wort "Fähigkeit" auch folgendermaßen fortfahren:

(e) ... – *Vermögen – (Fähigkeit)*
(f) ... – *Anlage – Veranlassung – Begabung – (Anlage)*
(g) ... – *Befähigung – Vermögen – Können – Leistung –*
– Anstrengung – Aufwand – ... .

Hier läßt sich nicht entscheiden, welche Kette "richtig" oder "besser" ist, so daß die Länge der Kette ausschließlich von dem subjektiven Urteil des Untersuchenden abhängt. Besondere Schwierigkeiten entstehen bei Wörtern mit wertenden Inhalten ("Leistung", "Anstrengung"). An den vorgestellten Möglichkeiten (e), (f) und (g) läßt sich leicht verdeutlichen, daß es differierende Auffassungen darüber gibt, ob "Geduld" und "Nach-

sicht" eine "Leistung" oder eine "Anstrengung" sind. Da aber die Kriterien zur unterschiedlichen Interpretation dieses Sachverhalts nicht klar zu umreißen sind, kann nur der persönliche Erfahrungsbereich des Untersuchenden eine Entscheidungshilfe sein. Letzteres zu erforschen ist jedoch nicht das Ziel der hier angestellten quantitativen Untersuchung.

Ähnliche Probleme sind bei der Bestimmung des genus proximum zu erwarten. Zwar bestehen die meisten Bedeutungserklärungen in Wörterbüchern u. a. aus genus proximum und differentia specifica, aber wegen der Polysemie vieler Lemmata obliegt es der persönlichen Auswahl des Untersuchenden, welches Wort als genus proximum zu bestimmen ist.

Veränderungen in den definitorischen Abgenzungen modifizierten somit die Daten nur geringfügig. Die Anzahl der Wörter pro Ebene wäre jedoch immer noch erheblich höher als bei Martin, so daß eine Bestätigung des mathematischen Modells von Altmann, Kind unwahrscheinlich bleibt.

Es ist aber auch denkbar, daß die Daten der deutschen bzw. polnischen Sprache den Prozeß der Klassenbildung und die Klassenzugehörigkeit eines Begriffs richtig wiedergeben. Dann hätten Altmann, Kind die Daten der französischen Sprache zu ernst genommen, und das Modell bedarf einer Modifikation.

Bei Martins Untersuchungsmethode kann man vermuten, daß er seine Ketten allzuoft nach eigener Kompetenz komplettiert hat und auf den höheren Ebenen erheblich weniger Wörter gefunden hat, weil der Wortschatz einer Einzelperson natürlich kleiner ist als der eines Wörterbuchs. Somit hätte Martin mehr seine eigene Sprachkompetenz als die Substantive des Wörterbuchs untersucht.

Hier müßten psycholinguistische Experimente, in denen Versuchspersonen, die Begriffsketten ohne die Zuhilfenahme eines Wörterbuchs bilden, Aufschluß darüber geben können, ob Martins Daten stärker mit der Begriffskettenbildung von Individuen als mit den Ergebnissen aus anderen Sprachen korrelieren.

3.3 Unterschiede in der Substantivbildung

Ein Typikum der deutschen Sprache ist die Möglichkeit der (fast beliebigen) Komposition von Substantiven,[4] die dazu führen kann, daß das Modell Altmann, Kinds in der deutschen Sprache nicht gültig ist.

Im Polnischen und Französischen werden statt Komposita nominale Satzglieder mit Genitivattribut formuliert, z.B. wird im Französischen

"bras" mit "partie du corps" erklärt; im Deutschen benutzt man stattdessen für "Arm" "Körperteil" als Erklärungswort. Genauso werden aus "Teil der Maschine" und "Teil der Pflanze" im Französischen "Maschinenteil" und "Pflanzenteil" im Deutschen. Addiert man für diese drei Beispiele die Anzahl der Erklärungswörter, so erhält man im Deutschen drei Wörter, im Französischen aber nur ein Wort. Die vielen verschiedenen Komposita führen dazu, daß sich nur wenige Wörter in einer Ebene wiederholen und sehr viele unterschiedliche Wörter auf den einzelnen Ebenen auftreten.

Diesem Phänomen kann man durch Auflösung der Komposita begegnen. Dabei kann das jeweils semantisch relevante Wort für die Kettenbildung berücksichtigt werden, so daß in den genannten Beispielen ("Körperteil", "Maschinenteil", "Pflanzenteil") dreimal das Grundwort "Teil" zu verwenden wäre.

Diese Operation würde zu einer drastischen Abnahme der Wortzahl pro Ebene führen. Zwar würden viele intuitive Entscheidungen bei der Auflösung benötigt; entsprechende Definitionen zur Kompositaauflösung könnten dabei aber möglicherweise Abhilfe schaffen.[5]

Von größerer Bedeutung ist jedoch die Frage, ob man durch die Auflösung der Komposita die deutsche Sprache nicht eines ihrer wesentlichen Merkmale berauben würde. Derartige Prozeduren können letztlich dazu führen, die gewonnenen Daten der Theorie anzupassen und damit eine Untersuchung wertlos zu machen.

3.4 Der Untersuchungsgegenstand

Leider war vor Beginn dieser Untersuchung nicht bekannt, welche eklatanten Mängel das Untersuchungsobjekt, das dtv-Wörterbuch von Gerhard Wahrig aufweist (vgl. Schierholz 1988, 466 ff.).

Betrachtet man den gesamten Wortschatz der Substantive, die in den Bedeutungserklärungen des Wörterbuchs als Lemma oder als Erklärungswort auftreten, so ist die Zahl der Erklärungswörter, die selbst nicht erklärt werden, größer als die Zahl der lemmatisierten Erklärungswörter. Dabei soll bei der Konzipierung des Wörterbuchs gerade auf diesen Sachverhalt besonderer Wert gelegt worden sein (vgl. Wahrig 1981, 6).

Da sehr häufig vorkommende Wörter wie "Darstellung" und "Fähigkeit"[6] im Wörterbuch nicht lemmatisiert sind, müssen alle Ketten, in denen diese Substantive erscheinen, abgebrochen werden. Dies fiele noch deutlicher ins Gewicht, wenn in der Untersuchung die Ketten nach eige-

ner Kompetenz vervollständigt worden wären, weil davon alle Wörter, die nicht im Lexikon als Stichwörter eingetragen sind, betroffen wären. Im Endeffekt hätte das zu einer Verlängerung der Ketten geführt.

Es ist somit festzuhalten, daß die derzeit vorliegenden Ergebnisse durch die Wahl eines Wörterbuchs, das seinen eigenen Ansprüchen nicht gerecht wird, erheblich beeinträchtigt worden sind. Gerade die selbstgestellten Ansprüche hatten aber zu der Wahl eben dieses Wörterbuchs als Datenbasis geführt.

4. Schlußbemerkungen

Geht man trotz der erwähnten Mißerfolge bei der Formulierung eines gesetzmäßigen Zusammenhangs zwischen den Abstraktionsebenen von der Annahme aus, daß Sprache in all ihren Erscheinungsformen Gesetzen folgt, so ergeben sich aus den bisherigen Überlegungen einige Konsequenzen:

Für eine erneute Untersuchung in der deutschen Sprache muß möglichst ein den genannten Anforderungen besser entsprechendes Wörterbuch gefunden werden, und die Methode der Datenermittlung bedarf einiger Modifikationen.[7] So ist unbedingt darauf zu achten, daß die Substantive auf den höheren Ebenen alle Substantive in den darunter liegenden Ebenen erklären können. Einer Vervollständigung der Begriffsketten aus eigener Kompetenz kann prinzipiell nicht zugestimmt werden; die Kriterien zur Bildung der Ketten müssen eindeutig operationalisierbar sein, um dem Anspruch der intersubjektiven Überprüfbarkeit gerecht zu werden.

Falls jedoch auch nach der Erarbeitung eines deutlich verbesserten Datensatzes weiterhin Probleme mit dem Gesetzesvorschlag von Altmann, Kind bestehen bleiben, erscheint eine Neufassung des semantischen Gesetzes erforderlich zu sein.

Anmerkungen

1) Die Erklärungswörter, die berücksichtigt werden, sind in den Beispielen jeweils unterstrichen.
2) Zwar beruht die Beurteilung der Sinnhaftigkeit der gebildeten Sätze im Einsetztest auf der subjektiven Entscheidung des Untersuchenden, aber der Test läßt sich bei der überwiegenden Mehrzahl der Fälle problemlos durchführen, so daß eine intersubjektive Überprüfung gewährleistet ist.
3) Mündliche Mitteilung von G. Altmann.
4) Vgl. Zepic (1970, 15): "Das Hauptmerkmal des deutschen Vokabulars ist die beinahe unbegrenzte Möglichkeit, Zusammensetzungen zu bilden. Von rund 5000 Stammorphemen (...) läßt sich in der deutschen Sprache durch Komposition und Derivation eine in die Hunderttausende gehende Anzahl von Wörtern produzieren."
5) In den meisen Fällen wird das semantisch relevante Wort das zweite des Kompositums sein. Jedoch bleiben viele Problemfälle, wie die folgenden Beispiele zeigen: *Bauwerk, Freundeskreis, Kraftwerk, Lebewesen, Merkmal.*
6) Zu allen lemmatisierten Substantiven taucht "Fähigkeit" 42 mal und "Darstellung" 35 mal auf. Beide Worter gehören damit zu den 40 häufigsten Substantiven unter 12612 verschiedenen Erklärungswörtern.
7) Eine solche Untersuchung wird im Moment im Rahmen eines Promotionsvorhabens durchgeführt.

Literatur

Altmann, G. (1972), Status und Ziele der quantitativen Sprachwissenschaft. In: Jäger, S. (Hrsg.): *Linguistik und Statistik*, Braunschweig, Vieweg 1–9.

Altmann, G. , Kind, B. (1983), Ein semantisches Gesetz. *Glottometrika* 5, 1–13.

Hammerl, R. (1987), Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen. *Glottometrika* 8, 113–129.

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

Sambor, J. (1982), Lexikographische Definitionen. Bochum (unveröffentlichte Sammlung von 1000 Begriffsketten für die polnische Sprache unter Ausnutzung folgenden Wörterbuchs: Skorupka, S., Auderska, H., Łempicka, Z., Mały słownik jezyka polskiego. Warszawa, Panstwowe Wydawnictwo Naukowe 1968).

Schierholz, S. (1982), Untersuchungen zur Polysemie im Deutschen (Unveröffentlichte schriftliche Hausarbeit im Rahmen der fachwissenschaftlichen Prüfung für das Lehramt an Gymnasien). Göttingen.

Schierholz, S. (1988), Bedeutungswörterbücher als Grundlage empirischer Wortschatzuntersuchungen. *Germanistische Linguistik* 87-90, 463-478.

Wahrig, G. (Hrsg.) (1981), Wörterbuch der deutschen Sprache. 4. Aufl., München, dtv.

Žepić, S. (1970), Morphologie und Semantik der Nominalkomposita. Zagreb, Izdavački zavod Jugoslavenske akademije znanosti i umjetnosti.

Hammerl, R. (ed.), Glottometrika 10,1988.

Neue Perspektiven der sprachlichen Synergetik: Begriffsstrukturen - kognitive Gesetze*

Rolf Hammerl, Bochum/Warschau

1. Vorbemerkungen

In der traditionellen Sprachwissenschaft ist man bemüht, ein solches Begriffs- und Methodeninventar auszuarbeiten, welches für die Kennzeichnung und Deskription sprachlicher Erscheinungen geeignet erscheint. Solche Deskriptionen stellen jedoch lediglich den ersten Schritt wissenschaftlicher Erkenntnis dar, da diese erst über die Erklärung der Wirkungsmechanismen der Sprache gegeben ist, die für das Zustandekommen und die Veränderung sprachlicher Erscheinungen verantwortlich sind. Gerade dieser Grundgedanke findet seine Rechtfertigung in der Synergetik, einer Forschungstheorie, die heute sowohl in den Naturwissenschaften als auch in den Gesellschaftswissenschaften Anwendung findet. Es werden systeminterne Strukturbildungsprozesse untersucht, die sowohl deterministische als auch stochastische Prozesse sein können und gesetzesmäßig ablaufen; diese Prozesse müssen als Resultat des komplexen Zusammenwirkens von systeminternen und -externen Faktoren interpretiert werden, als Selbstregulationsprozesse, die in ihrer Summe einen solchen Systemzustand anstreben, der die optimale Erfüllung der Funktionen dieser Systeme garantiert. Auf der Grundlage dieser Erkenntnis können Modelle geschaffen werden, die den Einfluß der für die Erklärung der jeweiligen Erscheinung wesentlichen Faktoren berücksichtigt. Dies wiederum kann nur über die Anwendung exakter mathematischer Methoden geschehen (z.B. der Wahrscheinlichkeitsrechnung, Differential- und Integralrechnung).

Selbstregulationsprozesse, die im System "Sprache" ablaufen, sind u.a. abhängig von menschlichen Bedürfnissen (z.B. Sicherung des Informa-

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

tionsaustausches zwischen Sprecher und Hörer mit geringster Anstrengung bei gleichzeitiger Gewährleistung einer für die Aufrechterhaltung der sprachlichen Kommunikation optimalen Redundanz), welche sich wiederum in systeminternen Ordnungsparametern niederschlagen (z.B. Parameter der Redundanz einer Sprache) und natürlich von bestimmten Systemgrößen selbst (z.B. Buchstaben- oder Phoneminventar einer Sprache) abhängig sind.

Die Anwendung der synergetischen Forschungsmethode in der Linguistik wird im Aufsatz von Köhler und Altmann (1986) begründet, wobei sich diese Autoren auf eine Reihe von Voruntersuchungen stützen konnten, die vor allem die Modellierung und Validierung einzelner Sprachgesetze betreffen (z.B. das Menzerathsche Gesetz; vgl. Altmann 1980; Altmann, Beöthy, Best 1982), die Untersuchung bestimmter Ordnungsparameter (z.B. den Zipfschen Umfang; vgl. Orlov, Boroda, Nadarejsvili 1982).

Die untersuchten Sprachgesetze betreffen sowohl Prozesse (z.B. die diachronischen Untersuchungen von Altmann, von Buttlar, Rott, Strauß 1983) als auch bestimmte Prozeßstufen (Zustände). Im zweiten Falle müssen zunächst Verteilungsgesetze (z.B. Grotjahn 1982; Krylov 1982; Beöthy, Altmann 1983) von Gesetzen unterschieden werden, die die gegenseitigen Beziehungen von jeweils 2 sprachlichen Eigenschaften untersuchen. Auch hier kann man hinsichtlich der untersuchten Eigenschaften eine große Zahl unterschiedlicher Gesetze unterscheiden:

- Abhängigkeiten zwischen zwei Texteigenschaften (z.B. zwischen Satz- und Clauselänge; vgl. Heups 1983),
- Abhängigkeiten zwischen einer Texteigenschaft und einer Eigenschaft des Sprachsystems (z.B. zwischen Lexemlänge und mittlerer Bedeutungszahl im Text)
- Abhängigkeiten zwischen zwei Systemeigenschaften (z.B. zwischen der Lexemlänge und der Bedeutungszahl im Wörterbuch, vgl. Altmann, Beöthy, Best 1982).

Köhler (1986) konnte als erster den Selbstregulationsmechanismus zwischen den 4 Eigenschaften *Länge, Frequenz, Polylexie* und *Polytextie* aus einem allgemeinen synergetischen Modell heraus beschreiben und an deutschem Material überprüfen. Somit war der Grundstein gelegt für die Untersuchung von Selbstregulationssprozessen bei gleichzeitiger Berücksichtigung einer weit größeren Zahl von Eigenschaften sprachlicher Einheiten. Diese Untersuchungen werden innerhalb des von G. Altmann und R. Köhler geleiteten Forschungsprojektes "Sprachliche Synergetik" an der Ruhr-Universität Bochum realisiert.

In unserem bisherigen Kurzüberblick wurden Untersuchungen zum Martingesetz (Martin 1974, Altmann, Kind 1983) noch nicht berücksichtigt, da sie weder Text- noch Lexikoneigenschaften betreffen, sondern die wesentlich abstrakteren Strukturgesetze sprachlicher Begriffe, die als kognitive Strukturen interpretiert werden können (im Rahmen eines wesentlich allgemeineren synergetischen Modells bezüglich des Modells von Köhler 1986). Bevor wir uns aber diesem Problem zuwenden, wollen wir auch einen kurzen Überblick über die bisherigen Untersuchungen zum Martingesetz geben.

2. Kurzüberblick über die bisherigen Untersuchungen zum Martingesetz der Abstraktionsebenen

Der französische Forscher Martin (1974) ging davon aus, daß das in Bedeutungsdefinitionen einsprachiger Wörterbücher angegebene "genus proximum" Oberbegriff des zu definierenden Begriffes ist und daß man somit für jeden gewählten Ausgangsbegriff alle Oberbegriffe finden kann, bis zu einem Endbegriff, der am abstraktesten ist und nicht mehr weiter über die Angabe von Oberbegriffen definiert werden kann. Er nahm an, daß die Zahl der somit gefundenen Begriffe bezüglich der Zahl der 1., 2. usw. Unterbegriffe nicht nur vom Zufall bestimmt wird. Jedoch erst Altmann und Kind (1983) waren in der Lage, diese Relation zu modellieren und die abgeleitete mathematische Abhängigkeit an den von Martin (1974) gelieferten Daten zu überprüfen. Alle Forscher, die sich der Untersuchung und Überprüfung des somit gefundenen Martingesetzes der Abstraktionsebenen zugewendet haben, mußten sehr viele methodologische Einzelprobleme für die Suche von Oberbegriffen in den Bedeutungsdefinitionen der entsprechenden Wörterbücher lösen (vgl. Altmann, Kind 1983; Schierholz 1982; Sambor 1983; Hammerl 1987b). Deshalb sind auch die empirischen Daten zum Martingesetz für die deutsche und polnische Sprache (vgl. Sambor 1983; Hammerl 1988b) nicht mit denen der französischen Sprache vergleichbar. Schierholz (1982) dagegen hat eine solche Untersuchungsmethode angewandt, die die Relation "Hyponymie - Synonymie" nicht immer streng unterscheiden konnte, was natürlich zu völlig anderen Untersuchungsergebnissen führte (die natürlich auch nicht das Martingesetz der Abstraktionsebenen betreffen).

Es zeigte sich auch bald, daß das von Altmann, Kind (1983) vorgeschlagene Modell zur Beschreibung der Funktion $y = f(x)$ zwischen der Zahl der Begriffe y_x auf der Ebene x ($y_x = y_1$: Zahl der Ausgangsbe-

griffe; $y_x = y_2$: Zahl der Oberbegriffe der Ausgangsbegriffe usw.) und dem Index x dieser Ebene modifiziert werden muß. Es konnte zwar ein mathematisches Modell abgeleitet werden (vgl. Hammerl 1987a), welches diesen Zusammenhang für alle drei bisher untersuchten Sprachen (Französisch, Polnisch, Deutsch) gut beschreibt, jedoch war die gefundenene Abhängigkeit recht kompliziert und konnte theoretisch nicht zufriedenstellend validiert werden; außerdem zeigte sich in empirischen Untersuchungen, daß die (auf die oben beschrtiebene Weise) gebildete Begriffsstruktur modifizierungsbedürftig ist (vgl. Hammerl 1988 a,b); z.B. konnten die somit unterschiedenen Begriffsebenen (x = 1,2,...) nicht als Abstraktionsebenen interpretiert werden, da schon unter den Ausgangsbegriffen abstrakte neben konkreten Begriffen auftreten (und natürlich bezüglich konkreter Begriffe in abnehmender Zahl auch auf den weiteren Abstraktionsebenen).

Erst eine Neuordnung des empirischen Materials nach Kriterien, die aus einem synergetischen Modell abgeleitet wurden, welches Begriffsstrukturen als kognitive Strukturen interpretiert, konnte die genannten Schwierigkeiten überwinden.

3. Modell synergetischer Sprachgesetze

Wir wollen hier ein einfaches synergetisches Modell der Informationsbildung und -verarbeitung vorstellen, um daraus Konsequenzen für die Untersuchung von Begriffsstrukturen und deren Interpretation als kognitive Strukturen abzuleiten.

Es ist bekannt, daß die Dekodierung der vom Menschen aufgenommenen Informationen nach bestimmten - aus sprachinternen und -externen Faktoren resultierenden - Interpretationsregeln (-gesetzen) erfolgen muß, um das Funktionieren eines solch effektiven und komplizierten Informationsverarbeitungs- und Informationsbildungssystems überhaupt zu gewährleisten. Hierbei sind natürlich auch Sprachgesetze selbst von besonderer Bedeutung. Auf der ersten Interpretationsstufe finden Textgesetze Anwendung, die ja direkt an den zu interpretierenden Text anknüpfen, auf einer qualitativ höheren Stufe Verteilungsgesetze über Einheiten des Textes oder Lexikons (hier ist mindestens eine Eigenschaft eine reine Struktureigenschaft und kann nicht unmittelbar als Texteigenschaft interpretiert werden), und auf der höchsten Interpretationsstufe finden kognitive Gesetze Anwendung, die nur von bestimmten abstrakten

sprachlichen Eigenschaften beeinflußt werden, aber keinen direkten Zusammenhang mit Text- und Verteilungsgesetzen haben (müssen).

Bei Gültigkeit der Hypothese, daß die semantische Repräsentation sprachlicher Einheiten (lexikalischer Einheiten) im Langzeitgedächtnis in n-dimensionalen Begriffsnetzen erfolgt, wo sprachliche Begriffe hinsichtlich von n Eigenschaften charakterisiert und entsprechend eingeordnet werden, können die Begriffe dieses Raumes nach jeder der n Eigenschaften geordnet werden, d.h. es können mindestens n Begriffsstrukturen gebildet werden. Diese Begriffsstrukturen werden im Prozeß der sprachlichen Kommunikation aktiviert, beeinflussen und gestalten die gebildeten Kommunikate (auf Sprecherseite) und werden selbst wieder von diesen beeinflußt (auf Hörerseite). Sprachliche Begriffsstrukturen werden somit auch indirekt von Kommunikationsbedürfnissen beeinflußt (z.B. dem Bedürfnis nach Kommunikationsökonomie auf der einen Seite und dem Bedürfnis nach Einhaltung einer für die Informationsübertragung notwendigen Redundanz) und natürlich auch von bestimmten Ordnungsparametern der Begriffsstrukturen selbst (vgl. Punkt 4); sie können somit mit einem synergetischen Modell beschrieben und daraus erklärt werden. Hieraus lassen sich einige Konsequenzen für die Untersuchung von kognitiven Begriffsstrukturen ableiten:

a) Begriffsstrukturen sind immer individuelle Begriffsstrukturen, d.h. Strukturen konkreter Personen, die sich aber untereinander nur hinsichtlich bestimmter Parameterwerte von allgemeinen kognitiven Begriffsstrukturen (die alle individuellen Strukturen als Spezialfälle enthalten) unterscheiden.

b) Hyponymische Begriffsketten können nicht nur und nicht ausschließlich aus den Bedeutungsdefinitionen von Wörterbüchern gewonnen werden, sondern auch unter Anwendung spezieller anderer Verfahren (z.B. Respondentenbefragungen).

c) Da wir somit keine Begriffsstruktur der Lexik bestimmter Wörterbücher suchen, sondern konkreter Einzelpersonen, und der daraus ableitbaren Verallgemeinerungen, muß auch - bezogen auf die aus einem Wörterbuch gewonnenen Begriffsketten - eine Neuordnung der Begriffe (d.h. eine andere Zuordnung der Begriffe zu Begriffsebenen) vorgenommen werden.

Es ist leicht einzusehen, daß Begriffe bezüglich der Eigenschaft Abstraktheit eine obere Grenze besitzen, die leicht aus Bedeutungsdefini-

tionen in Wörterbüchern abgeschätzt werden kann. Auf dieser – bezüglich der Eigenschaft Abstraktheit – höchsten Begriffsebene müßten dann alle Endbegriffe der untersuchten Begriffsketten angeordnet werden, auf der nächstniedrigeren Ebene deren Unterbegriffe usw.

Im folgenden Abschnitt soll gezeigt werden, daß eine solche Begriffsstruktur leicht modelliert und validiert werden kann, aus synergetischen Ansätzen resultiert und gleichzeitig einen guten Ausgangspunkt für die Untersuchung vieler anderer struktureller Eigenschaften von Begriffen darstellt.

4. Begriffsstrukturgesetze als kognitive Gesetze

4.1. Ausgangspunkt sind die von Hammerl (1988c) durchgeführten Untersuchungen, die hier noch einmal kurz beschrieben werden sollen.

Untersucht wurden 1000 polnische und deutsche Begriffsketten, deren Begriffe so in ein hierarchisches Begriffssystem eingeordnet wurden, daß die jeweils abstraktesten Begriffe auf der höchsten Ebene mit dem Index x = 1 angeordnet wurden, deren Unterbegriffe auf der darunterliegenden Ebene mit dem Index x = 2 usw., wobei Begriffswiederholungen innerhalb ein und derselben Ebene eliminiert wurden. Die Ergebnisse dieser Untersuchungen führt Tabelle 1 auf.

Tabelle 1
Zahl der Begriffe y_x auf der Ebene x und Zahl der nach Gleichung (2) berechneten Begriffe $\hat{y}_x$

x	polnische Sprache y_x	polnische Sprache $\hat{y}_x$	deutsche Sprache y_x	deutsche Sprache $\hat{y}_x$
1	261	257.99	273	277.56
2	441	455.17	406	411.88
3	442	428.85	368	353.01
4	291	286.55	218	228.79
5	141	152.20	120	124.38
6	81	68.33	67	59.82
7	28	26.93	24	26.27
8	4	9.56	14	10.75
9	1	4.42	4	4.16
10	–	–	1	2.37
	k = 14.6901 p = 0.8799 $X^2(6) = 10.04$		k = 6.446 p = 0.7698 $X^2(7) = 4.29$	

Die Zahl der unterschiedlichen Endbegriffe der Begriffsketten ist in beiden Sprachen kleiner als die entsprechende Begriffszahl der Ebenen 2–3. Ab x = 4 ist eine stetige Abnahme der Zahl der unterschiedlichen Begriffe zu verzeichnen.

Intuitiv kann man auf Ähnlichkeiten in der Begriffsstruktur beider Sprachen schließen, was natürlich bestätigt werden muß.

Ausgehend von einem Sprecher–Hörer–Modell des Informationsaustausches wird angenommen:

a) Es gibt eine für alle Begriffsebenen gleiche maximale Besetzungszahl a (a = konstant) der Ebenen, bei deren Überschreitung die Begriffsstruktur zerstört und eine neue aufgebaut wird.

b) Nicht alle Ebenen sind aber gleichstark besetzt. Der Sprecher ist bestrebt, möglichst viele abstrakte Begriffe (d.h. Begriffe der Ebenen mit kleinem x) zu verwenden, da dadurch eine relativ große Zahl verschiedener Sachverhalte umfaßt werden kann. Er ist demzufolge bestrebt, die maximale Besetzungszahl a auf den tieferen Ebenen (d.h. mit wachsendem x) zu verkleinern, d.h. mit einer Kraft bx (b = konstant, x = 1,2,...) entgegenzuwirken.

c) Der Hörer dagegen ist bestrebt, daß möglichst viele konkrete Begriffe verwendet werden, daß somit Ebenen mit relativ hohem x auch am stärksten besetzt sind, d.h. er wirkt mit einer gegen die Kraft des Sprechers gerichteten Kraft cx (c = konstant, x = 1,2,...) auf die Besetzung der Begriffsebenen ein.

d) Beide Kräfte, die des Sprechers und Hörers, führen zu relativen Veränderungen der Besetzungszahlen y_x, was folgendermaßen geschrieben werden kann:

$$\frac{\Delta y_{x-1}}{y_{x-1}} = \frac{y_x - y_{x-1}}{y_{x-1}} = \frac{P_x - P_{x-1}}{P_{x-1}} = \frac{a-bx}{cx} \qquad (1)$$

Die Lösung dieser Differenzengleichung ergibt eine negative Binomialverteilung, deren Konstanten p und k aus den Konstanten a,b und c berechnet werden können:

$$P_x = \binom{k+x-1}{x} p^k q^x \qquad (x = 0,1,\ldots), \qquad (2)$$

Da wir die Ebenen mit 1, 2,... bezeichnen, ist es praktischer, die 1-verschobene negative Binomialverteilung zu benutzen, oder die Daten bei der Anpassung um eine Ebene tiefer setzen. Das Resultat ist identisch.

Wie aus Tabelle 1 ersichtlich wird, führt die negative Binomialverteilung zu einer für die polnische Sprache zufriedenstellenden und für die deutsche Sprache zu einer guten Anpassung. Es sei erwähnt, daß die negative Binomialverteilung auch bei der Untersuchung semantischer Diversifikationsprozesse (Altmann 1985) bestätigt werden konnte.

Bemerkenswert ist aber vor allem die Tatsache, daß sowohl alle Zusammenhänge zwischen den von Köhler in dessen Selbstregulationssystem angewandten Funktionen als auch das Menzerathsche Gesetz, unser Begriffsstrukturgesetz (Gleichung (2)) und noch mehrere andere Gesetze aus dem allgemeinen synergetischen Ansatz von Köhler, Altmann (1988) hergeleitet werden können:

$$D = \frac{P_x}{P_x} = \frac{Sx + b}{Hx + d} \qquad \text{für diskrete Variablen}$$

bzw.

$$D = \frac{dy}{y} = \frac{Sx + b}{Hx + d} \qquad \text{für stetige Variablen.}$$

S stellt hier das Bedürfnis des global wirkenden Sprechers mit dem global wirkenden Grundstörfaktor a dar, H das des lokal wirkenden Hörers und eines globalen Ebenenfaktors d, wobei a,b,S,H Konstanten sind, die auch negative Werte annehmen können.

Altmann (1988) hat gezeigt, daß nicht nur viele der schon bekannten Sprachgesetze unter Berücksichtigung der entsprechenden Sprecher-Hörer-Bedürfnisse abgeleitet werden können, sondern auch bisher noch nicht untersuchte Zusammenhänge (z.B. bezüglich der assoziativen Diversifikation, bezüglich des Synonymiegesetzes und der grammatischen Diversifikation).

4.2. Eine andere Struktur sprachlicher Begriffe im mehrdimensionalen Begriffsraum wird durch die Eigenschaft der "Begriffspotenz" erzeugt, d.h.

der Häufigkeit, mit der die jeweiligen Begriffe innerhalb aller untersuchten Begriffsketten auftreten.

Wir wissen schon, daß die Eigenschaft der Häufigkeit eine der wesentlichsten Struktureigenschaften verschiedener Kommunikationssysteme ist (vgl. Orlov, Boroda, Nadarejšvili 1982), wo Häufigkeit aber immer als Texthäufigkeit angesehen wird. Daraus folgt, daß auf die Beschreibung des Zusammenhanges zwischen dem Rang und der Begriffspotenz (die Begriffe müssen natürlich vorher in eine Ranganordnung nach der Eigenschaft der Begriffspotenz gebracht werden) nicht automatisch das Zipfsche Gesetz angewandt werden kann, das ja als Textgesetz nachgewiesen wurde und nicht als Gesetz des Sprachsystems.

4.3. In Hammerl (1988c) haben wir auch gezeigt, daß die unter 4.1. beschriebene Begriffsstruktur, wo ja Begriffswiederholungen auf verschiedenen Ebenen möglich waren, nicht allgemein als Abstraktionsstruktur der Begriffe angesehen werden kann. Dieses Problem wurde in Hammerl, Schulz (1988) untersucht.

4.4. Es kann natürlich auch der Zusammenhang zwischen zwei strukturellen Eigenschaften untersucht werden, z.B. der Zusammenhang zwischen Begriffspotenz und Abstraktheit. Voruntersuchungen lassen vermuten, daß die Begriffe mit mittlerer Abstraktheit gleichzeitig diejenigen mit der größten Begriffspotenz sind. Dies soll Gegenstand eines gesonderten Aufsatzes sein.

4.5. Begriffe sind stets an bestimmte formale Repräsentationen gebunden. Als solche Repräsentationen können z.B. die Formative von Lexemen gelten.

Es entsteht somit die Frage, ob Eigenschaften dieser Formative einen Einfluß auf die Begriffsstruktur ausüben, ob und in welchem Maße diese als strukturelle Eigenschaften des n-dimensionalen Begriffsraumes gelten können.

Auch diese Frage kann noch nicht eindeutig geklärt werden. In Voruntersuchungen konnte zunächst nicht bestätigt werden, daß die mittlere Länge der Formative aller Begriffe auf einer bestimmten Ebene des Begriffssystems als strukturelle Eigenschaft gelten kann, dagegen wurden bezüglich der Verteilung der Formativlängen (in Buchstaben) innerhalb einzelner Begriffsebenen Regularitäten festgestellt. Diese Probleme werden Gegenstand weiterer Untersuchungen sein.

4.6. Wie schon im Punkt 4.1 angedeutet wurde, können auch Synonymiegesetze untersucht werden. Bezogen auf unseren Untersuchungsgegenstand heißt das, daß z.B. zwischen den nach der Hyponymiestruktur geordneten Begriffen auch Synonymiebeziehungen bestehen können, die es zu untersuchen gilt. Dasselbe trifft auf Homonymie- und Polysemiebeziehungen zu. Hier müssen zwei Hypothesen überprüft werden:

a) Da sprachliche Begriffe semantisch eindeutig sind, stellen die im Sprachsystem bestehenden Synonymie- und Polysemiebeziehungen bei Nichtberücksichtigung der Tatsache, daß die jeweiligen Begriffe durch unterschiedliche bzw. durch dieselben Formative repräsentiert werden, nur eine einzige Beziehung dar, die der "Begriffsähnlichkeit" (Grad der semantischen Übereinstimmung). Homonyme brauchen hier nicht berücksichtigt zu werden, da ja die Begriffsähnlichkeit in diesem Falle gleich Null sein müßte.

b) Auf die Begriffsähnlichkeit hat auch die Tatsache der Übereinstimmung bzw. Nichtübereinstimmung der Formative der entsprechenden Begriffe Einfluß. Wenn das der Fall ist, so muß zwischen einer Synonymie-, Homonymie- und Polysemierelation unterschieden werden.

In beiden Fällen ist natürlich zu prüfen, inwieweit hier begriffsstrukturbildende Eigenschaften vorliegen.

5. Zusammenfassung

In diesem Artikel sollte gezeigt werden, daß nicht nur Abhängigkeiten zwischen Texteigenschaften, Text- und Systemeigenschaften bzw. zwischen Systemeigenschaften verschiedener sprachlicher Ebenen aus einem synergetischen Modell abgeleitet und im Rahmen dieses Modells interpretiert werden können, sondern daß auch kognitive Begriffsstrukturgesetze aus demselben Ansatz hergeleitet werden können.

Das läßt vermuten, daß alle genannten Gesetze Resultat eines allgemeinen synergetischen Modells sind, wobei somit auch bestimmte Beziehungen zwischen den Eigenschaften der verschiedenen Untersysteme (z.B. zwischen den Eigenschaften im System Köhlers und den Begriffsstruktureigenschaften) bestehen können.

Da bisher nur wenige strukturelle Eigenschaften von Begriffen bekannt sind, können neue Begriffsstrukturen eventuell über die Beziehungen zu Texteigenschaften der entsprechenden Lexeme erkannt werden. Auch dieser Zusammenhang wurde bisher nur an kleinen Stichproben untersucht, weshalb an dieser Stelle nicht weiter auf dieses Problem eingegangen wird.

In diesem Artikel sollte aber auch ein Überblick über die derzeit wichtigsten Untersuchungsprobleme gegeben werden, die bei Untersuchungen von Begriffsstrukturen gelöst werden müssen.

Literatur

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

Altmann, G. (1985) Semantische Diversifikation. *Folia Linguistica* 19, 177-200.

Altmann, G. (1988) Diversification processes of the word. In: Köhler,R. (ed.), *Studies in Language Synergetics* (erscheint).

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

Altmann, G., von Buttlar, H., Rott, W., Strauß, U. (1983) A Law of Language Change. In: Brainerd, B. (ed.), *Historical Linguistics*, Bochum, Brockmeyer 104-115.

Altmann, G., Kind, B. (1983) Ein semantisches Gesetz. *Glottometrika* 5, 1-13.

Beöthy, E., Altmann, G. (1984) Semantic Diversification of Hungarian Verbal Prefixes III. "föl-","el-","be-". *Glottometrika* 7, 45-56.

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

Hammerl, R. (1987a) Untersuchungen zur mathematischen Beschreibung des Martingesetzes. *Glottometrika* 8, 113-129.

Hammerl, R. (1987b) Voruntersuchungen zur Überprüfung des Martingesetzes der Abstraktionsebenen an deutschem Sprachmaterial. (Unv.)

Hammerl, R. (1988a) Neue modelltheoretische Untersuchungen im Zusammenhang mit dem Martingesetz der Abstraktionsebenen. *Glottometrika* 9, 105-119.

Hammerl, R. (1988b) Überprüfung des Martingesetzes an deutschem Sprachmaterial (erscheint).

Hammerl, R. (1988c) Synergetic aspects of the formation of definition chains. In: Köhler, R. (ed.), *Studies in Language Synergetics* (erscheint).

Hammerl, R., Schulz, K.-P. (1988) Untersuchungen von Strukturen sprachlicher Begriffe - am Beispiel von Abstraktheitsstrukturen. (Referat auf der Jahrestagung der Gesellschaft für Klassifikation, Darmstadt 1988).

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

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

Köhler, R., Altmann, G. (1988) Synergetic modelling of language phenomena, In: Köhler, R. (ed.), *Studies in Language Synergetics* (erscheint).

Krylov, Ju. K. (1982) Ob odnoj paradigme lingvostatističeskich raspredelenij. In: Soontak, Ja. (Hrsg.), *Lingvostatistika i vyčislitel`naja lingvistika.* Tartu, 80-102.

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

Orlov, Ju. K. (1982) Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik), In: Orlov, M., Boroda, M. G., Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen.* Bochum, Brockmeyer 1-55.

Sambor, J. (1983) O budowie tzw. ciagów definicyjnych (na materiale definicji leksykalnych). *Biuletyn Polskiego Towarzysztwa Językoznawczego* 40, 151-165.

Schierholz, St. (1982) Untersuchungen zur Polysemie im Deutschen. Hausarbeit im Rahmen der fachwissenschaftlichen Prüfung für das Lehramt an Gymnasien. Göttingen.

Hammerl, R. (ed.), Glottometrika 10, 1988

Untersuchung struktureller Eigenschaften von Begriffsnetzen*

Rolf Hammerl, Kielce

1. Einführung

Ausgangspunkt dieser Untersuchungen ist ein Modell, welches aussagt, daß sprachliche Begriffe Elemente eines mehrdimensionalen Begriffsraumes sind. Jeder Begriff besitzt n Eigenschaften und kann durch n Eigenschaftswerte charakterisiert werden. Jeder Begriff kann somit auch als n-dimensionaler Vektor dargestellt werden, der die Lage des jeweiligen Begriffes im Begriffsraum eindeutig bestimmt. Dieses Begriffssystem ist gesetzesartig aufgebaut und entwickelt sich auch gesetzesartig, da sonst ein effektives Funktionieren eines solch komplizierten Systems unmöglich wäre.

Bei der Untersuchung dieses Begriffsraumes unterscheiden wir 2 Typen von darin wirkenden Gesetzen:

a) ***Verteilunsgesetze***, die die Verteilung der Begriffe für jeweils nur eine Eigenschaft bestimmen und somit eindimensionale Begriffsstrukturen bilden können. Bei gleichzeitiger Betrachtung mehrerer Variablen erhält man mehrdimensionale Verteilungen. Die Variablen dieser Gesetze können ordinal oder höherskaliert sein. Ordinale Variablen ergeben sich bei den sogenannten Ranghäufigkeitsverteilungen.

b) ***Relationsgesetze***, die die Beziehungen zwischen einzelnen strukturellen Eigenschaften sprachlicher Begriffe bestimmen.

Beide Gesetzestypen sind Resultat des Wirkens bestimmter externer Systembedürfnisse und interner Systemgrößen (vgl. Köhler 1986; Köhler, Altmann 1986, 1988; Altmann 1988; Hammerl 1988).

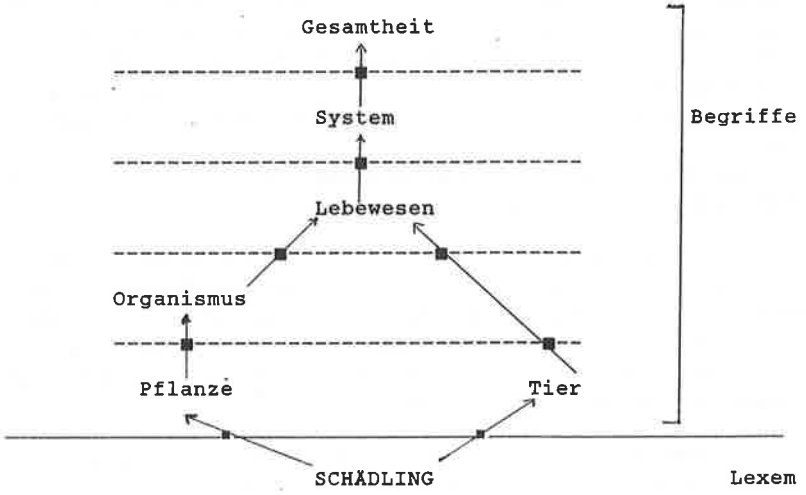
In diesem Aufsatz sollen zunächst Verteilungsgesetze sprachlicher Begriffe in speziellen Begriffsnetzen untersucht werden, die wir aus em-

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

pirischen Untersuchungen ableiten konnten. Diese Untersuchungen sollen jetzt genauer beschrieben werden:

Ausgangspunkt waren mono- bzw. polyseme Lexeme, genauer: die begriffliche Bedeutung dieser Lexeme, die aus einem einsprachigen Bedeutungswörterbuch ausgelost wurden. Zunächst wurde in jeder Bedeutungsdefinition dieser Lexeme das genus proximum als Ausgangsbegriff für die Suche deren Oberbegriffe (wiederum als genus proximum in den Bedeutungsdefinitionen der den Ausgangsbegriffen entsprechenden Lexemen) bestimmt. Somit konnte für jeden dem Ausgangslexem entsprechenden Ausgangsbegriff eine ganze Kette von Oberbegriffen gefunden werden bis hin zu den jeweils abstraktesten Begriffen.

Beispiel: Für das Ausgangslexem "Schädling", welches zwei Ausgangsbegriffe ("Tier", "Pflanze") umfaßt, wurde z.B. folgendes Begriffsnetz gefunden:



Wie man im obigen Beispiel sieht, entsprechen die Definitionsketten in einem Lexikon keineswegs den wissenschaftlichen Definitionsketten, sind aber ungefähr identisch mit der (durchschnittlichen) ethnowissenschaftlichen (d.h. natursprachlichen) Klassifikation der Welt.

Untersucht wurden also Begriffsnetze, die das jeweilige Ausgangslexem und die entsprechenden Oberbegriffe enthält, um die im Prozeß der Dekodierung sprachlicher Informationen ablaufende begriffliche Analyse von Lexemen, d.h. die Übergänge von polysemen Lexemen zu den mit

diesen Lexemen erfaßten Begriffen und deren Oberbegriffen, modellieren zu können.

Wie das obige Beispiel schon zeigt, können zwischen den einzelnen Oberbegriffen bestimmte Relationen bestehen, so daß z.B. bestimmte Unterbegriffe unter ein und demselben Oberbegriff subsumiert werden, so daß man auch auf mehreren Wegen mit unterschiedlicher Länge über eine unterschiedliche Zahl von Zwischenbegriffen zu einer auch veränderlichen Zahl von Endbegriffen gelangen kann.

Diese Eigenschaften von Begriffsnetzen der obigen Art können jedoch nicht willkürliche Werte annehmen, da sonst eine effektive Begriffsanalyse von Begriffen überhaupt nicht möglich wäre, sondern müssen - bezogen auf die Menge aller Begriffsnetze - bestimmten Gesetzmäßigkeiten folgen, die es nun zu untersuchen gilt.

2. Empirische Untersuchungen

Um erste Vorstellungen über Variablen, die die Struktur solcher Begriffsnetze beeinflussen können, zu erhalten, wurden Voruntersuchungen an einer noch relativ kleinen Stichprobe von 100 aus dem Handwörterbuch der deutschen Gegenwartssprache (1984) ausgelosten Ausgangslexemen durchgeführt. Dabei konnten zunächst folgende strukturellen Eigenschaften solcher Begriffsnetze unterschieden werden:

(a) e = Zahl der Endbegriffe in einem Netz

(b) w = Zahl der möglichen Wege vom Ausgangslexem zu den jeweiligen Endbegriffen

(c) z = Zahl aller Begriffe (einschließlich der Endbegriffe) im Netz

(d) h = mittlere Höhe der Begriffsnetze

$$h = \frac{\sum_{i=1}^{w} (1 + h_i)}{w}$$

wo h_i die Zahl der Begriffe auf dem konkreten Weg i darstellt

(e) b = mittlere Breite der Begriffsnetze

$$b = \frac{w_b}{w_z},$$

wo w_z die Zahl der Begriffe auf dem Weg mit maximaler Begriffszahl und w_b die Zahl der Schnittstellen von horizontal angebrachten Linien zwischen den Begriffen des längsten Weges mit den Pfeilen, die die einzelnen Wege des Begriffsnetzes kennzeichnen, darstellt (vgl. unser obiges Beispiel zum Ausgangslexem "Schädling").

Für unser obiges Beispiel (Ausgangslexem "Schädling") gilt:

e = 1 (Endbegriff "Gesamtheit"),

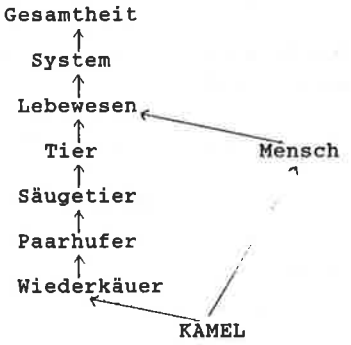
w = 2 (zwei Wege zum Endbegriff),

z = 6 (6 Begriffe insgesamt),

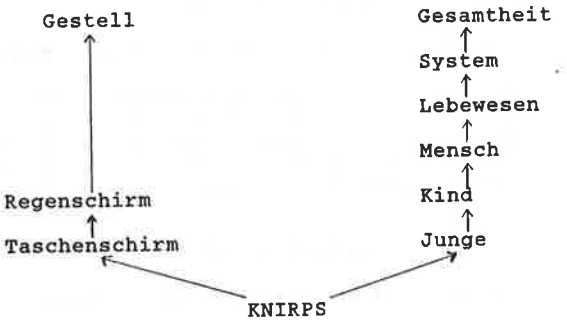
h = [(1+5)+(1+4)]/2 = (6+5)/2 = 5.5 und

b = (2+2+2+1+1)/5 = 1.6.

Es sei erwähnt, daß in unseren Untersuchungen alle im Wörterbuch aufgeführten Bedeutungen der Ausgangslexeme berücksichtigt wurden, d.h. auch umgangssprachliche Bedeutungen, da ja gerade diese Begriffe die Gestaltung der Begriffsnetze stark beeinflussen können, was unser Beispiel zeigt:



Auch lexikalische Homonyme wurden berücksichtigt, da sie ja in synchroner Hinsicht einen Grenzfall der Polysemie darstellen.



Es sei auch betont, daß in den so dargestellten Begriffsnetzen kein Vergleich der Begriffe auf verschiedenen Wegen hinsichtlich deren Abstraktheit vorgenommen werden darf, d.h., daß der Begriff "Taschenschirm" nicht genauso abstrakt/konkret wie der Begriff "Junge" sein muß, nur weil beide Begriffe im obigen Begriffsnetz auf derselben Höhe angeordnet wurden. Auf welcher Höhe der Begriff "Taschenschirm" innerhalb der möglichen Ebenen berücksichtigt wird, ist eine rein subjektive Entscheidung und hat auf die Ermittlung der Werte für die oben genannten Eigenschaften keinen Einfluß.

Wie unsere Beispiele schon zeigen, können Untersuchungen dieser Art zur genaueren Analyse der semantischen Relationen zwischen den begrifflichen Bedeutungen polysemer Lexeme beitragen, auch zu einer genaueren Abgrenzung von Polysemie und Homonymie.

In unseren Untersuchungen erhielten wir folgende Resultate (der Übersichtlichkeit halber wurden die obigen Symbole durch die Variable x ersetzt):

Tabelle 1

Verteilung der Zahl der Endbegriffe in 100 Netzen

x	1	2	3	4	5
f_x	67	22	7	2	2

Tabelle 2

Verteilung der Zahl der Wege

x	1	2	3	4	5	6
f_x	53	27	11	4	4	1

Tabelle 3

Verteilung der Zahl der Begriffe

x	1	2	3	4	5	6	7	8	9	10	11
f_x	4	13	13	16	12	14	9	10	4	2	2

Tabelle 4

Verteilung der mittleren Breite

x	1.0-1.99	2.0-2.99	3.0-3.99	4.0-4.99	5.0-5.99
f_x	60	28	7	4	1

Tabelle 5

Verteilung der mittleren Höhe

x	<2-3)	<3-4)	<4-5)	<5-6)	<6-7)	<7-8)	<8-9)	<9-10)
f_x	15	26	16	11	12	11	6	3

Bei der Modellierung der angeführten Verteilungen gehen wir von einem allgemeinen sprachlich-synergetischen Modell von Köhler, Altmann (1988) aus. Dabei wird von einer Analyse der gegenseitigen Sprecher-Hörer-Bedürfnisse in der sprachlichen Kommunikation ausgegangen und deren Auswirkungen auf die untersuchten Eigenschaften modelliert.

Der Sprecher wirkt in diesem Modell stets mit dem Bedürfnis nach maximaler Ökonomie bei der Bildung, Kodierung und Abgabe von Informationen, der Hörer mit dem Bedürfnis nach maximaler Ökonomie bei der Aufnahme, Dekodierung und Speicherung sprachlicher Informationen. Diese Bedürfnisse rufen Unifikations- und Diversifikationskräfte hervor, die man als sogenannte Zipfsche Kräfte bezeichnet. Verteilungsgesetze sprachlicher Einheiten können aber auch von systeminternen Ordnungsparametern beeinflußt werden (das kann z.B. die Zahl der Phoneme einer Sprache sein, die z.B. einen Einfluß auf die mittlere Wortlänge hat usw.).

Die hier für Eigenschaften sprachlicher Einheiten getroffenen Aussagen gelten genauso für sprachliche Begriffe als Elemente eines mehrdimensionalen Begriffssystems.

3.1. Modellierung der Verteilung der Zahl der Endbegriffe

Wie schon bei der Untersuchung der semantischen Diversifikation gezeigt wurde (vgl. Zipf 1949; Beöthy, Altmann 1984; Altmann 1985), äußert sich das ***Sprecherbedürfnis*** in einer formalen Unifikation und semantischen Diversifikation, d.h. in dem Bestreben, möglichst wenig sprachliche Einheiten mit möglichst vielen Bedeutungen zu verwenden. Das Streben nach maximaler Polysemie muß sich dann auch in einer großen Zahl von Endbegriffen der entsprechenden Begriffsnetze äußern. Je größer die semantische Diversifikationskraft, desto größer die Variable x (Zahl der Endbegriffe); diese Kraft wird im folgenden mit b gekennzeichnet. Dieser Kraft wirkt die Gegenkraft des Hörers entgegen, die wir mit c kennzeichnen.

Neben diesen Zipfschen Kräften können natürlich noch wesentlich mehr Faktoren und Eigenschaften sprachlicher Einheiten Einfluß auf die Zahl der Endbegriffe haben, z.B. die Stärke der semantischen Beziehungen zwischen den Begriffen auf verschiedenen Wegen des Begriffsnetzes. Je stärker diese Beziehungen sind, desto größer ist die Wahrscheinlichkeit, daß sie unter ein und demselben Oberbegriff subsumiert werden und somit die Zahl der Endbegriffe verkleinern.

Wir werden bei der Modellierung aller oben genannten Eigenschaften von Begriffsnetzen zunächst nur die nach unserer Meinung wesentlichsten und aus den Sprecher-Hörer-Bedürfnissen und Systemgrößen resultierenden Kräfte berücksichtigen, da erst empirische Untersuchungen an relativ großen Stichproben Hinweise darüber liefern können, ob und inwieweit weitere Faktoren zu berücksichtigen sind. Falls dann eventuelle Modifi-

kationen bei der Modellierung der jeweiligen Abhängigkeit vorgenommen werden müssen, dann sollten zunächst die von uns bei der Untersuchung der jeweiligen Eigenschaft beschriebenen Faktoren, deren Einflüsse zunächst nur in Form einer konstanten Kraft (Symbol a) berücksichtigt werden, genauer analysiert werden.

Die oben beschriebenen Kräfte a, b und c bewirken spezielle Veränderungen der relativen Besetzungswahrscheinlichkeiten einzelner Klassen von Endbegriffen, was zu folgender Differenzengleichung führt:

$$\frac{P_x - P_{x-1}}{P_{x-1}} = \frac{a + bx}{cx}, \qquad x = 1,2,\ldots K. \tag{1}$$

Offensichtlich kann die Anzahl der Endbegriffe nicht unendlich sein, sondern höchstens eine Zahl K, daher der angegebene Definitionsbereich von X.

Die Lösung der Differenzengleichung (1) führt zu der verschobenen, rechts gestutzten negativen Binomialverteilung, die man nach einigen Umformungen folgendermaßen darstellen kann:

$$P_x = \binom{k+x-2}{x-1} q^x p^k / G, \qquad x = 1,2,\ldots,K, \tag{2}$$

wo $G = \Sigma P_x$ $(x = 1,2,\ldots,K)$, $0 < p < 1$, $q = 1-p$, $k > 0$.

Wie man sich anhand von Tabelle 6 überzeugen kann, ist die Anpassung der negativen Binomialverteilung an die empirischen Daten gut.

Tabelle 6

Anpassung von (2) an die Verteilung der Zahl der Endbegriffe

x	f_x	NP_x
1	67	67.14
2	22	21.56
3	7	7.57
4	2	2.73
5	2	1.00
	k = 0.8435 p = 0.6193 FG = 1 X^2 = 0.07 P = 0.79	

3.2. Modellierung der Verteilung der Zahl der Wege

Auch hier bewirkt der Sprecher durch das Streben nach maximaler Polysemie eine Vergrößerung der Zahl der Wege mit einer Kraft b. Der Hörer bewirkt durch das Streben nach minimaler Polysemie eine Verkleinerung der Zahl der Wege mit einer Kraft c. Zusätzlich wird noch eine globale Störgröße a berücksichtigt.

In die Störgröße a können Kräfte eingehen, die aus den Relationen zwischen der Polysemie des Ausgangslexems und der mittleren Abstraktheit der diesem Lexem entsprechenden Begriffe entstehen (bei großer Abstraktheit ist die Wahrscheinlichkeit, durch Übergänge zwischen den einzelnen Wegen neue Wege zu schaffen, relativ klein, da ja die Begriffsketten relativ kurz sind) und aus der Stärke der semantischen Relationen zwischen den Begriffen auf den Wegen (wenn diese stark sind, werden Übergänge zwischen den einzelnen Wegen, d.h. neue Wege, geschaffen).

Mit demselben Ansatz wie unter § 3.1 kann wiederum die obige negative Binomialverteilung abgeleitet werden, die die Verteilung der Zahl der Wege aus unseren Voruntersuchungen (Tabelle 7) gut beschreibt.

Tabelle 7

Anpassung von (2) an die Verteilung der Zahl der Wege

x	f_x	NP_x
1	53	52.89
2	27	25.28
3	11	12.05
4	4	5.74
5	4	2.74
6	1	1.30
	k = 1.0045 p = 0.5242 FG = 3 X^2 = 1.39 P = 0.71	

3.3. Modellierung der Verteilung der Zahl der Begriffe

Auch hier gilt für das Sprecherbedürfnis: starke Polysemie des Ausgangsbegriffs und somit auch relativ große Zahl von Begriffen.

Für das Hörerbedürfnis gilt dann: kleine Polysemie und relativ kleine Zahl von Begriffen.

Die globale Störgröße a steht hier stellvertretend für Faktoren, die in Relation zur Eigenschaft der Abstraktheit des Ausgangslexems stehen (je abstrakter das Ausgangslexem, desto kleiner die erwartete Zahl von Oberbegriffen) und für eine Systemgröße, die aussagt, ob das Ausgangslexem einem semantisch stark differenzierten Begriffsfeld angehört und wie stark diese Differenzierung ist. Wir konnten in unseren empirischen Untersuchungen beobachten, daß z.B. alle Begriffe des semantischen Feldes der "Lebewesen" eine relativ große Zahl von Oberbegriffen haben (vgl. die obigen Beispiele).

Die Modellierung der hier untersuchten Abhängigkeit verläuft analog der Modellierung in den §§ 3.1 und 3.2 und führt wiederum zur negativen Binomialverteilung. Die Anpassung dieser Verteilung an unsere empirischen Daten (Tabelle 3), wie in Tabelle 8 dargestellt, ist auch hier gut.

Tabelle 8

Anpassung von (2) an die Verteilung der Zahl der Begriffe im Netz

x	f_x	NP_x
1	4	4.67
2	13	10.69
3	13	14.75
4	16	15.86
5	12	14.64
6	14	12.18
7	9	9.39
8	10	6.83
9	4	4.75
10	2	3.18
11	2	2.07
	k = 4.9072 p = 0.5330 FG = 8 X^2 = 3.59 P = 0.89	

3.4. Modellierung der Verteilung der mittleren Breite

Die Eigenschaften der mittleren Breite und Höhe sind stetige Eigenschaften, was ein etwas anderes Vorgehen bei der Modellierung der entsprechenden Verteilungen verlangt.

Analog den bisherigen Ausführungen wirkt wiederum eine auf maximale Polysemie und somit auch relativ große Breite des Begriffsnetzes gerichtete semantische Diversifikationskraft b des Sprechers und eine entsprechende Gegenkraft c des Hörers.

Als Störgröße a wirkt hier vor allem eine Kraft, die die Stärke der semantischen Beziehungen zwischen den Begriffen auf den einzelnen Wegen betrifft. Falls diese Kraft groß wird, wird die mittlere Breite klein.

Tabelle 9

Anpassung von (4) an die Verteilung der mittleren Breite

x	f_x	NP_x
<1.0-1.99>	60	59.52
<2.0-2.99>	28	26.59
<3.0-3.99>	7	9.69
<4.0-4.99>	4	3.20
<5.0-5.99>	1	1.00
A = 1.0993; B = -1.3884 FG = 1; X^2 = 0.98; P = 0.32		

Deshalb gehen wir von der Gleichung

$$\frac{df(x)}{f(x)} = \frac{bx + a}{cx}dx \qquad (3)$$

aus, deren Lösung die Funktion

$$f(x) = Cx^{a/c}e^{bx/c} = Cx^{A}e^{Bx}, \quad 1 \leq x \leq K \qquad (4)$$

ist (vgl. Köhler, Altmann 1988). Würde unsere Variable bis ins Unendliche laufen (mit $B < 0$), dann würde (4) eine Gammaverteilung darstellen. Mit $B < 0$ und C als Normierungsgröße

$$C = \left(\int x^{A} e^{Bx}\, dx\right)^{-1}$$

kann man sie wohl als eine zweiseitig gestutzte Gammaverteilung bezeichnen.

Auch die Anpassung dieser Gammaverteilung an die empirischen Daten aus Tabelle 4 ist gut (vgl. Tabelle 9).

3.5. Modellierung der Verteilung der mittleren Höhe

Eine große semantische Diversifikationskraft hat hier keinen direkten Einfluß auf die mittlere Höhe der Begriffsnetze und wird somit in der Störgröße a (auch die entsprechende Gegenkraft des Hörers) berücksichtigt.

Einen großen Einfluß, und das konnte in unseren empirischen Untersuchungen ganz deutlich beobachtet werden, hat vor allem eine Kraft, die aussagt, ob das Ausgangslexem einem semantisch stark differenzierten Begriffsfeld angehört und wie stark diese Differenzierung ist. Diese Kraft soll mit b bezeichnet werden. Außerdem wirkt eine aus der Abstraktheit der Ausgangslexeme resultierende Kraft k, die um so größer ist, je kleiner die Abstraktheit der Ausgangslexeme.

Der Sprecher ist bestrebt, die Höhe der Begriffsnetze möglichst klein zu halten, d.h. die Kraft b wirkt der konstanten Störkraft a entgegen, die Kraft k wirkt dagegen unterstützend auf die Kraft a ein.

Der Hörer dagegen wirkt der Verkleinerung der Begriffsnetze entgegen, indem er mit einer Kraft c auf den Ausbau relativ differenzierter Begriffsfelder wirkt und mit einer Kraft l auf die Vergrößerung der Zahl der Begriffe mit relativ geringer Abstraktheit.

Hieraus folgt folgender Modellierungsansatz:

$$\frac{df(x)}{f(x)} = \frac{a + kx - bx}{cx + lx}dx$$

$$= \frac{a + (k - b)x}{(c + l)x}dx$$

$$= \frac{a + mx}{rx}dx. \qquad (5)$$

Auch die Lösung dieser Gleichung führt auf die obige Gleichung (4), welche auch hier die empirischen Daten in Tabelle 5 gut beschreibt (vgl. Tabelle 10).

Tabelle 10

Anpassung von (5) an die Verteilung der mittleren Höhe

x	f_x	NP_x
<2-3)	15	17.01
<3-4)	26	20.27
<4-5)	16	18.69
<5-6)	11	15.16
<6-7)	12	11.39
<7-8)	11	8.13
<8-9)	6	5.60
<9-	3	3.80

A = 1.5337; B = -0.5922
FG = 5; X^2 = 4.61; P = 0.46

4. Zusammenfassung

Es wurde gezeigt, daß aus einem allgemeinen synergetischen Ansatz spezielle Modelle resultieren, die die Verteilung sprachlicher Begriffsnetzeigenschaften beschreiben. Diese Modelle betrafen in unseren Untersuchungen sowohl diskrete als auch stetige Eigenschaften und führten zu Modifizierungen der negativen Binomialverteilung bzw. der Gammaverteilung.

In allen Modellen wurden nur die der bisher erkannten Kräfte berücksichtigt, deren Einfluß uns am stärksten erschien. Es wurde aber für alle untersuchten Eigenschaften angegeben, welche anderen Kräfte bei eventuellen Modifikationen der vorgestellten Modelle besonders zu berücksichtigen sind.

Um die Frage nach einer eventuell weiteren notwendigen Modifikation der oben vorgestellten Modelle beantworten zu können, müssen zunächst zeitaufwendige, empirische Untersuchungen an mehreren Sprachen durchgeführt werden.

Literatur

Altmann, G. (1988), Diversifikationsprozesse des Wortes. In: Köhler, R. (Ed.), *Studies in language synergetics* (erscheint).

Altmann, G. (1985), Semantische Diversifikation. *Folia Linguistica* 19, 177 -200.

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

Handwörterbuch der deutschen Gegenwartssprache. In zwei Bänden. (1984) Berlin, Akademie-Verlag.

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

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

Köhler, R., Altmann, G. (1988), Synergetic modelling of language phenomena. In: Köhler, R. (Ed.), *Studies in language synergetics* (erscheint).

Hammerl, R. (1988), Synergetic aspects of the formation of definition chains. In: Köhler, R. (Ed.), *Studies in Language Synergetics* (erscheint).

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

Hammerl, R. (ed.), Glottometrika 10, 1988

Quantitative Lexicology of Finnish

Raimo Jussila, Pauli Saukkonen, Tuomo Tuomi, Helsinki

0. The Finnish Research Centre for the Domestic Languages and the Department of Finnish and Lappish at the University of Oulu concluded a contract in Spring 1987 with the Institute of Linguistics at the University of Bochum (Ruhr-Universität) concerning participation in an international project "Language Synergetics". In Finland two projects were started, one dealing with Finnish, the other with Lappish: "Quantitative Lexicology of Finnish" and "Quantitative Lexicology of Lappish". This article deals with the content and goals of the project devoted to Finnish. It is being implemented by a working group which includes Professors Tuomo Tuomi and Pauli Saukkonen, and Raimo Jussila, M.A.

The Finnish lexicon is an exceptionally open system. The vocabulary may be increased either by compounding or derivation virtually without limitation. Very few constraints limit the formation of ***compounds***, these constraints being of semantic or syntagmatic nature. The most important semantic constraint is that the result cannot be a nonsense word. The syntagmatic restrictions are stronger. Verbs cannot be generally joined to form a composite with representatives of other morphological classes. Only by means of so-called nominalization is it possible for a verb to form a compound with a noun. The few compound verbs which do exist are usually derivatives from compounds (*valokuva* [*valo* 'light' + *kuva* 'picture'] 'photograph' --> *valokuvata* 'to photograph') or technical terms calqued from some other language (*kylmävalssata* 'to cold-roll', *kaasuhitsata* 'to torch weld'). Neither do particles normally form compounds. On the other hand, nominals (substantives and adjectives) may, within the boundaries of the semantic restrictions mentioned above, freely form compounds. Since the members of a compound may themselves be compounds, the formation of purely nominal-based compounds already opens vast possibilities for augmenting the vocabulary.

A somewhat larger number of restrictions exists concerning ***derivation***. These are both phonotactic and morphological in nature. Here we can only cite a few examples. With very few exceptions it is not possible

to derive verbs from derived nominal bases, certain derivational elements can only be added to roots with a specific phonological structure, etc. Since Finnish disposes of over approximately 200 derivational morphemes (140 for deriving nouns and 60 for deriving verbs) , and since it is usually possible to derive further derivatives from a derivative, the process of derivation offers a wide range of possibilities for augmenting the lexicon.

The basis for our research is provided by the most complete dictionaries of Finnish. Not even they cover all of the potential vocabulary of Finnish. Everything presented in those sources is actual vocabulary in the sense that the words have been selected from actual texts. This set of words represents the Finnish lexicon as well as a collection of samples that can be representative of an open system. The total size of the sample is approximately 215,000 entries. It is in a form suitable for data processing purposes.

According to the general coding system of the project every word will be provided with the information in 14 columns (files) described below (1. Lemma, 2. Syllabification etc.).

1. The primary research unit is the graphemic form of the lexeme (***lemma***). The phonematic nature of Finnish orthography makes it possible to study the phonological structure of the lexeme at the same time. Orthography only differs from phonematic structure in two instances. They are the following:

(1) An *n* appearing before an orthographical *k* is actually realized phonologically as *ŋ*. The Finnish *ŋ* first became an independent phoneme when the cluster *ŋk*, which previously did not participate in consonant gradation, accommodated itself to this alternation in the same manner as other clusters of nasal plus occlusive (*nt : *nd > nn*; *mp : *mb > mm*), resulting in the gradation *ŋk : ŋŋ*. The cluster *ŋŋ* is represented orthographically as *ng*.

(2) Aside from a few exceptions involving neologisms, proper names, and affective vocabulary (e.g. *ale* 'sale', *Kalle* 'Charlie', *nalle* 'teddy bear', *nukke* 'doll') the word final *−e* contains an element which either lengthens the vowel of the stem preceding the suffix, or doubles the initial sound of a morpheme beginning in a consonant (*vene: venee−n: venet−tä* 'boat nomSg, genSg, partSg'; *purje + tuuli: purjet−tuuli* 'sail'+ 'wind': 'fair wind, breeze').

On the basis of the material we have used, the phonotactic structure of the Finnish vocabulary is already rather well known (see Karlsson 1983). Consequently, our research on the morphological word (formative, lexical items in orthographic or phonemic form) concentrates on certain suprasegmental features such as vowel harmony, syllable structure, and the positioning of different syllable types within the word in addition to normal phonotactics. Rendering the picture of Finnish phonotactics more precise is, of course, also one of the goals of our research.

An inverse dictionary of the type so important for the study of derivation has been compiled for Finnish and it is specifically based on the material we have used, see Tuomi (1972). The inverse dictionary only contains the bases and derivatives; the compound words are stored as a separate computerized data file.

2. An extremely powerful program exists for Finnish ***syllabification***. With two exceptions, it syllabifies a word in its basic form with absolute reliability. Problems are caused by: (1) the interpretation of certain V + i sequences in non-initial syllables, and (2) compounds. Sequences of V + i in non-initial syllables may also be interpreted as diphtongs. On the other hand the second component of the diphtong may, from a historical and morphological standpoint, be a part of the derivational affix or it may be a marker (*a−vain, a−va+in* 'key'; *lauloim−me − laulo+i+mme* 'we sang'). In Finnish dialects it is quite common for sequences ending in *−i* of this type to be diphtongs on the basis of gemination criteria (*avvain*; a consonant is geminated only before long vowels and diphtongs). In our research we interpret sequences of this type to be diphtongs.

The syllabification of compound words is made difficult by compound words which exhibit ambivalence with respect to the location of the juncture e.g. in

kaivosaukko (1) 'mine opening' (*kaivos + aukko*)
(2) 'well otter' (*kaivo + saukko*)

or in

koululaiskuri (1) 'a person who is lazy at school' (*koulu + laiskuri*)
(2) 'discipline of school children' (*koululais + kuri*).

Particularly problematic from the standpoint of automatic syllabification are those cases in which the first component of the compound ends in a

consonant, with the second component beginning with a vowel. Difficulties of this type are avoided because the morpheme boundary or boundaries between the components of compounds are marked in our material (the marking also distinguishes primary and secondary divisions). For this reason the syllable boundaries can be indicated automatically.

The study concentrates on an analysis of the structure of the syllables and of the syllable structure of the words.

3. ***Stress*** in Finnish is of no interest, since the main stress always falls on the vowel of the first syllable of a word, with the secondary stresses falling on the vowels of successive odd numbered syllables. In certain predictable cases the secondary stress may also fall on the fourth and subsequent even numbered syllables.

4. ***Word class*** is marked in all the basic material. Nine word classes are distinguished:

A = adjective
D = adverb
I = interjection
K = conjunction
N = numeral
P = post/preposition
R = pronoun
S = noun
V = verb.

Combinations of different word classes can also be used (e.g. NR = numerical pronoun, AS = adjective and noun). The classification is based upon the concept of word class in traditional grammar.

5. The information concerning ***inflection*** is extracted automatically from the file for the *Reverse Dictionary of Standard Finnish* which, in turn, is based upon the *Nykysuomen sanakirja*, the closest equivalent to an unabridged, academy-type dictionary available for contemporary Finnish. Declensional type is indicated by a number referring to a declensional scheme. Nominals (nouns, adjectives, numerals, and pronouns) have their own declensional scheme, and verbs each have their own specific conjugational scheme. Additionally, the following information obtained from the *Reverse Dictionary of Standard Finnish* is indicated in this field by a code:

& = alternative inflection
E = irregular declension
M = plurale tantum
P = the word occurs in the plural
T = conjugation as both an impersonal and normal verb
V = defective declension
Y = conjugation as an impersonal verb
X = the possessive suffix is obligatory
Z = the possessive suffix may occur optionally
A = gradation.

6. ***Morphological status*** covers the following eight classes:

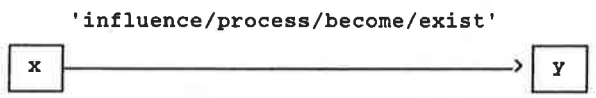
S = stem
D = derivate
C = compound
DC = derived compound (= compound + derivative suffix)
CD = compounded derivates (= derivative + derivative)
L = compound written orthographically separately
N = status unknown
U = uncertain status.

The compound words and the majority of the derivatives may be defined automatically from the computerized material for the *Reverse Dictionary of Standard Finnish*. The boundaries for derivatives, bases, and different types of compound words have to be checked manually.

7. The ***semantic and syntactic properties*** of substantives and verbs are coded in the ***function*** column. The purpose is to determine such things as the degree of regularity and interplay governing the mutual combination of these basic word classes at the sentence level and the manner in which the properties to be studied correlate with the other properties of the vocabulary.

7.1. Verbs are described along with their arguments as a series of ***valence*** roles. The description is based on the sample sentences and definitions in the *Nykysuomen sanakirja*. Different types of usage are represented as different combinations of arguments. Obligatory and facultative arguments are not strictly differentiated; in other words, arguments are included which are not absolutely obligatory.

The theoretical basis for the description is the fact that a verb expresses relationships (R) between the nominal members of the sentence (arguments x, y,...): xRy. The following are considered to be the primary meanings of the relation: 'influence', 'process', 'vary, become', 'be, exist'; these can account for all the meanings of all verbs (cf. Jämsä 1986):



Meanings thus also express the direction of a relationship, so that *x* is the beginning of 'influencing/processing/becoming/existing' and *y* is its end. The initial argument *x* thus has the role 'that which influences' or 'that which processes' or 'that which is changing, becoming something' or 'that which exists, is something', or some combination of these. The basic system of initial and terminal roles is in general the following:

Initial roles

Code	Role	Example
i	influencer	*Väsymys* aiheutti onnettomuuden '*Fatigue* caused the accident'
p	processor	*Lapsi* putosi tuolista '*The child* fell from the chair'
v	variable, one which is changing, becoming, transforming something	*Elämä* tuli helpommaksi '*Life* became easier'
e	existent	*Sunnuntai* oli eilen '*Sunday* was yesterday'

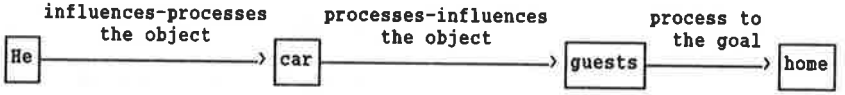
Terminal roles

Code	Role	Example
r	result	Väsymys aiheutti *onnettomuuden* 'Fatigue caused the *accident* ' Elämä tuli *helpommaksi* ' Life became *easier* '
n	patient	Hän hoitaa *lasta* 'She is taking care of *the child*'
j	object	Hän lukee *kirjaa* 'She is reading *a book*'
g	goal	Juna saapui *asemalle* 'The train arrived *at the station*'
o	origin, source	Lapsi putosi *tuolista* 'The child fell *from the chair* '
l	location	Subjekti on *ennen objektia* 'The subject is *before the object* Sunnuntai oli *eilen* ' Sunday was *yesterday* '
q	quality, sort	Hän on *arkkitehti* 'He is *an architect* '

When several valence roles can be associated with a verb a consistent chain of relations is formed. For example, the sentence

Hän ajaa vieraat autolla kotiin
'He drives the-guests in-the-car home'

may be represented by the following process chain in which the arrows indicate the meaning of the verb *ajaa* 'drive':



The roles of the arguments between the initial and terminal arguments of the chain of relations are usually double or triple roles formed from the basic ones. An initial role can also be double. Each role type has its own position in the coding. The total system is as follows:

LEMMA	INITIAL	AUXILIARY	INITIAL/MAIN	TERMINAL
ajaa	ip	jpi	jp	g
ajaa	hän	autolla	vieraat	kotiin
'drive'	'(s)he'	'by car'	'the guests'	'(to) home'
teettää	i	jip	jv	r
teettää	hän	apulaisella	työn	valmiiksi
'have made'	'(s)he'	'with assistant'	'the work'	'ready'

The following are theoretical double or triple roles: ip, iv, ie, ji, jip/jpi, jp, jv, je. The combinations of the roles given in the previous example are:

ip = influencer-processor (agent, agentive)
jpi = object-processor-influencer (instrument)
jip = object-influencer-processor
jp = object-processor
jv = object-variable.

(The overall model has also been influenced by Autio 1986.)

The trend is that the more general and shorter the verb lexeme is, the more roles are connected with it. For example, *ajaa* 'to drive, force etc.', a very common verb, has the following argument structures:

ajaa	ip	jpi	jp	g/o
ajaa	ip		jp	q
ajaa	ip		ji	r
ajaa	ip		jv	r
ajaa			jv	r
ajaa			ip	g
ajaa			ip	q
ajaa			ip	l
ajaa			i	r
ajaa			v	l
ajaa			p	g
ajaa			p	l
ajaa			p	q

More specific verbs may have implicit "arguments" incorporated within them, thus limiting the possibility of explicit arguments. For example, the verb *aavikoitua* 'to become a desert' can only have one argument role, the role v (= variable = that which becomes a desert) which belongs to the INITIAL/MAIN position :

aavikoitua v

7.2. Substantives are coded according to abstraction level (cf. Lyons 1977: 495) and intentionality/animation level. Compound words are not included at this stage. The scale is as follows:

Abstraction level	Intentionality/animation level	
c = concrete		
	h = human	e.g. *arkkitehti* 'architect'
	f = fauna & flora	e.g. *ki*ssa 'cat'
	m = material	e.g. *kirja* 'book'
t = time		e.g. *sunnuntai* 'Sunday'
d = dynamic energy		e.g. *juoksu* 'run'
s = static energy		e.g. *lämpö* 'heat'
a = abstract		e.g. *objekti* 'object'.

When the substantive belongs to several main categories the codes are combined with a plus sign. The subcategories h, f, and m are added directly after the main categories, e.g.

liike	'motion, tendency, business enterprise'	d+a+cmh
lukeminen	'reading, lesson (abst. and concr.)'	d+a+cm
käsi	'hand, authority, possession'	cm+a+s.

7.3. The categorization of substantives is also applied to the arguments of verbs so that in addition to the roles, the arguments also have a code or a code series, immediately following, which accords with their inherent meaning (without the plus sign) in the following manner:

LEMMA	INITIAL AUXILIARY	INITIAL/MAIN	TERMINAL
aiheuttaa		is	rd
aiheuttaa		väsymys	onnettomuuden
'cause'		'fatigue'	'accident'
pudota		ph	om
pudota		lapsi	tuolista
'fall'		'child'	'from the chair'
tulla		vd	ra
tulla		elämä	helpommaksi
'become'		'life'	'easier'
olla		et	lt
olla		sunnuntai	eilen
'be'		'sunday'	'yesterday'
olla		ea	la
olla		subjekti	ennen objektia
'be'		'subject'	'before the object'
olla		eh	qh
olla		hän	arkkitehti
'be'		'(s)he'	'architect'

lukea	iph	jm
lukea	hän	kirjaa
'read'	'(s)he'	'book'
saapua	pm	gm
saapua	juna	asemalle
'arrive'	'train'	'at the station'

A direct result of the analysis is a list and typology of the basic sentence types in Finnish.

8. The eighth column indicates the ***literary source*** of the lemma. This is needed during the stage when the material is being collected and the lemmas of a specific source are being extracted for manual coding. Additionally, it allows research to be concentrated on, for example, data from a specific source. Information about sources is also useful when doing such things as supplementing the basic information. The sources are as follows:

T = Tuomi, *Suomen kielen käänteissanakirja*
Y = Tuomi, *Suomen kielen käänteissanakirjan yhdyssanatiedosto*
T+Y = *Nykysuomen sanakirja*
S = *Suomen kielen taajuussanasto* ('A Frequency Dictionary of Finnish')
U = *Uudissanasto 80* ('Dictionary of Neologisms 80')
P = *Nykysuomen perussanakirja* ('Basic Dictionary of Finnish') (to appear in 1989).

9. The information concerning the ***number of meanings*** is based upon the lexicographical classification in the *Nykysuomen sanakirja* (scope approx. 207,000 words) and the *Uudissanasto 80* (scope approx. 6,000 words). The dictionaries have been gone through article by article, and their classification has been stored in coded form. All words have been included which have two or more classificational groups. Each classificational group has been regarded as being a meaning group regardless of the semantic weight and level in the classificational hierarchy. Only evident typographical errors have been corrected.

The number of meanings was calculated using numerical codes. Codification of the information concerning classification is an intermediate step, the purpose of which was to expedite the extraction of data. All of the markings in the internal arrangement of the article which make use

of boldface letters or numbers have been taken into consideration in the codification. The code consists of eight (in a few rare cases ten) numbers which, in turn, form four (five) sets of numbers as follows:

1. pair of numbers: indicates the number of groups in the level of classification marked with capital letters (A, B, C,...).
2. pair of numbers: indicates the number of groups in the level of classification marked with Roman numerals (I, II, III,...).
3. pair of numbers: indicates the number of groups in the level of classification marked with Arabic numerals (1, 2, 3,...).
4. pair of numbers: indicates the number of groups in the level of classification marked with small letters (a, b, c,...).
5. pair of numbers: indicates the number of groups in the level of classification marked with two small letters (aa, bb, cc,...) (very rare).

For example, the code might be 03082214, this indicating that the highest level of the hierarchy has three classificational groups, the second eight groups, the third twenty-two, and the fourth fourteen giving a total of 47 classificational groups. The code 00000200 indicates that the word has two groups of meanings which are indicated by Arabic numerals.

The code numbers were used to calculate the following information for each word:

(1) *The number of groups of meaning* (= the nodes of the lowest level in the tree diagram for the classificational hierarchy; these contain the data in the dictionary providing examples),

(2) *the number of classificational groups* (= all nodes of the tree diagram; in addition to the foregoing we also included here dominant nodes; these usually contain only classificational information, but no example material), and

(3) *the number of hierarchical levels.*

For some of the groups of meanings it was necessary to calculate the number manually. These identification numbers are indicated in the ninth column and separated by a slash (/), e.g. *ajaa* 16/19/2).

The goal of the study is to determine the structure of the vocabulary of the contemporary Finnish literary language from the standpoint offered by lexicographical classification. Purely statistic aspects have provided the first object of research: How many words in Finnish have one, two, etc. meanings? How many words are there in the classificational

hierarchy with one, two, etc. levels? It is the intention in the next phase to supplement this with the other variables in the database: Does the number of meanings have any connection with the length of the word, the word class, the fact that the word is derived or compounded, or with a feature such as the word's age? A central object of research will be the manner in which the number of word meanings is a function of frequency.

An additional goal will be to determine what the lexicographical classification used in the dictionaries which served as the sources of information exactly describes: the semantics of the word, grammatical usage, usages connected with extralinguistic associations, or something else. A second goal will be to determine whether the lexicographical classification systems used in the dictionaries which served as sources are uniform and consistent. A third goal will be an attempt to determine the conditions under which studies of the same type made for different languages can be compared.

10. The ***literal age*** of a word is indicated by a date or period between two dates in the tenth field. Dating the approximately 200,000 words which were used as data is an overwhelming task. For this reason dating is concentrated specifically on those words which have been established in the literary language for the longest time and which have shown themselves to be used with the greatest frequency. The age of some of the source data can be determined within a specific period (e.g. the vocabulary in *Uudissanasto 80* is from the years 1960–1979).

The Finnish literary language is of quite recent origin. The oldest literary sources are from the mid 16th century. The archives for the dictionary of old literary Finnish have numerous examples of usage for the period preceding the beginning of the 19th century. The entire production of Mikael Agricola, the founder of the Finnish literary language, is available in computerized form, making the vocabulary he used completely accessible (see *Index Agricolaensis*). In contrast, vocabulary collections for the period during which the present literary language was emerging (the 19th century) are quite incomplete. A brief collection of entries relevant to this period has been published (Rapola 1960).

The Research Centre for the Domestic Languages has begun to collect the first appearances of Finnish words. The *Vanhan kirjasuomen sanakirja* (Dictionary of Old Literary Finnish), the first part of which appeared in 1985, as well as archives have provided information on the first recorded appearance of words. This includes the date, source, and sometimes also the meaning, and it is being compiled in computerized

form. After this information on literary age has been collected, the computer will be used to combine it with other databases. There are plans to publish a list concerning the first appearance of words. A thesis is also being written on this data.

11. Information concerning ***etymology*** (column 11) has not yet been compiled. The sources will be the *Suomen kielen etymologinen sanakirja I–VII* (Etymological Dictionary of Finnish, vol. I–VII) and the work *Suomen sanojen alkuperä* (The Origin of Finnish Words) which is presently being compiled. It is planned to determine the approximate distribution of all root words and at least the most important derivatives in the languages related to Finnish, as well as the source language of loan-words. The degree to which compound words are to be etymologized has not yet been determined.

The etymologies will be presented in the following manner:

F = only in the Finnish language
I = equivalents in the Finnic languages
L = equivalents in the Lappish language
V = equivalents in the Volgaic languages
P = equivalents in the Permian languages
U = equivalents in the Ugrian languages
S = equivalents in the Samoyedic languages
B = Baltic loan
G = Germanic loan
R = loan from current Swedish
E = loan from current English
V = Russian loan
D = Indo-European loan
A = international word, originally from Latin
K = international word, originally from Greek
C = loan from Lappish
M = loan from other languages
N = loan of uncertain origin
H = distribution in related languages obscure
J = obscure origin.

12. The information concerning ***frequency*** of the Finnish vocabulary is based on Saukkonen, Haipus, Niemikorpi, Sulkala (1979), which was compiled at the department of Finnish and Lappish at Oulu University. It contains the following material:

(1) Original Finnish literary works which appeared during the period 1961–67 (700 textual samples)

(2) talk shows originally intended for the radio Sept. 29, 1968 – May 26, 1969 (1,100 samples)

(3) newspapers and magazines which appeared in 1967 (1,600 samples)

(4) Finnish non-fiction which appeared during the period 1961–67 (2,300 samples).

Each sample consists of five sentences and at least sixty words. The samples comprise a total of 5,700 fragments of randomly selected text. The organizational and orthographical principles of the *Nykysuomen sanakirja* were followed in the lemmatization of the material.

The basic material comprises 408,301 running words and 43,670 different words. The material is in computerized form, and in addition to the entries and their frequencies it contains other information (such as the rank numbers of words, the word classes, the percentage sum). Field 12 will contain an indication of frequency extracted by computer. A frequency of zero will be indicated for those words of the basic data which do not appear in the material of the frequency dictionary.

13. ***Polytexty*** will be indicated in column 13. The information will be extracted automatically from the *Frequency Dictionary of Finnish*. Four contextual types will be distinguished:

K = fiction
R = radio
L = newspaper
T = non-fiction
U = no occurrence in a frequency list.

14. The last column will contain the ***specification of subject or style***. The characterization will be taken from the *Nykysuomen sanakirja*. Examples of these are vulgar, professional, colloquial, used in the slang of school children, disparaging sense, rejectable form, evasively, provincial, figurative, used in child language, humorous, dialectal, preferable form, conversational, poetry term, in poems, archaic; biological, zoological, religious, term used in fishery, used in the metal industry, military, railroading term. There are approximately ninety terms altogether.

Bibliography

Autio, Risto (1986), Suomen kielen elementaarilauseiden rakennetyypit (Structural Types of Finnish Elementary Sentences). Unpublished licentiate thesis, Department of Finnish and Lappish, University of Oulu 1986.

Häkkinen, Kaisa (1981), Johtaminen ja leksikaalistuminen - lingvistin vai tavallisen kielenkäyttäjän ongelmia? (Derivation and Lexicalization - Problems for the Linguist or the Normal Language User?). *Publications of the Finnish Linguistic Society 9,*

Index Agricolaensis I-II. *Publications of the Research Centre for Domestic Languages* 11, 1980 (Helsinki).

Jämsä, Tuomo (1986), Suomen kielen yleisimpien verbien semantiikkaa (On the Semantics of the Most Common Finnish Verbs). *Acta Universitatis Ouluensis, B 12, Philologica 5* (Oulu).

Karlsson, Fred (1983), Suomen kielen äänne- ja muotorakenne (The Phonology and Morphology of Finnish). WSOY, Helsinki-Juva.

Lyons, John (1977), Semantics 2. London, Cambridge University Press.

Nykysuomen sanakirja 1-6 (Dictionary of Contemporary Finnish, vol. 1-6). WSOY, Porvoo 1951-1961.

Rapola, Martti (1960), Sanojemme ensiesiintymiä Agricolasta Yrjö Koskiseen (The First Appearance of our Words from Agricola to Yrjö Koskinen). *Tietolipas 22.* Suomalaisen Kirjallisuuden Seura, Helsinki.

Saukkonen, Pauli; Haipus, Marjatta; Niemikorpi, Antero; Sulkala, Helena (1979), Suomen kielen taajuussanasto (A Frequency Dictionary of Finnish). Porvoo 1979.

Suomen kielen etymologinen sanakirja I-VII (Etymological Dictionary of Finnish, vol. I-VII). *Lexica Societatis Fenno-ugricae XII,1.* Helsinki 1955-1981.

Tuomi, Tuomo (1972), Suomen kielen käänteissanakirja - Reverse Dictionary of Modern Standard Finnish. Suomalaisen Kirjallisuuden Seura, Hämeenlinna 1972, 2. revised edition 1980.

Uudissanasto 80 (Dictionary of Neologisms 80), Edited by the Language Bureau of the Research Centre for Domestic Languages. WSOY, Porvoo 1979.

Vanhan kirjasuomen sanakirja 1 (Dictionary of Old Literary Finnish). Publication 33 of the Research Centre for Domestic Languages. Helsinki 1985.

R. Hammerl (ed.), Glottometrika 10, 1988.

Polnische Version des Projekts "Sprachliche Synergetik. Teil I. Quantitative Lexikologie"*

Jadwiga Sambor, Warschau

1. Einführung

Schon seit mehreren Jahren ist zu beobachten, daß man sich innerhalb der quantitativen Linguistik für statistische Gesetze des Vokabulars interessiert, die von den bekannten Zipfschen Gesetzen (Zipf 1935, 1949) einen wichtigen Impuls erhielten. Das führte vor allem zu einer präziseren Formulierung des Begriffs des statistischen Sprachgesetzes (Altmann 1978).

Den Gegenstand genauerer Analysen bildeten später einzelne quantitativ-lexikalische Sprachgesetze, die entweder den Text oder das Lexikon betrafen (oder Text und Lexikon gleichzeitig). Unter den Gesetzen, die den synchronischen Aspekt der Lexik betrafen, wurde vor allen Dingen das bekannteste Gesetz von Zipf-Mandelbrot umformuliert und neu interpretiert (Orlov 1982a,b,c; Guiter, Arapov 1982), ein neues Modell für die Verteilung der Wortlängen nach der Silbenzahl im Text gefunden (Fucks 1955, Grotjahn 1982) und von neuem die Verteilung der Lexeme im Lexikon nach der Bedeutungszahl untersucht, die gegenwärtig als Krylovgesetz bezeichnet wird (Zipf 1949; Papp 1967; Krylov 1982, 1984). Die größte Zahl der Arbeiten wurde jedoch dem Menzerathschen Gesetz gewidmet, welches die Abhängigkeit zwischen der Länge der sprachlichen Konstrukte (x) und der (mittleren) Länge deren Bestandteile (y) erfaßt. Auf der Suche nach einem adäquaten Modell für die Beschreibung dieses Gesetzes formulierte G. Altmann eine einfache Annahme, die zu einer Funktion führt, welche den Typ der Relation zwischen den untersuchten Variablen x und y bestimmt (Altmann 1980). Die genannte Arbeit von Altmann eröffnete eine neue Etappe in der Erforschung der einzelnen

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

Sprachgesetze - die Modellierung dieser Gesetze stellte nunmehr ein echt deduktives Vorgehen dar.

Das Menzerathsche Gesetz wurde bisher im Bereich der Phonologie (vgl. Menzerath 1954; Altmann 1980; Geršić, Altmann 1980), der Morphologie (Gerlach 1982) und der Syntax (Köhler 1982; Heups 1983) überprüft. Es wurde auch die Abhängigkeit zwischen der Lexemlänge und der mittleren Bedeutungszahl der Lexeme im Lexikon untersucht (Altmann, Beöthy, Best 1982; Fickermann, Markner-Jäger, Rothe 1984; Sambor 1984), jedoch ist die Interpretation dieser Abhängigkeit als eine Manifestation des Menzerathschen Gesetzes in der Semantik zumindest diskutabel (beide Variablen: das Konstrukt (Lexem) und deren mittlere Bedeutungszahl sind nicht einheitlich, da sie unterschiedlichen Ebenen angehören, der formalen und der semantischen, worauf Köhler hingewiesen hat; vgl. Köhler 1986, 10-11).

2. Synergetischer Aspekt der Untersuchung der Sprache. Das Modell von Köhler als synergetische Beschreibung des lexikalischen Systems

Als Ergebnis des wachsenden Interesses an den oben genannten - und vielen anderen - lexikalischen Gesetzen wurden zumindest einige von ihnen an relativ reichem sprachlichen Material untersucht (z.B. das Menzerathsche Gesetz oder Shermans Gesetze, vgl. Altmann 1988), jedoch gab es bisher zwischen diesen Gesetzen kaum Querverbindungen, sie wurden meistens isoliert betrachtet. Die in den letzten Jahren formulierten Thesen der Synergetik als Wissenschaft von der Kooperation einzelner Teile von Systemen (gr. synergòs 'mit-wirkend') (Haken 1978) eröffneten neue Perspektiven weiterer Untersuchungen der quantitativen Sprachgesetze.

Der synergetische Aspekt linguistischer Forschung ist nichts anderes als die Suche nach den gegenseitigen Abhängigkeiten der einzelnen Gesetze, d.h. nach einem Netz, einem System von Gesetzen, oder mit anderen Worten, nach einer Sprachtheorie. In der Linguistik hat bereits Zipf die Annahme geäußert, daß die Sprache ein selbstregulierendes System darstellt, dessen quantitativ-qualitative Strukturen Ergebnis des Wirkens der sogenannten Zipfschen Kräfte sind, d.h. der miteinander oder gegeneinander wirkenden Bemühungen der Sender und Empfänger zur Minimierung ihres Aufwandes, den die Aussendung und der Empfang eines sprachlichen Textes erfordert.

G. Altmann und R. Köhler (Köhler, Altmann 1986; 1988) haben diese Konzeption aufgenommen und begannen mit der Erforschung der Lexik unter synergetischem Aspekt, obwohl unter den Linguisten die Überzeugung verbreitet ist, daß das Vokabular - im Gegensatz zur Phonologie und Grammatik - keine Struktur besitzt.

Köhler (1986) hat ein Projekt eines synergetischen Lexikonmodells formuliert, das die gegenseitigen Abhängigkeiten zwischen verschiedenen quantitativen Eigenschaften von Lexemen, die als Variable aufgefaßt wurden, beschreiben und erklären sollte.

In allen oben genannten lexikalisch-statistischen Gesetzen waren die grundlegenden quantitativen Eigenschaften der Lexeme, d.h. die grundlegenden Variablen: die Länge (L) (gemessen in Silben, Phonemen und Buchstaben), Frequenz (F) und die Bedeutungszahl im entsprechenden Wörterbuch, welches zur Beschreibung dieser Eigenschaft ausgewählt wurde (PL - Polylexie). Im Modell Köhlers wurde noch eine vierte Eigenschaft berücksichtigt, nämlich die Polytextie (PT) als Zahl verschiedener Kontexte, in denen das jeweilige Lexem auftreten kann. Außerdem wurden zwei quantitative Größen gewählt, die das Phoneminventar (Phonemzahl PZ) und den Umfang des Lexikons (LG) betrafen - in den konkreten Untersuchungen der Lexik der jeweiligen Sprache sind diese Größen konstant.

Köhler verfolgte bei der Erstellung seines Modells folgende Ziele:

(a) Beschreibung der gegenseitigen Abhängigkeiten zwischen den genannten vier strukturellen Eigenschaften der Lexeme und Untersuchung des Einflusses der Phonemzahl (PZ) und der Lexikongröße auf die Parameter der entsprechenden Funktionen und auch des Wirkens der entsprechenden Zipfschen Kräfte.

(b) Ableitung weiterer quantitativer Abhängigkeiten im Lexikon zwischen anderen, bisher noch nicht berücksichtigten Lexemeigenschaften.

(c) Einbeziehen der schon bekannten Gesetze in ein allgemeines System lexikalischer Gesetze.

(d) Erklärung dieser Gesetze und deren Struktur durch deren Einbeziehung in eine übergeordnete Gesetzmäßigkeit, die das menschliche Verhalten steuert.

In Anlehnung an die Annahme Altmanns (1980) bezüglich der Proportionalität von Veränderungen, die zwischen den Variablen x und y vor sich gehen, beschrieb Köhler alle untersuchten Abhängigkeiten zwischen den vier genannten Variablen L, F, PL und PT mit Hilfe der Potenzfunktion $y = ax^b$. Dieses Modell wurde am Material des deutschen Lexikons

überprüft, wobei der Autor diese Anpassung nur als die erste Probe ansieht; das untersuchte Material weist in einigen Fällen große Divergenzen beim Vergleich mit dem vorgeschlagenen Modell auf (vgl. Köhler 1986: 116, 125 u.a.).

Kurz nach dem Erscheinen der Arbeit Köhlers wurde in Bochum ein Forschungsprogramm unter dem Namen "Sprachliche Synergetik. Tei I. Quantitative Lexikologie" formuliert, welches die Untersuchung – nach einer mehr oder weniger einheitlichen Liste von Eigenschaften – der Lexik verschiedener Sprachen zum Ziel hatte, die auf der Grundlage von einsprachigen Wörterbüchern untersucht werden, was wiederum in der Zukunft ermöglichen würde, Materialien für quantitative Untersuchungen lexikalischer Gesetze unter konfrontativem Aspekt (d.h. unter verschiedenen Randbedingungen) zu erhalten.

Eine wichtige Eigenschaft des besprochenen Projektes ist die Formulierung einer nicht begrenzten Zahl von Eigenschaften für die Untersuchung des jeweiligen Lexikons. Neben den vier im Modell von Köhler berücksichtigten strukturellen Eigenschaften (Länge L, Frequenz F, Bedeutungszahl PL und Polytextie PT) von Lexemen kann diese Beschreibung auch um solche Eigenschaften erweitert werden wie das Alter, die Herkunft, der morphologische Wortbildungsstatus der Lexeme usw. Ein größeres Repertoire von Eigenschaften würde z.B. die Untersuchung diachronischer quantitativer Sprachgesetze zulassen, die schon von Zipf formuliert wurden (Abhängigkeit zwischen der Frequenz, dem Alter und der Herkunft der Lexeme, vgl. Zipf 1949) und in den letzten Jahren gründlich von sowjetischen Forschern analysiert wurden (Arapov, Cherc 1983).

Interessante Ergebnisse dürfte auch die Untersuchung aller oben genannten synchronischen Sprachgesetze hinsichtlich einzelner Wortarten liefern, was das Modell Köhlers überhaupt nicht berücksichtigt.

Die Sprachwissenschaftler aus Bochum realisieren gegenwärtig das von ihnen erarbeitetet Forschungsprojekt, indem Materialien für die deutsche Sprache vorbereitet werden. Als Grundlage wurde das im Wörterbuch von Wahrig (1981) enthaltene Vokabular gewählt. Mit dem Forschungsprogramm wurden gleichzeitig verschiedene linguistische Zentren, die sich mit quantitativer Linguistik beschäftigen, bekannt gemacht; somit wurden in verschiedenen Ländern analoge Untersuchungen des Lexikons mehrerer Sprachen aufgenommen (darunter einer relativ großen Gruppe exotischer Sprachen).

Bezüglich der slawischen Sprachen wurden bisher (August 1988) Beschreibungen der Lexik der slowakischen und polnischen Sprache angemeldet.

3. Die polnische Version des Projekts "Sprachliche Synergetik"

3.1. Quellen

Aufgrund der Relevanz, die in quantitativ-lexikalischen Untersuchungen die Frequenz hat, wurde als Materialgrundlage für die polnische Sprache das Vokabular gewählt, welches im allgemeinen Häufigkeitswörterbuch der polnischen Schriftsprache enthalten ist. Dieses Vokabular beruht auf einem Textkorpus mit einem Umfang von n = 500000 Wörtern und umfaßt fünf Funktionalstile (das Korpus für jeden Stil umfaßt 100000 Wörter; vgl. Słownictwo 1974-77, Band 1-5). Das gesamte Vokabular umfaßt 38468 Lexeme (Kaminska-Szmaj 1983: 133); in unseren Untersuchungen wurden jedoch Eigennamen, fremdsprachige Zitate und Abkürzungen nicht berücksichtigt, wodurch die Zahl der berücksichtigten Lexeme der polnischen Sprache auf 30059 reduziert wurde (diese Zahl ist kleiner als die aus den Daten der Arbeit von Kaminska-Szmaj 1983:133 resultierende Zahl; die Differenzen sind Ergebnis der im gegenwärtig vorbereiteten synthetischen Band angewandten nicht so strengen Kriterien bei der Qualifikation der Lexeme als Entlehnungen, der besseren Differenzierung lexikalisch-grammatischer Homonyme und in einigen Fällen Ergebnis der Korrektur von Fehlern).

Im folgenden besprechen wir die in der Beschreibung der Lexeme angewandten quantitativ-qualitativen Eigenschaften. Zur besseren Übersichtlichkeit weisen wir gleichzeitig auf die entsprechende Spalte im beigefügten Musterformular hin, das die jeweiligen Daten enthält.

3.2. Graphische und phonologische Schreibweise der Lexeme (Spalte 1)

Die Lexeme wurden in graphischer und phonologischer Schreibweise aufgenommen. Als Grundlage für die phonologische Schreibweise wurde die im polnischen phonologischen System von W. Jassem (1966) verwendete Beschreibung gewählt, wo 37 Phoneme unterschieden werden – ohne Nasalvokale /ǫ/, /ę/ und ohne weiche Labiale. Somit werden in diesem System die Phoneme /l/, /y/ und /j/, auch /c/, /ʒ/, /n/, /ŋ/, /ɲ/ als unterschiedliche Phoneme behandelt (entsprechend der Interpretation des Typs

/pasek/ : /pjasek/, /pysk/ : /pisk/ und /droge/ : /droʒe/, /franka/ : /fraʒka/).

In der phonologischen Transkription der Lexeme wurden folgende Regeln angewandt:

(a) Die phonologische Schreibweise wurde nach der Aussprache im Tempo lento vorgenommen und entsprechend der Norm der Warschauer Aussprache, z.B. /pomyçny/, /attyka/, /xfawa/, /kfjat/, dargestellt.

(b) Die Gruppen /oN+S/, /eN+S/ in Morphemen der Muttersprache wurden konsequent als /ow+S/, /ew+S/ transkribiert (also wurde z.B. die Schreibweise /mewsci/, /sowçat/ angenommen). In fremden Morphemen wurden diese Gruppen entsprechend als /oNS/, /eNS/ interpretiert, d.h. es wurde die Schreibweise des Typs /konstrukcja/, /agens/ angewandt.

(c) Nasalkonsonanten in der Position vor den hinteren Verschlußkonsonanten wurden als /n/ nur in morphologischen Knoten transkribiert: /sucen-ka/, /ʃklan-ka/; in allen anderen Fällen als /ŋ/ (vgl. /raŋga/, /kleŋkatç/, /puŋkt/).

(d) In entlehnten Lexemen mit dem Affix *-logia* und in Wörtern des Typs *alergia, ironia, autarkia* wurde die phonologische Schreibweise /-loɲa/, /iroŋia/, /awtarcja/, /alerɲa/ gewählt.

(e) In der Schreibweise wurde die Neutralisierung der phonologischen Opposition im Inlaut berücksichtigt (/vjeçtç/ als Schreibweise für das Verbpaar *wieść: wieźć*) und im absoluten Auslaut (/kot/, /ʃef/ als Schreibweise für die entsprechenden Paare *kot: kod, szef: szew*; vgl. im Deutschen /ra:t/ für die Lemmata *Rat: Rad*). Diese Paare wurden unterschieden, indem die homophonen Lexeme numeriert wurden (/kot/$_1$, /kot/$_2$).

(1) Lemma Graph.	Phonolog.	(2) Silb.	(3) WA	(4) Flex.	(5) Funk.	(6) Polys.	(7) Frequ.	(8) Kont.	(9) MS
WIOŃÁ7	vjoɲɲtç	2	5	Vb	dk	1	2		
WIOSENNO-LETNI	vjosenno-letɲi	5	2			1	1		K_2
WIOSENNY	vjosenny	3	2			1	20	45,80	
WIOSKA	vjoska	2	1	fIII		1	20	62,82	
WIOSŊO	vjoswo	2	1	nIII		2	5	55,28	
WIOSŊOWA7	vjoswovatç	3	5	IV	ndk	1	1		
WIOSNA	vjosna	2	1	fIV		1	32	83,91	
WIOŚLARSKI	vjoɕlarsci	3	2			1	5	0,00	
WIOŚLARSTWO	vjoɕlarstfo	3	1	nIII	sgt	1	1		
WIOŚLARZ	vjoɕlaʃ	2	1	mII		1	4	27,11	
WIOŃCZE7	vjotʧetç	2	5	III	ndk	1	1		
WIOTKI	vjotci	2	2			2	3		
WIÓR	vjur	1	1	mIV		1	1		
WIÓROWY	vjurovy	3	2			1	1		
WIR	vir	1	1	mIV		1	2		
WIRAŻ	viraʃ	2	1	mII		2	1		
WIRNIK	virɲik	2	1	mIII		1	6	0,00	
WIROLOT	virolot	3	1	mIV		1	2		K_1

3.3. Quantitative Eigenschaften der Lexeme (Spalten 2,6,7 und 8)

3.3.1. Lexemlänge (Symbol L)

Diese Eigenschaft wurde in Spalte 2 berücksichtigt. Angegeben wurde die Zahl der Silben als diskrete Variable mit den Werten 0,1,2,3,..., da es im Polnischen Präpositionen wie *w, z* mit Null Silben gibt. Die Berechnung der Lexemlänge in Buchstaben und Phonemen und die Berechnung der mittleren Silbenlänge in Buchstaben/Phonemen für das jeweilige Lexem wurde maschinell durchgeführt.

3.3.2. Frequenz der Lexeme (Sympol F)

Diese Daten wurden in Spalte 7 aufgenommen in Anlehnung an die synthetische Liste des allgemeinen Häufigkeitswörterbuches der polnischen Schriftsprache (der synthetische Band wird gegenwärtig unter Berücksichtigung aller Bände von Słownictwo 1974–1077 vorbereitet).

3.3.3. Bedeutungszahl der Lexeme (Polylexie, Symbol PL)

Diese Eigenschaft wird in Spalte 6 aufgeführt. Die Bedeutungszahl wurde auf der Grundlage von Skorupka, Auderska, Lempicka, *Mały Słownik Języka Polskiego* (1968) angegeben. Gezählt wurden nur die Lexembedeutungen, die im untersuchten Wörterbuch als gleichgestellte Bedeutungen angesehen wurden, d.h. als die 1., 2., 3., ... Bedeutung aufgeführt wurden. Die Bedeutungsbeschreibung einer kleinen Zahl von Lexemen, die wohl im Häufigkeitswörterbuch, aber nicht im Wörterbuch von Skorupka et al. auftraten, wurde auf der Grundlage der Wörterbuchs von M. Szymczak (1978–1981) ergänzt.

3.3.4. Zahl der unterschiedlichen Kontexte der Lexeme (Polytextie, Symbol PT)

Diese Eigenschaft wurde in Spalte 8 aufgeführt. Aufgrund der sehr atomistischen Struktur des Korpusses des polnischen Häufigkeitswörterbuches, das sich aus 2000 Proben mit je 50 Wörtern zusammensetzt und für jeden Stil aus einer großen Menge verschiedener Titel ausgelost wurde (das ganze Korpus umfaßt somit 10000 Proben), war es unmöglich, diese Eigenschaft über die Angabe der Zahl verschiedener Kontexte zu erfassen, in denen das jeweilige Lexem auftrat (was mit der Erfassung dieser Eigenschaft von Köhler 1968:63 übereinstimmen würde). Solche Daten führen z.B. – aufgrund einer anderen Konstruktion des Korpusses – die Häufigkeitswörterbücher der russischen Sprache auf (vgl. Steinfeldt o.D.) und der slowakischen Sprache (Mistrík 1969).

Die Variable PT soll den Grad der Allgemeinheit erfassen, d.h. die stilistische Kennzeichnung der Wörter. Die Lexeme *atom, cybernetyka, facet* treten nur in einer kleinen Zahl stilistisch unterschiedlicher Texte auf, da sie fast ausschließlich in wissenschaftlichen Texten oder umgangssprachlichen Texten auftreten, dagegen werden die Lexeme des Typs *człowiek, być, chodzić* in einer sehr großen Zahl verschiedener Texte verwendet, d.h. sie sind stilistisch neutral.

Aufgrund der Spezifik des polnischen Korpusses wurde die Eigenschaft der Polytextie nicht als Zahl der Texte ausgedrückt, sondern unter Anwendung des Parameters D von Juilland, der den Grad der Gleichverteilung der Häufigkeit der Lexeme in den untersuchten n Funktionalstilen des Häufigkeitswörterbuchs angibt (vgl. Juilland, Edwards, Juilland 1965: XLV). Dieser Parameter nimmt Werte im Intervall <0,1> an, wobei D = 0 bedeutet, das die Polytextie gleich Null ist (das jeweilige Lexem tritt nur in einem Stil auf; z.B. das Lexem *ambasada* hatte in 5 Stilen unseres Wörterbuches folgende Häufigkeitsverteilung: 0, 25, 0, 0, 0 d.h. dieses Lexem trat nur in Texten der Pressenachrichten auf), und D = 1 bedeutet eine Gleichverteilung der Häufigkeit der Lexeme auf alle Stile. Solche Verteilungen (D fast 1) haben vor allem das grammatische Vokabular und der Grundwortschatz, welcher in allen Stilen mit etwa derselben Häufigkeit auftritt.

Die Zahlenwerte für D wurden nach

$$D = 1 - \frac{v}{\sqrt{n - 1}} \qquad (1)$$

berechnet, wo n die Zahl der Funktionalstile bezeichnet und v die Variable der Veränderlichkeit (vgl. Juilland, Edwards, Juilland 1965: XLV).

Die in Prozent angegebenen Werte für D wurden nur für Lexeme mit $F \geq 4$ berechnet.

3.5. Paradigmatische Beschreibung der Lexeme (Spalten 3,4,5)

Die im synergetischen Modell Köhlers beschriebenen Abhängigkeiten zwischen den Variablen L, F, PL und PT kann man sowohl im gesamten Lexikon untersuchen als auch im Bereich einzelner Wortarten. Für die weitere Analyse ist auch eine genauere Flexionsbeschreibung des Lexikons wichtig, d.h. die Kennzeichnung der einzelnen Lexeme nach deren Flexionsparadigmen (Deklination, Konjugation), der Bestimmung des Aspekts der Verben usw.

In Spalte 4 (Symbol WA) wurde die Klassifikation der Lexeme nach der Wortart vorgenommen - entsprechend dem im polnischen Häufigkeitswörterbuch angenommenen Zahlenkode (*Słownictwo* 1974-1977, Einführungsteil in den einzelnen Bänden). In unserer Beschreibung wurden Partikel und Adverbien unterschieden, in *Słownictwo* wurden diese als eine Wortart aufgefaßt.

Bei der Klassifikation der Lexeme nach Wortarten wurden vor allem formale Kriterien angewandt; so wurden z.B. alle Lexeme adjektivischer Deklination als Adjektive aufgefaßt (das betrifft Ordinalia, adjektivische Pronomen und Partizipien); diese Wortart erhielt das Symbol 2. Substantivierte Adjektive (*chory, pospieszny*) wurden als Substantive beschrieben (Symbol 1). Zur Klasse der Verben wurden auch Prädikativa des Typs *można, trzeba* usw. gezählt.

Unterschieden wurde auch die syntaktische Homonymie des Typs *brak* (1)//(5) (Substantiv//Verb), *koło* (1)//(6) (Substantiv//Präposition), *wolno* (5)//(8) (Verb//Adverb) und die lexikalisch-grammatische Homonymie des Typs $brak_1$ (Genitiv auf -u)//$brak_2$ (Genitiv auf -a), $boleć_1$ (*boli*)//$boleć_2$ (*boleje*). Eine genauere Beschreibung der von uns angewandten Klassifikationskriterien der Wortarten und der syntaktischen Homonymie ist in *Słownictwo* (1974-1977, Einführung) zu finden.

Das Ausmaß der im untersuchten Lexikon berücksichtigten lexikalisch-grammatischen Homonymie wird in einer gesonderten Publikation besprochen.

Die Spalten 4 und 5 (Symbole: Flex. und Funk.) enthalten eine Beschreibung der substantivischen und verbalen Lexeme (mit den Symbolen 1 und 5 in Spalte 3) nach deren Flexionsparadigmen. Hier wurde die Beschreibung der polnischen Flexion nach J.Tokarski (1968, 1973) angewandt, wobei dessen Klassifikation in einigen Fällen erweitert wurde.

Bei der Beschreibung der Verben wurden von Tokarski 11 Flexionsparadigmen vorgeschlagen (I-XI), die in unseren Analysen um das zusätzliche Paradigma XII erweitert wurden (unregelmäßige Verben des Typs *być, chcieć, iść, stać* u.a.) und um das Paradigma XIII (Prädikativa des Typs *trzeba, można, wolno, brak* usw.). Zusätzliche Symbole wurden für einige Verben gewählt, die 2 Flexionsparadigmen besitzen (z.B. *dokonywać:* I/VIIIa).

Eine kompliziertere Beschreibung erforderte die Flexion der Substantive. Neben 17 Deklinationsparadigmen, die von Tokarski unterschieden wurden, wurde ein zusätzliches Symbol für Substantive mit gemischter Deklination gewählt (z.B. *hrabia, sędzia,* auch *książę* usw.).

Außerdem wurde in Spalte 4 mit gesonderten Symbolen die Art der unflektierbaren Substantive gekennzeichnet (*kapo* - mO, *netto, brutto* nO), die Art der substantivierten Adjektive (*chory* m, *miła* f, *cudze* n) und auch die Art der Substantive mit adjektivischer Deklination (*chorąży* - mAdj, *woźna* - fAdj, *komorne* - nAdj).

Die Spalte 5 (Funk) wurde für zusätzliche Informationen zur Flexion reserviert. Hier wurden Angaben zum Aspekt für Verben (WA 5) gemacht (ndk//dk - unvollendeter//vollendeter Aspekt). Für Substantive gelten folgende Symbole:

plt//sgt - Pluralia tantum//Singularia tantum
m - Maskulina mit Deklination der Feminina in Spalte 4 (*artysta, radca:* in Spalte 4 als fIV)
m,f - Substantive mit doppeltem Genus (*kaleka, włoczega* u.a.)
Gen.-a//Gen.-u als Flexionsendungen für grammatische Homonyme (vgl. bal_1, Gen.-a//bal_2, Gen.-u).

3.6. Wortbildungsstatus der Lexeme, Komposita

Für die Beschreibung der quantitativen Struktur des Lexikons wichtig ist auch die Untersuchung von drei grundlegenden Lexemtypen hinsichtlich deren Wortbildungsbau - es handelt sich hier um Wörter, die unter Anwendung der Wortbildungsregeln nicht weiter teilbar sind (stems), Derivate und Komposita als Wörter mit mehreren Stämmen.

Eine Klassifizierung des polnischen Lexikons hinsichtlich dieser drei Kategorien war nicht möglich, da bisher für die polnische Sprache noch keine hinreichend genauen Kriterien der Unterscheidung synchronischer Derivate von synchronisch nicht teilbaren Wörtern vorliegen.

Aufgrund der letztens von G. Altmann (1988a) formulierten Hypothesen über die quantitative Struktur der Komposita haben wir uns entschieden, bei der Beschreibung des polnischen Lexikons Kompositionsklassen zu unterscheiden, um diese Hypothesen am polnischen Material überprüfen zu können.

Die Daten zur Komposition enthält Spalte 9. Eine genaue Klassifikation der polnischen Komposita für quantitative Untersuchungen geben wir in einer gesonderten Arbeit an (Sambor 1989), hier werden wir lediglich diese Klassifikation zur besseren Orientierung umreißen.

Als übergeordnete Klasse wurden Wörter mit mehreren Lexemen angesehen, die sich somit aus mindestens 2 lexikalischen Morphemen zusammensetzen. Fremdsprachige Morpheme des Typs *bio-*, *fono-*, *geo-*, *hydro-*, *tele-* usw. wurden als Präfixe angesehen, da sie in der polnischen Sprache nicht selbständig sind.

Diese Klasse teilen wir weiter in zwei Unterklassen:

- Komposita im weiteren Sinne, d.h. Wörter mit mehreren Lexemen, die formale Kennzeichen von zusammengesetzten Wörtern aufweisen,

- Kompositionsderivate (D_k), d.h. Wörter mit mehreren Lexemen, die Wortbildungsmorpheme besitzen, z.B. *parowoz-ownia* (n.loci) < *parowóz*.

Die Komposita im weiteren Sinne kann man wiederum in drei Untergruppen teilen:

- eigentliche Komposita (die Kompositionsmorpheme enthalten),

- Zusammenrückungen (K_4) als zusammengesetzte Wörter, die infolge der Zusammenwachsung von syntaktischen Gruppen entstanden, wobei in der Regel nur ein Glied flektiert wird (*wiarygodny* < *godny wiary*, *Wielkanoc* < *wielka noc*),

- Zusammensetzungen (K_5) - in den Untersuchungen wurden nur Zusammensetzungen von 2 Substantiven berücksichtigt, die mit einem Bindestrich verbunden waren und somit einen mit S-S darstellbaren Bau besitzen (*teatr-laboratorium, teatr-muzeum, chłop-robotnik*). In diesem Typ besteht nicht die Möglichkeit, die Zahl der Glieder zu erweitern.

Die eigentlichen Komposita haben charakteristische Kompositionsmorpheme des Typs

-o- (*par-o-wóz*, 'lokomotywa poruszana parą')
-i- (*łam-i-strajk*, 'ten, kto łamie strajk')
-ø- (*drog-o-wskaz-ø*, 'słup wskazujący drogę').

Sie teilen sich in zwei Klassen:

- semantisch reguläre (endozentrische): es besteht semantische Übereinstimmung zwischen der Bedeutung des ganzen Kompositums und der Bedeutung zumindest eines seiner Glieder (*drogowskaz, parowóz*),

- semantisch nicht reguläre (exozentrische): es besteht keine semantische Übereinstimmung zwischen der Bedeutung des ganzen Kompositums und der Bedeutung eines beliebigen seiner Glieder (K_3). Als Beispiel können die Komposita *pierwsz-o-rzęd-ny* 'bardzo dobry, wspaniały', *koci-o-kwik* 'zły nastrój psychiczny' dienen.

Die semantisch regulären Komposita können noch einmal in zwei Klassen unterteilt werden hinsichtlich der Möglichkeit der Erweiterung um weitere Glieder:

- semantisch reguläre und nicht erweiterbare Komposita (K_1) (z.B. *parowóz, drogowskaz, samobójstwo*),

- semantisch reguläre und erweiterbare Komposita (K_2); es handelt sich hier um einen Typ, der in der polnischen Sprache nur adjek-

tivische Komposita des Typs Adj + (Adj +...+ Adj) + Adj betrifft (z.B. *biało-czerwony, biało-czerwono-zielony, rolno-przemysłowy, rolno-spożywczo-przemysłowy* usw.).

Die für quantitative Beschreibungen angenommene Klassifikation der Wörter mit mehreren Lexemen kann man folgendermaßen darstellen (s. Abb. 1):

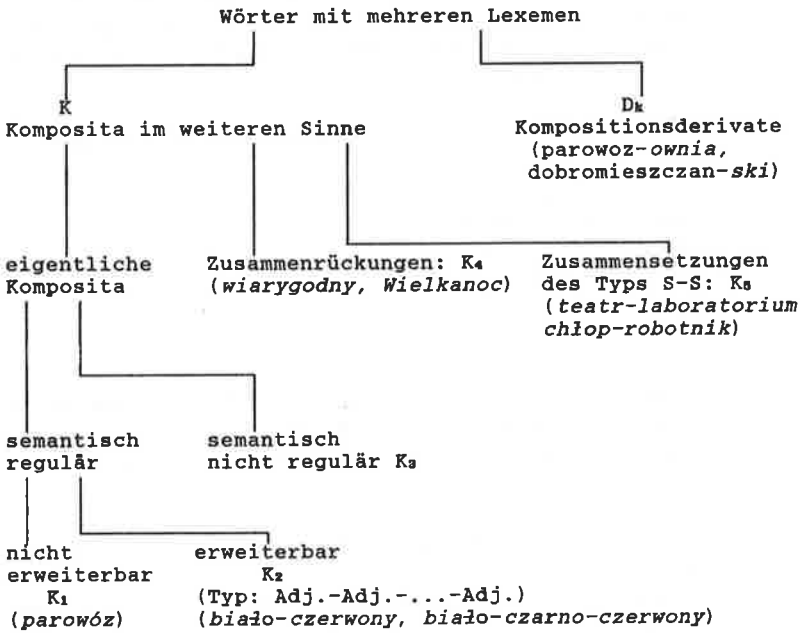


Abb. 1. Struktur der polnischen Wörtern mit mehreren Lexemen

Bei der Klassifikation des Materials wurde ein ziemlich bequemes und operatives Kriterium für die Unterscheidung der Komposita (K) von den Kompositionsderivaten (D_k) angenommen: ein Wort ist ein Kompositum, wenn in der Wortbildungsparaphrase beide Glieder getrennt auftreten (*wielkomiejski = związany z wielkim miastem)*, oder ein Kompositionsderivat, wenn in der Paraphrase das ganze Kompositum auftritt (*drobnomieszczanski = taki, jak drobnomieszanin; wielkanocny = związany z Wielkanocą*).

Bei der detaillierten Beschreibung des Materials traten Schwierigkeiten vor allem bei der Unterscheidung der adjektivischen Komposita der Klassen K_1 und K_2 auf und auch bei der Unterscheidung der Derivate D_k von den Komposita in Paaren Adjektiv:Adverb (*pierwszorzędny-pierwszorzędnie*). Spezielle Entscheidungen werden in der schon genannten Arbeit, die den polnischen Komposita gewidmet ist, besprochen (Sambor 1989).

Material, welches für die Überprüfung der Hypothesen G.Altmanns hinsichtlich der Komposita dienen kann, sind ausschließlich die Klassen K_1 – K_5, dagegen werden die Klassen der Derivationskomposita (D_k) nicht berücksichtigt, da diese Hypothesen keine Derivate betreffen. Der Umfang der Klasse D_k läßt es aber zu, die ***Wortbildungskraft*** der polnischen Komposition zu bestimmen, d.h. die Fähigkeit, Derivate zu bilden (es handelt sich um eine wichtige Variable für Vergleiche bei weiteren Untersuchungen der Komposita und Derivate, die in Texten und im Lexikon anderer slawischer Sprachen auftreten).

3.7. Weitere Perspektiven quantitativer Untersuchungen des Lexikons

Wie schon erwähnt wurde, sollen die Daten aus Spalte 9, die den Wortbildungsstatus der Lexeme betreffen, in Zukunft durch die Klasse der unter Anwendung von Wortbildungskriterien nicht weiter teilbaren Lexeme (stem, Wurzel) und der eigentlichen Derivate (D) ergänzt werden.

Die Untersuchung der gegenseitigen Abhängigkeiten der sprachlichen Eigenschaften können durch Hinzufügung weiterer Eigenschaften in den folgenden Spalten erweitert werden, z.B. durch die Eigenschaften "Alter" und "Herkunft" der Lexeme, also um den diachronischen Aspekt. Die wichtigen Arbeiten von Arapov und Cherc, die an die bahnbrechenden Arbeiten Zipfs (1949) anknüpfen und die Modellierung der Abhängigkeit zwischen der Häufigkeit der Wörter und deren Alter und Herkunft (Arapov, Cherc 1983) betreffen, können als Vorbild dienen bei der Ausnutzung der auf diese Weise erhaltenen neuen Materialien.

Der große Vorteil des Forschungsprogramms "Sprachliche Synergetik" ist die Möglichkeit, daß die Linguisten die Beschreibung noch um weitere Lexemeigenschaften erweitern können, die für die jeweilige Sprache als wichtig für die quantitative Beschreibung des Lexikons angesehen werden. Die Auswahl der Eigenschaften hängt nämlich vor allem vom Typ der jeweiligen Sprache ab.

4. Daten für das polnische Lexikon. Beschreibung des Formulars

Zur besseren Orientierung des Lesers in der Struktur der Beschreibung des polnischen Lexikons wird ein Formular mit vollständigen Angaben für 8 quantitative-qualitative Eigenschaften angeführt, die in den Punkten 3.2 bis 3.6 besprochen wurden. In dieser Form wurde das Formular für die mechanische Verarbeitung angefertigt (s. S. 177).

In Spalte 1 wird die graphische und phonologische Schreibweise der Lexeme angeführt. In der graphischen Schreibweise wurden anstelle einiger polnischer Buchstaben Ersatzzeichen aus dem deutschen Alphabet oder aus dem Kode der Ziffern verwendet (ą = ä, ó = ö, ć = 7, ł = 9 usw.).

Andere Ersatzzeichen wurden in der phonologischen Transkription angewandt, falls das entsprechende Zeichen im Repertoire der deutschen Kleinbuchstaben fehlte; um jedoch dem Leser die Identifikation der phonologischen Schreibweise zu erleichtern, wurden diese Zeichen im angeführten Formular in der internationalen Transkription angegeben.

Spalte 2 enthält die Silbenzahl der Lexeme, Spalte 3 die Wortart nach dem in *Słownictwo* (1974-1977) angewandten Kode. In Spalte 4 (Flex.) wurden für Substantive und Verben (Symbole 1 und 5 in Spalte 3) die entsprechenden Deklinations- und Konjugationsparadigmen angegeben.

Spalte 5 (Funk.) enthält Informationen über den Aspekt von Verben und eine zusätzliche Beschreibung für einige substantivische Klassen (z.B. Singularia tantum//Pluralia tantum). In Spalte 6 (Polys.) und 7 (Frequ.) werden die Bedeutungszahl der Lexeme, die aus dem ausgewählten Wörterbuch abgelesen wurde, und die Frequenz aus *Słownictwo* (1974-1977) angeführt. In Spalte 8 (Kont.) werden die Werte für die Dispersion D von Juilland (in Prozent) angeführt, die den Grad der Gleichverteilung der Häufigkeiten der Lexeme in den 5 untersuchten Funktionalstilen der polnischen Schriftsprache angibt (*Słownictwo* 1974-1977). Es sei daran erinnert, daß der Index D nur für Lexeme berechnet wurde, deren Häufigkeit $F \geq 4$ ist.

Im analysierten Formular hat das Lexem *wiosna* einen relativ hohen Grad der Gleichverteilung der Häufigkeit (100D = 83.91), d.h., daß dieses Lexem in einer großen Zahl verschiedener Kontexte auftritt. Den geringsten Grad der Gleichverteilung (100D = 0) erhielten wir für die Lexeme *wiośłarski* und *wirnik*, was bedeutet, daß diese Lexeme in einer sehr begrenzten Zahl verschiedener Kontexte auftraten (*wiośłarski* nur in

Sportnachrichten des Stils der Pressenachrichten, *wirnik* nur in wissenschaftlichen Texten).

Die letzte Spalte 9 (MS) enthält gegenwärtig lediglich Informationen bezüglich der oben besprochenen polnischen Kompositionstypen und der Derivationskomposita. In Zukunft werden diese Daten um Informationen erweitert, die die Unterscheidung von Wurzeln und eigentlichen Derivaten betreffen.

5. Erste Ergebnisse

Diese Ergebnisse werden in der Reihenfolge der einzelnen Spalten des besprochenen Formulars angeführt. Die Bearbeitung und Interpretation dieser Daten ist Gegenstand weiterer Publikationen.

Spalte 1: Graphische und phonologische Schreibweise der Lexeme

Auf der Grundlage der Daten aus Spalte 1 erhielten wir:

(a) die Statistik der einzelnen Phoneme und die Matrix der Verbindungen von Phonemen des Typs VV, CC, CV, VC;

(b) die Verteilung der Lexeme hinsichtlich ihrer Länge (in Zahl der Buchstaben).

Die statistischen Daten für Phoneme werden in einem gesonderten Artikel angeführt und interpretiert; hier soll nur die empirische Verteilung der Lexeme hinsichtlich deren Länge (in Buchstaben) angeführt werden (vgl. Tabelle 1).

Spalte 2: Silbenzahl

Als erste Daten erhielten wir die empirische Verteilung der Lexemlänge in Silben (vgl. Tabelle 2). Diese Ergebnisse wurden verwendet, um Wortlängenverteilungen in Silben im Häufigkeitswörterbuch und im ganzen Lexikon zu vergleichen (vgl. Hammerl, Sambor in diesem Band).

Tabelle 1

Wortlängenverteilungen in Buchstaben für das ganze Lexikon (Textwörterbuch) und für 4 Wortarten

Wortlänge x_i	Häufigkeit n_i				
	Alle	Substantive	Verben	Adjektive	Adverbien
1	12	1	0	0	0
2	55	3	0	2	2
3	321	218	16	11	10
4	906	701	69	52	32
5	1865	1296	222	185	95
6	2760	1653	487	410	152
7	3580	1758	869	767	143
8	4097	1699	1033	1126	195
9	4126	1622	955	1343	183
10	3696	1373	891	1241	168
11	2956	1098	643	1054	151
12	2084	766	404	805	98
13	1384	478	251	577	74
14	874	294	121	410	46
15	503	175	49	251	25
16	300	104	30	155	10
17	169	56	12	95	6
18	109	28	7	72	2
19	68	13	2	53	
20	51	9		42	
21	45	8		37	
22	26	4		22	
23	15	1		24	
24	10	0		10	
25	14	0		14	
26	6	1		5	
27	4	0		4	
28	5	0		5	
29	2	0		2	
30	0	0		0	
31	3	0		3	
32	0	0		0	
33	0	0		0	
34	1	0		1	
35	0	0		0	
36	0	0		0	
37	0	0		0	
38	2	1		1	
	30059	13360	6061	8779	1392

Spalte 3: Wortarten

Infolge der Berechnungen erhielten wir 2 Ranglisten: nach der Lexemzahl und nach deren Frequenz (vgl. Tabelle 3). Diese Daten können mit der Statistik der Wortarten in *Słownik Jezyka Polskiego* unter Leitung von W. Doroszewski (Warszawa 1958–1968, in 10 Bänden) verglichen werden, d.h. mit den Daten des größten polnischen Lexikons (diese Daten wurden von Z. Saloni berechnet und bisher nicht veröffentlicht).

Tabelle 2

Wortlängenverteilungen in Silben für das ganze Lexikon (Textwörterbuch) und für 4 Wortarten

Wortlänge x_i	Häufigkeit n_i				
	Alle	Substantive	Verben	Adjektive	Adverbien
0	7	1	0	0	0
1	1240	894	148	24	35
2	6151	3755	1328	574	326
3	9800	4284	2543	2332	522
4	7824	2894	1532	2997	368
5	3395	1122	413	1744	113
6	1110	311	78	696	25
7	331	78	18	233	2
8	127	10	1	116	
9	43	9		34	
10	20	1		19	
11	7	0		7	
12	2	0		2	
13	0	0		0	
14	0	0		0	
15	1	0		1	
16	0	0			
17	1	1			
Σ	30059	13360	6061	8779	1391
$\bar{x}$	3.39	3.07	3.17		
s^2	1.66	1.46	0.98		

Tabelle 3

Statistik der Wortarten: Ranglisten

Wortart	Lexemrangliste		Wortart	Frequenzrangliste	
	Anzahl	Frequenz		Frequenz	Anzahl
1 Subst	13361	144188	1 Subst	144188	13361
2 Adj	8779	79995	2 Adj	79995	8779
5 Verb	6061	71988	5 Verb	71988	6061
8 Adverb	1391	21428	6 Präp	56812	75
10 Part	141	18757	9 Konj	33605	83
9 Konj	83	33605	4 Pron	31833	26
7 Int	79	650	8 Adverb	21428	1391
6 Präp	75	56812	10 Part	18757	141
3 Num	63	8976	3 Num	8076	63
4 Pron	26	31833	7 Int	650	79

Im polnischen Material wurden auch Lexeme, die zwei, drei oder vier Wortarten angehören, gesondert gekennzeichnet (sog. syntaktische Homonyme). X soll nun die Zahl der Wortarten bezeichnen, denen das jeweilige Lexem angehören kann, und f_x die Zahl der Lexeme, die den Werten der Variablen X entsprechen. Die Verteilung der Lexeme nach deren Zugehörigkeit zu X Wortarten zeigt dann Tabelle 4.

Tabelle 4

Verteilung der syntaktischen Homonyme im polnischen Häufigkeitswörterbuch

Zahl der Wortarten X	Zahl der Lexeme f_x	Frequenz
1	29632	397145
2	204	60928
3	4	6772
4	1	2486

Beispiele für Lexeme, die zwei Wortarten angehören, können die Adjektive und substantivierten Adjektive (Adj//Subst) sein vom Typ

pospieszny, chory, Lexeme des Typs *brak, żal* (Subst//Verb), die als Substantive oder Prädikativa auftreten, u.a.

Lexeme, die drei oder vier Wortarten angehören, gibt es in der polnischen Sprache nur wenige. Zur ersten Klasse gehört z.B. das grammatische Lexem *to*, das die Funktion eines substantivischen Pronomens (Symbol 4) (*to 4 mi sie podoba*), einer Konjunktion (Symbol 9) (*to 9 wchodził, to 9 wychodził*) und einer Partikel (Symbol 10) (*coraz to 10 ciemniej*) ausübt.

Nur das grammatische Morphem *co* trat in vier Wortarten auf, als substantivisches Pronomen (Symbol 4) (*co tam widzisz?*), als Präposition (Symbol 6) (*co 6 roku, co 6 godzinę, co 6 kilka metrów*), als Konjunktion (Symbol 9) (*ten, co 9 nie miał szczęścia*) und als Partikel (Symbol 10) (*co 10 sie tak patrzysz?*).

Als erste Ergebnisse führen wir auch eine Zusammenstellung für die drei größten Gruppen der syntaktischen Homonyme auf, d.h. für die Typen Subst//Adj (1//2), Subst//Verb (1//5) und Präp//Adv (6//8). Dies zeigt Tabelle 5.

Tabelle 5

Statistik der drei größten Gruppen von Lexemen, die 2 Wortarten angehören

Typ der syntaktischen Homonymie: Wortarten	Anzahl f_x	Frequenz
Subst//Adj (1//2)	122	6436
Subst//Verb (1//5)	14	1861
Präp//Adv (6//8)	19	1150

Eine genaue Analyse dieser und anderer Lexemgruppen, die mehreren Wortarten angehören, wird Gegenstand einer selbständigen Arbeit sein.

Spalte 4: Flexion

Als erste Ergebnisse führen wir nur die Rangliste für die Deklinations- und Konjugationsparadigmen (Tabellen 6 und 7) an. Diese Daten entsprechen dem polnischen Häufigkeitswörterbuch und können - wie im Falle der Wortarten - mit den entsprechenden Daten aus einem großen polnischen Lexikon verglichen werden (Materialien von Z. Saloni zur Fle-

xion in *Słownik Języka Polskiego* unter Leitung von W. Doroszewski; diese Daten wurden bisher nicht veröffentlicht).

Tabelle 6

Flexionsparadigmen der Substantive im polnischen Häufigkeitswörterbuch

Symbol des Flexionsparadigmas	Anzahl y_x	Frequenz
m (*chory*)	159	1081
m0 (*kapo*)	5	6
mAdj (*woźny*)	31	171
mI	357	6021
mII	569	6121
mIII	1473	14755
mIV	2557	34288
mV	40	107
mVI	8	73
f (*chora*)	54	341
f0	1	7
fAdj (*generałowa*)	27	84
fI	1011	11677
fII	273	2553
fIII	1196	8326
fIV	1407	20034
fV	729	7626
fVI	32	1263
n (*cudze*)	21	38
n0 (*netto*)	20	39
nAdj (*komorne*)	4	4
nI	2624	17370
nII	149	2121
nIII	403	6475
nIV	14	175
nV	8	245
nVI	56	683
0 (*bantu*)	9	25

Die Informationen zur Flexion, die Spalte 5 (Funktion) enthält, werden in diesem Text umgangen, da diese Daten, die z.B. den Aspekt der Verben betreffen, den Anteil der Substantive des Typs Singularia tantum//Pluralia tantum, in eine gesonderte Arbeit aufgenommen werden, die die polnische Flexion betrifft.

Tabelle 7

Flexionsparadigmen der Verben im polnischen Häufigkeitswörterbuch

Symbol des Flexionsparadigmas	Anzahl y_x	Frequenz
I	1728	13865
II	8	450
III	78	317
IV	970	4804
Va	218	1011
Vb	78	541
Vc	108	633
VI	8	132
VIa	1115	10413
VIb	456	3935
VIc	1	1
VII	5	529
VIIa	61	2485
VIIb	56	1084
VIII	1	1
VIIIa	203	1107
VIIIb	62	244
IX	331	3120
Xa	112	886
Xb	54	251
Xc	63	1184
XI	200	3224
XII	103	19292
XIII	16	1859
I/IX	9	15
I/VIIIa	6	57
Va/Vc	2	6

Die Daten der Spalten 6, 7 und 8 enthalten drei grundlegende quantitative Eigenschaften von Lexemen: die Polylexie, Frequenz und Polytextie. Diese Daten erlauben bei zusätzlicher Berücksichtigung der Daten aus Spalte 2 (Wortlänge) eine Untersuchung der gegenseitigen Abhängigkeiten zwischen diesen Eigenschaften, was auch Gegenstand einer gesonderten Arbeit ist.

Spalte 9: Morphologischer Status

Bisher wurden nur zusammengesetzte Wörter der Gruppen K_1 – K_5 untersucht (vgl. Abbildung 1). Eine exakte Analyse der erhaltenen Daten und eine Überprüfung der Hypothesen von Altmann bezüglich Komposita wird Gegenstand weiterer Publikationen sein.

Jetzt wollen wir nur die ersten Ergebnisse zeigen, die die Anzahl und Frequenz der Komposita betreffen (Tabelle 8 und Abbildung 2).

Tabelle 8

Komposita (Typen K_1 - K_5) im polnischen Häufigkeitswörterbuch

Typ	Anzahl	Frequenz
K_1	1029	4406
K_2	149	226
K_3	77	626
K_4	3	137
K_5	50	55

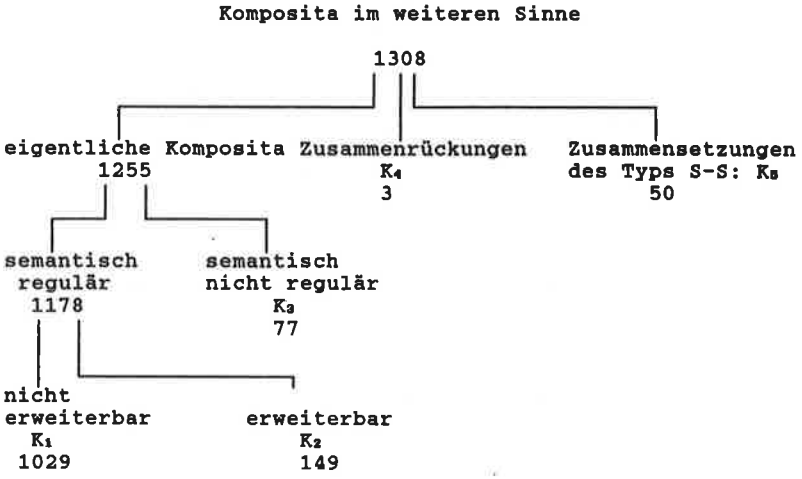


Abb. 2. Quantitative Struktur der Komposita im polnischen Häufigkeitswörterbuch

Alle oben dargestellten Daten stellen die Grundlage für weitere Untersuchungen dar. Ziel dieser Untersuchungen ist vor allem die Überprüfung des synergetischen Modells von R. Köhler im Zusammenhang mit der Flexionsstruktur im Lexikon.

Literatur

Altmann, G. (1978), Towards a theory of language. *Glottometrika 1, 1-12.*

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

Altmann, G. (1988a), Hypotheses about compounds. In diesem Band.

Altmann, G. (1988b), Verteilungen der Satzlängen. *Glottometrika 9, 147-169.*

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

Arapov, M.V., Cherc, M.M. (1983), Mathematische Methoden in der historischen Linguistik. Bochum, Brockmeyer.

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984), Wortlänge und Bedeutungskomplexität. *Glottometrika 6, 116-126.*

Fucks, W. (1955), Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Köln.

Gerlach, R. (1982), Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. *Glottometrika 4, 93-102.*

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

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

Guiter, H., Arapov, M.V. (eds.) (1982), Studies on Zipf's law. Bochum, Brockmeyer.

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

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

Jassem, W. (1966), The distinctive features of the Polish phoneme system. *Biuletyn Polskiego Towarzystwa Językoznawczego 24, 87-108.*

Juilland, A., Edwards, P.M., Juilland, I. (1965), Frequency dictionary of Rumanian words. The Hague, Mouton.

Kaminska-Szmaj, I. (1983), Części mowy w słowniku i w tekście pięciu stylów funkcjonalnych polszczyzny pisanej na materiale słownika frekwencyjnego. *Biuletyn Polskiego Towarzystwa Językoznawczego 40, 127-136.*

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

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

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

Köhler, R., Altmann, G. (1988), Synergetic modelling of language phenomena. In: Köhler, R. (ed.), Studies in language synergetics (erscheint).

Krylov, Ju. K. (1982), Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: Guiter, H., Arapov, M.V. (eds.), Studies on Zipf's law. Bochum, Brockmeyer, 234-262.

Krylov, Ju. K. (1984), Ob odnoj paradigme lingvostatisticeskich raspredelenij. In: Soontak, J. (Hrsg.), Lingvostatistika i vycislitel'naja lingvistika. Tartu, TGU, 80-97.

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

Mistrík, J. (1969), Frekvencia slov v slovenčine. Bratislava, Vydavatel'stvo Slovenskej Akadémie vied.

Orlov, Ju. K. (1982a), Dynamik der Häufigkeitsstruktur. In: Orlov, Ju.k., Boroda, M.G., Nadarejšvili, I.Š., Sprache, Text, Kunst. Quantitative Analysen. Bochum, Brockmeyer 82-117.

Orlov, Ju. K. (1982b), Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, Ju.K., Boroda, M.G., Nadarejsvili, I.S., Sprache, Text, Kunst. Quantitative Analysen. Bochum, Brockmeyer, 1-55.

Papp, F. (1967), Bearbeitung des ungarischen Wortschatzes auf Lochkartenmaschine. *Acta Linguistica Academiae Scientiarum Hungaricae* 17, 141-172.

Sambor, J. (1984), Menzerath's law and the polysemy of words. *Glottometrika 6, 152-176.*

Sambor, J. (1989) Struktura statystyczna wyrazów złożonych - zarys problematyki. Poradnik Językowy (erscheint).

Skorupka, S., Auderska, H., Łempicka, Z. (1968), Mały słownik jezyka polskiego. Warszawa, Państwowe Wydawnictwo Naukowe.

Słownictwo współczesnego języka polskiego. Listy frekwencyjne (1974-1977), Warszawa, Polska Akademia Nauk.

Steinfeldt, E. (o.D.), Russian word count. Moscow, Progress Publishers.

Szymczak, M. (1978-1981), Słownik języka polskiego. B. I-III. Warszawa, Państwowe Wydawnictwo Naukowe.

Tokarski, J., (1968), Formy fleksyjne. In: Skorupka, Auderska, Łempicka (1968), IX-XXI.

Tokarski, J. (1973), Fleksja polska. Warszawa, Państwowe Wydawnictwo Naukowe.

Wahrig, G. (1981), Wörterbuch der deutschen Sprache. München, Bertelsmann.

Zipf, G.K. (1935), The psycho-biology of language. Boston, Houghton Mifflin.

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

Hammerl, R. (ed.), Glottometrika 10, 1988

Vergleich der Längenverteilungen von Lexemen nach der Silbenzahl – im Lexikon und im Textwörterbuch

Rolf Hammerl, Kielce
Jadwiga Sambor, Warschau

1. Einführung

In der polnischen Version des Forschungsprojektes "Sprachliche Synergetik", die in diesem Band besprochen wird (vgl. Sambor, S. 171 ff.), wurde darauf hingewiesen, daß als Grundlage der Untersuchungen des Vokabulars nicht das gesamte Lexikon eines bestimmten einsprachigen Wörterbuchs diente, sondern das Vokabular des polnischen Häufigkeitswörterbuches (Słownictwo 1974–1977). Das polnische lexikalische Material umfaßt somit nicht das Lexikon, sondern ein Textwörterbuch (HWB).

In dem genannten Artikel wurden – als erste Ergebnisse – die empirischen Längenverteilungen der Lexeme im polnischen HWB angeführt, die sowohl das gesamte Vokabular als auch Substantive, Verben, Adjektive und Adverbien betreffen (vgl. Tabelle 2 im Artikel von Sambor in diesem Band).

Es entsteht somit die Frage, ob sich die von uns erhaltenen Längenverteilungen der Lexeme nach der Silbenzahl, die sich auf das Textwörterbuch beziehen, signifikant unterscheiden und ob diese Unterschiede auch die mittleren Lexemlängen (gemessen in der Zanl der Silben) betreffen.

Als Vergleichsdaten zu den empirischen Verteilungen der Lexeme im HWB (vgl. Tabelle 2 im Artikel von Sambor in diesem Band) wurden 3 empirische Lexemverteilungen gewählt, die wir aus 3 Stichproben, welche aus dem Wörterbuch von Skorupka, Auderska, Łempicka (1968) ausgelost wurden, erhalten haben. Diese Proben mit einem jeweiligen Umfang von etwa 1000 Lexemen umfassen a) das gesamte Vokabular, b) Substantive, c) Verben; es handelt sich also hier um empirische Längenverteilungen der Lexeme nach der Silbenzahl im polnischen Lexikon (vgl. Tabelle 1).

Tabelle 1

Empirische Längenverteilungen der Lexeme nach der Silbenzahl im polnischen Lexikon

Silbenzahl x_i	Ges. Vokabular n_i	Substantive n_i	Verben n_i
1	107	93	55
2	328	338	342
3	378	371	362
4	177	177	193
5	49	43	67
6	6	12	12
7	1	1	1
Σ	1046	1035	1032
$\bar{X}$	2,7657744	2,7864735	2,9370156
s^2	1,0933219	1,0974016	1,1074671

2. Differenzen zwischen den mittleren Längen der Lexeme im HWB und im Lexikon

Unter Anwendung der Teststatistik u mit der Verteilung N(0,1) und

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \qquad (1)$$

die für den Vergleich von Mittelwerten im Falle großer Proben angewandt wird, wurden die in Tabelle 2 angeführten Werte u berechnet (unter Anwendung der Daten aus Tabelle 1 für die Längenverteilung im Lexikon und der Daten aus Tabelle 2 des Artikels von Sambor in diesem Band (S. 189) für die Längenverteilungen im HWB). Die Ergebnisse dieser Berechnungen führt Tabelle 2 auf.

Tabelle 2

Beurteilung der Differenzen der mittleren Lexemlängen im HWB und im gesamten Lexikon

Typ der Probe		x_i	s_i^2	n_i	u_{emp}
Gesamtes Lexikon	HWB	3.3909	1.6609	30059	18.84
	Lexikon	2.7658	1.0933	1046	
Substantive	HWB	3.0722	1.4602	13360	9.36
	Lexikon	2.7865	1.0974	1035	
Verben	HWB	3.1724	0.9798	6061	6.70
	Lexikon	2.9370	1.1075	1032	

Die Differenzen zwischen den mittleren Lexemlängen sind signifikant in allen drei Proben, wobei die größten Differenzen in der Probe auftreten, die alle Lexeme des Vokabulars umfaßt. Dies wird vermutlich durch den relativ großen Anteil von zusammengesetzten Adjektiven im Textwörterbuch verursacht, die in der polnischen Sprache potentielle Bildungen sind und nicht im System (im Lexikon) auftreten. Diese Adjektive können im Text sehr lange Ketten bilden - z.B. Farbbezeichnungen wir *brudnobrązowozielonożółty* usw.

3. Differenzen zwischen den Längenverteilungen der Lexeme nach der Silbenzahl im HWB und im Lexikon

Zur Überprüfung, ob zwei Stichproben aus Grundgesamtheiten mit derselben Verteilung stammen, werden in der Fachliteratur mehrere Tests vorgeschlagen (z.B. der Serientest, der Test von Smirnow, der Medianentest, der Wilcoxon-Test - vgl. Gren 1987:490 ff.; Metody statystyczne 1980: 140 ff.; Platt 1978:160 ff.; Storm 1974:248 ff.), die jedoch nur stetige Verteilungen in der Grundgesamtheit betreffen. In unserem Beispiel haben wir es jedoch mit diskreten Verteilungen zu tun.

Da für die Beschreibung der Verteilung von Eigenschaften mit diskreten Variablen im Bedarfsfall auch die Anwendung von stetigen

Verteilungen zulässig ist (besonders dann, wenn es keine andere Beschreibungsmöglichkeit gibt), soll für den Vergleich der Längenverteilungen der Lexeme im HWB und im Lexikon der schon oben genannte Test von Smirnow angewandt werden. Dies ist möglich, weil bei der Anwendung dieses Tests eine endliche Zahl von Intervallen der stetigen Variablen, repräsentiert durch eine endliche Zahl von Zahlenwerten, verglichen wird. Wir nehmen somit an, daß die Zahlenwerte unserer diskreten Variablen stellvertretend für bestimmte Intervalle einer stetigen Variablen stehen (z.B. steht unser Wert $x_i = 1$ - die Länge des Lexems beträgt 1 Silbe - für das Intervall $(0,1\rangle$ einer stetigen Variablen y - die Länge des Lexems liegt im Intervall $0 < y \leq 1$).

Tabelle 3

Differenzen zwischen den Längenverteilungen der Lexeme nach der Silbenzahl im HWB und im Lexikon

x_i	Empirische Häufigkeiten		Empirische Distribuanten		Differenzen der Distribuanten
	HWB	Lexikon	WB	Lexikon	
0	7	0	0.00023	0.00000	0.00023
1	1240	107	0.04149	0.10229	0.06080
2	6151	328	0.24611	0.41586	0.16975
3	9800	378	0.57214	0.77724	0.20510
4	7824	177	0.83242	0.94646	0.11404
5	3395	49	0.94537	0.99330	0.04793
6	1110	6	0.98230	0.99904	0.01674
7	331	1	0.99331	1.00000	0.00669
8	127	0	0.99753	1.0	0.00247
9	43	0	0.9989	1.0	0.0011
10	20	0	0.9996	1.0	0.0004
11	7	0	0.9999	1.0	0.0001
12	2	0	0.9999	1.0	0.0001
13	0	0	0.9999	1.0	0.0001
14	0	0	0.9999	1.0	0.0001
15	1	0	0.9999	1.0	0.0001
16	0	1	0.9999	1.0	0.0001
17	1	0	1.0000	1.0	0.0000
Σ	30059	1046			

Die Anwendung eines solchen verteilungsfreien Tests hat den Vorteil, daß außer der Stetigkeit der Verteilung keine weiteren Annahmen über diese Verteilung gemacht werden müssen, wie es oft bei anderen Tests der Fall ist. Die Gütefunktion dieser Tests ist aber etwas schlechter als die von parametrischen Tests, d.h., daß die Nullhypothese über das Fehlen von signifikanten Differenzen zwischen den zu überprüfenden Verteilungen in vielen Fällen nicht abgelehnt werden kann, wo andere Tests signifikante Unterschiede aufzeigen. "Wird jedoch bereits mit einem verteilungsfreien Test ein Unterschied erkannt, d.h. die Nullhypothese abgelehnt, wenn sie falsch ist, so kann man mit dem entsprechenden Parametertest kein anderes Ergebnis erzielen" (Storm 1974:249).

Ähnlich wie beim Vergleich der Mittelwerte überprüfen wir die Hypothese der Übereinstimmung der Verteilungen für 3 Proben: a) für alle Lexeme, b) für Substantive, c) für Verben.

Die Anwendung des Smirnow-Testes zeigen wir am Beispiel der ersten Probe (alle Lexeme). Die wichtigsten Daten zur Berechnung des Zahlenwertes der Teststatistik werden in Tabelle 3 aufgeführt.

Der Test wird wie folgt durchgeführt:

1. Man berechnet die empirischen Distribuanten; dazu errechnet man für jedes x_i die kummulierten empirischen Häufigkeiten und dividiert diese durch die jeweilige Gesamthäufigkeit. Den zweiten Wert der empirischen Distribuanten für das HWB erhält man z.B. aus:

$$\frac{n_1 + n_2}{n_a} = \frac{7 + 1240}{30059} = 0.04149.$$

2. Man bildet die Differenzen der entsprechenden Werte der beiden empirischen Distribuanten (absoluter Betrag) und sucht den größten Wert unter allen Differenzen heraus. Dieser Wert, der mit dem Symbol $d(n_1 n_2)$ bezeichnet wird, beträgt bei uns 0.2051.

3. Man berechnet einen Wert n' nach

$$n' = \frac{n_a n_b}{n_a + n_b} . \qquad (2)$$

Bei uns gilt:

$$n' = \frac{30059(1046)}{30059 + 1046} = 1010.8251.$$

4. Man berechnet die Testgröße

$$X_{emp} = n'^{1/2} \cdot d(n_1 n_2). \qquad (3)$$

Für unser Beispiel gilt wiederum:

$$X_{emp} = (1010.8251)^{0.5} \cdot 0.2051 = 6.5208.$$

5. Diese Testgröße X_{emp} vergleicht man nun mit einem Wert, den man für ein vorgegebenes Signifikanzniveau (bei uns: $\alpha = 0.05$) aus den Tabellen der Grenzverteilung der Statistik ablesen kann. Für $\alpha = 0.05$ beträgt dieser Wert etwa 1.36.

6. Da $X_{0.05} < X_{emp}$ ist, wird die Nullhypothese abgelehnt, d.h., beide Proben (Vokabular im HWB, Vokabular im Lexikon) stammen aus Grundgesamtheiten mit unterschiedlichen Verteilungen, d.h. aus unterschiedlichen Grundgesamtheiten.

Dasselbe Ergebnis erhält man bei der Anwendung dieses Tests auf die beiden anderen Proben (Substantive, Verben), wobei die empirischen Testwerte folgende Größenordnung annehmen:

Substantive : $X_{emp} = 3.29$
Verben : $X_{emp} = 4.19.$

Das Ergebnis dieser Untersuchungen, welches durch das Ergebnis des Vergleichs der Mittelwerte im Kapitel 2 noch unterstützt wird, lautet: Das Vokabular des Textwörterbuchs und das Vokabular des Lexikons gehören zwei verschiedenen Grundgesamtheiten an. Dies wurde an drei Stichproben (alle Lexeme, Substantive, Verben) gezeigt. Dieses Ergebnis muß in weiteren Untersuchungen überprüft werden, auch unter Anwendung anderer Tests. Der von uns angewandte Test, der ja eigentlich für stetige Variablen gilt, kann uns somit nur erste Informationen liefern, die es in Zukunft zu verifizieren gilt.

4. Schlußbemerkung

Wenn das oben erhaltene Ergebnis in anderen Untersuchungen bestätigt wird, so müssen in allen lexikalischen Untersuchungen zwei Grundgesamtheiten unterschieden werden: das Vokabular das Textwörterbuches und das Vokabular des Lexikons. Ergebnisse, die in Untersuchungen einer der beiden Grundgesamtheiten gewonnen werden, dürfen dann nicht ohne weiteres auf entsprechende Sachverhalte der anderen Grundgesamtheit übertragen werden. Somit wären dann die Ergebnisse der Untersuchungen zur polnischen Sprache innerhalb des Projekts "Sprachliche Synergetik. Quantitative Lexikologie" nicht mit entsprechenden Ergebnissen aus Untersuchungen des Lexikons anderer Sprachen vergleichbar, da sich die polnischen Untersuchungen auf ein Textwörterbuch stützen.

Literatur

Gren, Jerzy (1987), Statystyka matematyczna. Warzawa: Państwowe Wydawnictwo Naukowe.

Krzysztofiak, M., Urbanek, D. (1980), Metody statystyczne. Warszawa: Państwowe Wydawnictwo Naukowe.

Platt, C. (1987), Problemy rachunku prawdopodobienstwa i statystyki matematycznej. Warszawa: Państwowe Wydawnictwo Naukowe.

Skorupka, S., Auderska. H., Łempicka. Z. (1968), Mały słownik języka polskiego. Warszawa: Państwowe Wydawnictwo Naukowe.

Słownictwo współczesnego języka polskiego (1974-77), Listy frekwencyjne. Tom I-V. Warszawa.

Storm, R. (1974), Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. Leipzig: VEB Fachbuchverlag.

Hammerl, R. (ed.), Glottometrika 10, 1988.

A syntactic variable on the text level

Luděk Hřebíček, Prague

The question to be dealt with in the present paper is, whether the sentence syntax paticipates in some way in the structure of the text. Syntax can be understood as a feature which does not surpass the limits of the sentence, and it is usually comprehended in this sense. However, when the text is understood as an entity being in growth, then its dynamics must effect all linguistic levels. Consequently, we seek an expression for mutual relations of different linguistic levels in a growing text.

For this reason a simple variable is introduced which characterizes certain properties of the sentence structure. Its textological (i.e., suprasentence) properties are given as the sum of its values for all sentences of an analyzed text. This variable is related to an expression for co-references derived elsewhere (Hřebíček 1985) and tested on a sample of Turkish texts. Expressions for the growing text are formulated on the basis of the observed variables, as are the number of lexical units, number of sentences and the text length in number of words. The theory is tested here on a corpus of Turkish texts.

1. Syntactic variables

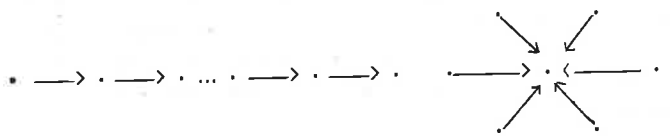
The sought after syntactic variable should be derived from a given grammatical theory, from a theoretically consistent description of sentence structure. However, we want to be independent of grammatical theories; we seek a syntactic variable applicable within the framework of different descriptions.

Let us suppose that a description results in a graph expressing a sentence structure. Such graphs have been drawn by representatives of different linguistic positions encompassing the classical theory in terms of Subject-Predicate relations, as well as by followers of Noam Chomsky.

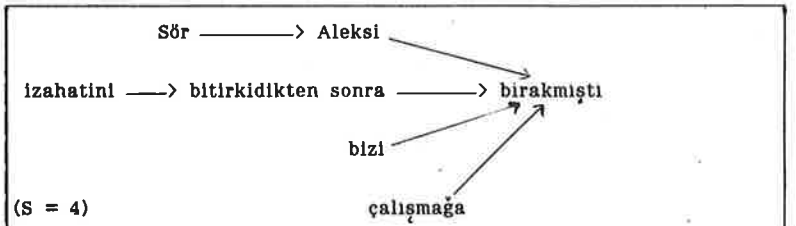
Regardeless of the type of the graph, the variable can be visualized as a ***concentration of a graph***; certain points (syntactic units) of the graph are incident to more arrows or lines than other points of the same graph; these units are highly syntactically determined in comparison with other units; and consequently, a sentence structure provides a higher concentration in comparison with another sentence. It is possible to introduce a variable characterizing this property which varies from sentence to sentence.

Let us describe, how this variable is derived for Turkish sentences. Each text is a sequence of sentences free of everything that is not a sentence (figures, graphs, complicated numerical or algebraical expressions, isolated interjections, etc.). Each basic syntactic unit is either a sentence nucleus, or a unit directly or indirectly (through other syntactic units) determining the nucleus. Nucleus is *verbum finitum* in a broader sense, it is defined by predication; this syntactic function is fulfilled by verbal or non-verbal forms. The structure of a sentence thus consists of syntactic pairs, each pair having a determined unit and a determining unit (the latter also in a zero-form). This hypotactic structure is described in detail in Hřebíček (1971). It is evident that such an understanding is not in concordance with the traditional linguistic treatement of syntax; it is a simplification worth of being confronted with what was written on hypotaxis by L. Johanson (1975) and by other authors quoted in his paper. For example, the Subject-Predicate relation is described by us as Subject -> Predicate, Adverb-Predicate as Adverb -> Predicate, etc.; instead of classical concepts, morphological characteristics are used. All syntactic pairs of a sentence are unified into a directed graph with a finite verb as its sink (no arrow goes out of it), cf. Fig. 1 containing the graph of the first sentence of RN (cf. Supplement 2).

From a viewpoint of concentration, two terminal types of graphs can be indicated:



All other graphs are certain combinations of the two indicated types. It is evident that the degree of concentration of a graph is in direct relation to the number of units into which no arrow enters. Let us call it S. The same variable S expresses the number of ***complete paths*** of a graph, each path leading from a terminal unit (not entered by an arrow) to the finite verb. This syntactic formation can be understood as a ***complete construction*** (syntagm). S is, consequently, an expression of structural properties os a sentence. Its advantage consists of the fact that it is a fairly observable quantity. It can be also understood as an additive quantity; a text or part of a text can be characterized by the sum of values proper to its sentences.



Sör Aleksi, izahatini bitirkidikten sonra bizi çalişmağa birakmisti.
"Sister Aleksi let us work after finishing her explanation."

Figure 1. The structure of a sentence by R.N. Güntekin

2. Co-references

One of the observed textological quantities is the number of co-references *z* (or in short: references). An example of the way in which co-references and other textological units were identified in Turkish text is presented in Supplement 1. The analyzed Turkish texts are listed in Supplement 2. It must be stressed that other rules for identification of co-references and other variables can be stated, but the relations of the textological variables discussed below should not change with the exception of coefficients of proportionality.

According to our conception, co-references represent semantic identities between the basic textological units. Furthemore these units are

endowed with grammatical and lexical properties; each of them can be qualified as lexical and grammatical forms. In short, they are word-forms.

The determination of identities between these units requires a subjective semantic evaluation. A correct analysis should be therefore done by a group of native analysts whose results are averaged for the purpose of obtaining objective number of z. Unfortunately, this approach could not be applied here; the analysis was made by the author himself; certain rules were applied in order to minimize faults and incorrect decisions, cf. Supplement 1.

Table 1 contains the values of variables observed in the corpus of Turkish texts; these cumulative data were obtained from samples consisting at first of sentences 1 to 10, then of 1 to 20, etc., and finally of 1 to 100 of the given text. The number of sentences in each sample is marked k.

The expected number of co-references z_e was derived in Hrebícek (1985) and is given by the relation

$$z_e = \frac{akv}{n}, \qquad (1)$$

where k = number of sentences
v = number of lexical units
n = text length in number of word-forms
a = coefficient of proportionality.

The derivation of (1), to put it briefly, starts with two suppositions:

(1) The more lexical units are repeated in a text, the fewer co-references occur in it.

(2) The more sentences occur in a text, the more references occur in it.

Together with observed z and expected z_e , the values of z_s are also presented in Table 1; this variable is defined as follows:

$$z = \frac{Sv}{n}, \qquad (2)$$

where S is the sum of complete paths of a given text.

It has been observed that between

z and z_e
z and z_s
z_e and z_s

there exists a mutual relation. With the help of the F-test, the significance of differences between the variances of the indicated pairs of variables were tested. In none of the thirty tests (three pairs in ten texts) was a significant difference ascertained.

The same observed values were further tested by the help of the t-test for difference of means observed in two samples (with the presumption of $\sigma_1 = \sigma_2$ proved by F-tests). In all thirty tests no significant difference between the tested means was ascertained.

Hence it follows that z, z_e and z_s are variables which are statistically equivalent.

Table 1

Variables observed in the corpus of Turkish texts

(The values in lines of this table represent parts of the texts; each begins from its first sentence and reaches k = 10, 20,...,100; consequently, the values are cumulative. z_e = akv/n; z_s = Sv/n)

RN (a = 4.52)

k	v	n	z	S	z_e	z_s
10	42	45	19	24	42.19	22.40
20	150	190	53	90	71.37	71.05
30	202	298	82	133	91.92	90.15
40	247	378	105	171	118.14	111.74
50	315	525	137	238	135.60	142.80
60	348	588	151	270	160.51	159.80
70	396	719	182	335	174.16	184.51
80	422	779	196	364	195.89	197.19
90	465	886	218	416	213.50	218.33
100	485	957	233	452	229.07	229.07

NC (a = 2.67)

k	v	n	z	S	z_e	z_s
10	89	117	25	57	20.31	43.36
20	137	185	42	92	39.54	68.13
30	173	234	61	117	59.22	86.50
40	195	272	76	141	76.57	101.08
50	213	302	91	155	94.16	109.32
60	245	367	106	188	106.95	125.50
70	268	417	120	211	120.12	135.61
80	286	457	129	230	133.68	143.94
90	300	495	137	247	145.64	149.70
100	321	538	146	267	159.31	159.31

RI (a = 3.51)

k	v	n	z	S	z_e	z_s
10	89	114	21	52	27.40	40.60
20	140	198	47	94	49.64	66.46
30	176	266	64	124	69.67	82.05
40	214	337	87	161	89.16	102.24
50	254	406	105	196	109.80	122.63
60	286	478	130	228	126.01	136.42
70	309	539	149	263	140.86	150.77
80	348	612	168	298	159.67	169.45
90	389	687	182	330	178.87	186.86
100	405	727	192	351	195.54	195.54

MÖ (a = 4.02)

k	v	n	z	S	z_e	z_s
10	102	126	21	58	32.54	46.95
20	174	232	46	106	60.30	79.50
30	235	327	78	157	86.67	112.83
40	270	387	104	190	112.19	132.56
50	314	476	128	234	132.59	154.36
60	358	552	157	273	156.43	177.05
70	380	614	181	305	174.16	188.76
80	410	687	199	337	191.93	201.12
90	437	781	217	373	202.44	208.71
100	458	837	237	402	219.97	219.97

YG (a = 1.81)

k	v	n	z	S	z_e	z_s
10	36	43	12	25	15.15	20.93
20	62	76	25	46	29.53	37.53
30	87	108	40	66	43.74	53.17
40	115	152	57	87	54.78	65.82
50	133	178	70	102	67.62	76.21
60	147	204	81	119	78.26	85.75
70	155	223	90	131	88.07	91.05
80	172	255	98	147	97.67	99.15
90	195	299	114	165	106.24	107.61
100	211	329	129	181	116.08	116.08

DC (a = 4.86)

k	v	n	z	S	z_e	z_s
10	67	82	33	41	39.71	33.50
20	140	197	65	94	69.08	66.80
30	212	314	100	149	98.44	100.60
40	265	421	120	199	122.37	125.26
50	291	502	134	234	140.86	135.65
60	330	610	153	286	157.75	154.72
70	357	680	171	321	178.61	168.53
80	396	794	186	370	193.91	184.53
90	430	907	209	421	207.37	199.59
100	470	1057	232	486	216.10	216.10

MA (a = 4.68)

k	v	n	z	S	z_e	z_s
10	87	121	22	56	33.65	40.26
20	163	273	57	124	55.89	74.04
30	204	378	77	177	75.77	95.52
40	250	504	99	233	92.86	115.58
50	305	643	120	291	111.00	138.03
60	331	734	138	333	126.63	150.17
70	341	758	141	347	147.38	156.10
80	363	844	165	392	161.03	168.60
90	374	905	178	424	174.06	175.22
100	397	998	202	468	186.17	186.17

OP (a = 3.72)

k	v	n	z	S	z_e	zs
10	51	62	19	32	30.60	26.32
20	114	172	48	80	49.31	53.02
30	151	236	63	113	71.41	72.30
40	203	319	83	148	94.69	94.18
50	251	409	111	195	114.15	119.67
60	287	503	132	244	127.35	139.22
70	304	546	151	268	144.98	149.22
80	329	611	172	301	160.25	162.08
90	348	667	192	329	174.68	171.65
100	382	743	207	372	191.26	191.26

NM (a = 2.52)

k	v	n	z	S	z_e	zs
10	33	34	12	19	24.46	18.44
20	79	85	28	46	46.84	42.75
30	112	137	49	73	61.80	59.68
40	154	198	71	102	78.40	79.33
50	186	255	86	129	91.91	94.09
60	242	345	104	176	106.06	123.46
70	261	381	119	197	120.84	134.95
80	275	408	136	214	135.88	144.24
90	291	440	153	236	150.00	156.08
100	299	464	169	252	162.39	162.39

YK (a = 2.10)

k	v	n	z	S	z_e	zs
10	27	29	10	16	19.55	14.90
20	54	71	23	39	31.94	29.66
30	92	127	40	66	45.64	47.81
40	117	170	54	86	57.81	59.19
50	130	206	67	105	66.26	66.26
60	145	239	76	123	76.44	74.62
70	162	283	87	147	84.15	84.15
80	178	316	95	165	94.63	92.94
90	192	354	108	187	102.51	101.42
100	216	400	119	210	113.40	113.40

3. Explanation

The insignificance of differences between z and z_e is explained by the theory quoted in connection with (1). As far as the other two relations are concerned, from the statistical equivalence mentioned in the preceding section it can be deduced that the following relation holds:

$$S = ak. \tag{3}$$

This means that S is a linear function of the number of sentences k.

The empirical validity of (3) was proved by the Wilcoxon-White test, in which the values of S were compared with those of ak. The coefficient a was estimated as a constant for each text of our corpus from the value corresponding to k = 100; for this reason, for each text only 9 pairs of values were tested, i.e., the values corresponding to k = 10, 20,...,90. In all ten texts the observed u was less then $u_{0.05}$; this means that between the tested variables, S on the one hand and ak on the other hand, there is no significant difference. It seems to be evident that S really is a linear function of k.

Expression (3) is an equation, which can be multiplied by an arbitrary value, for example, by 1/n. Then, if we put

$$b = 1/a,$$

(3) can be correctly written in the form:

$$\frac{n}{S} = \frac{bn}{k} \tag{4}$$

The left-hand side of (4) is evidently a characteristic of concentration expressing the mean number of words which occur in a complete path (the complete construction) of a given text. The fraction n/k is the mean number of words in a sentence.

With respect to its empirical verification, formula (1) can be written as

$$\frac{v}{z} = \frac{1}{a}.\frac{n}{k} \; . \tag{5}$$

The right-hand side of (5) equals to that of (4). Therefore it holds that

$$\frac{n}{S} = \frac{v}{z} \,. \qquad (6)$$

We try to explain (6) on the basis of two hypothetical assumptions which may not seem to be too evident and convincing, but if certain textological imagination is used, they may be taken as acceptable.

The first assumption is as follows: With constant z, the increase of vocabulary Δv of a text is proportional to the increase of syntactic characteristics $\Delta(n/S)$. New lexical units of a growing text are not included in relations of reference - cf. the supposition of constant z; this situation requires more ample syntactic structure which results in greater increase $\Delta(n/S)$.

The second assumption is: With constant vocabulary v, the increase $\Delta n/S$ is proportional to the given value of n/S; however, with respect to the supposed indirect relation between n/S and z, the negative value of z must be taken into account in the relation between $\Delta(n/S)$ and Δz.

The assumption can be written as

$$\Delta\left(\frac{n}{S}\right) \approx \frac{1}{z}\,\Delta v$$

and

$$\Delta\left(\frac{n}{S}\right) \approx \frac{n}{S}\cdot\frac{1}{z}\,(-\Delta z) = \frac{v}{z}\cdot\frac{1}{z}\,(-\Delta z)\,.$$

The right-hand side of this equation implies the following assumption: The variable $\Delta(n/S)$ is proportional to the mean number of lexical units falling upon one reference; this means, it is proportional to v/z. We thus obtain two differential equations:

$$\frac{\delta(n/S)}{\delta v} = \frac{1}{z} \qquad \text{and} \qquad \frac{\delta(n/S)}{\delta z} = -\frac{v}{z^2}\,. \qquad (7)$$

Their solutions are

$$\frac{n}{S} = \frac{v}{z} + C_1 \qquad \text{and} \qquad \frac{n}{S} = \frac{v}{z} + C_2, \qquad (8)$$

where C_1 and C_2 are integration constants. With assumption $C_1 = C_2 = C$ from the initial value $n_0/S_0 = 0$ we obtain $C = 0$. Thus from our assumptions equation (6) is obtained. According to our conviction, this relation is important as it comprises relations between the syntactic, lexical and semantic characteristics of a text.

Let us add that values of bn/k and v/z were tested by t-tests for paired values (cf. Table 2) at the $\alpha = 0.05$ level; all their results were insignificant. Consequently, between these pairs of values there are no significant differences.

Table 2

Comparison of (1) bn/k and (2) v/z in the corpus of Turkish texts

Text \ k		10	20	30	40	50	60	70	80	90	100
RN	(1)	1.00	2.10	2.20	2.09	2.32	2.17	2.27	2.15	2.17	2.12
	(2)	2.21	1.67	2.46	2.35	2.30	2.30	2.18	2.15	2.13	2.08
NC	(1)	4.38	3.46	2.92	2.55	2.26	2.29	2.23	2.14	2.06	2.01
	(2)	3.56	3.26	2.83	2.57	2.34	2.31	2.23	2.22	2.19	2.20
RI	(1)	3.25	2.82	2.53	2.40	2.31	2.27	2.19	2.18	2.17	2.07
	(2)	4.24	2.98	2.75	2.46	2.42	2.20	2.07	2.07	2.14	2.11
MÖ	(1)	3.13	2.89	2.71	2.41	2.37	2.29	2.18	2.14	2.16	2.08
	(2)	4.86	3.78	3.01	2.60	2.45	2.28	2.10	2.06	2.01	1.93
YG	(1)	2.38	2.10	1.99	2.10	1.97	1.88	1.76	1.76	1.83	1.81
	(2)	3.00	2.48	2.18	2.01	1.90	1.81	1.72	1.76	1.71	1.64
DC	(1)	1.69	2.03	2.15	2.17	2.07	2.09	2.03	2.07	2.07	2.18
	(2)	2.03	2.15	2.06	2.21	2.17	2.17	2.09	2.13	2.06	2.03
MA	(1)	2.59	2.92	2.69	2.69	2.75	2.61	2.31	2.25	2.15	2.13
	(2)	3.95	2.86	2.65	2.53	2.54	2.40	2.42	2.20	2.10	1.97
OP	(1)	1.67	2.31	2.11	2.14	2.20	2.25	2.10	2.05	1.99	2.00
	(2)	2.68	2.38	2.40	2.45	2.26	2.17	2.01	1.91	1.81	1.85
NM	(1)	1.35	1.69	1.81	1.96	2.02	2.28	2.16	2.02	1.94	1.84
	(2)	2.75	2.82	2.29	2.17	2.16	2.33	2.19	2.02	1.90	1.77
YK	(1)	1.38	1.69	2.02	2.02	1.96	1.90	1.93	1.88	1.87	1.90
	(2)	2.70	2.35	2.30	2.17	1.94	1.91	1.86	1.87	1.78	1.82

4. Paradoxes of an infinitely increasing text

Let us suppose an ideal text increasing *ad infinitum*. With a sufficiently high n, the increase of vocabulary approaches zero: $\Delta v \to 0$. Thus v changes into a constant.

As far as the relation k/n (which is the inverted value of a mean sentence length) is concerned, in a sufficiently large text it can be also supposed to be a constant.

Thus the relation (1) is transformed into product of three constants: a, k/n and v; consequently, z becomes also a constant quantity.

This indicates that the supposition of an infinitely increasing text is unacceptable. If a text increases, z must increase, too. Under the condition of $\Delta z \to 0$, the text ceases its existence as a text, its structure disintegrates. The mentioned assumption is incompatible with the dynamic understanding of a text.

In a natural language, changes of cocabulary (Δv can be positive or negative) are inevitable. The vocabulary of a language varies in time. Thus time enters into the linguistic theories as an unexplicit but non-negligible constituent. The supposition of a language with a fixed number of lexical units is thus unacceptable. Language is not a simple set of complicated symbols and rules of their usage; time is one of its basic dimensions.

We obtain the same result when the starting supposition is connected with the expression $z = Sv/n$.

Conclusions

Syntax appears to be a phenomenon surpassing the limits of the sentence. However, what was valid for S should be valid for each variable appearing to be a function of a number of sentences, cf. (3). We believe that the indicated properties of the text were formulated in a sufficiently general way, so that we may expect they will also prove valid in other languages and their texts.

SUPPLEMENT 1

Rules used in the statistic inquiry of Turkish texts

The rules are formulated ad hoc; they can be modified in the case of another inquiry and the described relations should not change with the exception of coefficients of proportionality. We present here only several examples of these rules.

1. Sentence is defined by verbum finitum.
2. Gerundial constructions are functionally adverbs related to some further verbal forms in the sentence, mostly to the finite verb.
3. Conditional forms with *-sa* are heads of constructions related to the finite verb, or they are transformed into the finite verb.
4. Direct speech, together with indirect speech which introduces it, is taken as two sentences.
5. The verbal base distinguishes lexical units: *oturmak* and *oturtmak* are two different lexical units.
6. *olmak*, regardless its meaning, is always counted as one lexical unit.
7. Diminutives do not form different lexical units.
8. *-li* and *-siz* derive different lexical units.
9. Intensive forms of adjectives represent one lexical unit: *upuzun* is the same unit as *uzun*.
10. Plurals of personal pronouns are taken as different lexical units: *onlar* is not identical with *o*.
11. Adverbs with the suffix *-ca* form new lexical units (nevertheless, *boyunca* is a postposition).
12. Compound numerals are taken as one unit: *on iki* is one unit of the text length n and one unit of the vocabulary v.
13. Indefinite article *bir* is not taken as a unit.
14. Proper nouns composed of several nouns are counted as one unit.
15. Compound verbs are taken as one unit if written together, and as two units if written separately; consequently, orthography is decisive.
16. Doubled forms are taken as one unit.
17. Indefinite numerals of the type *üç dört* are counted as one unit.
18. *ne var ki...* is a construction composed of three units; it has an adverbial function and *ne* represents a co-reference to the context.
19. *Kuzey-Batı* is one unit.

20. Grammatical relations within a sentence are not taken as co-references; the sentence *Ben geldim* contains one reference (*ben* refers to the speaker or to the author of the analyzed text). Relations surpassing the limits of the sentence are taken as references: in the sentence *Eve geldim* the predicative suffix of the verb *-im* represents the occurrences of one reference.

Etc.

SUPPLEMENT 2

If not indicated otherwise, each text was analyzed from its beginning to its 100th sentence inclusive.

RN - Resat Nuri Güntekin, Çalikuşu, Onuncu Baskı, Istanbul 1957.
NC - Necati Cumali, Yağmurlar ve Topraklar, Istanbul 1973.
RI - Rifat Ilgaz, Karartma Geceleri, Istanbul 1974.
MÖ - Mehmet Önder, Mevlâna Celâleddîn-i Rumî, Ankara 1986.
YG - Necati Cumali, Yarali Geyik, Oyun (A manuscript. Only sentences of drammatical dialogs were analyzed.)
DC - Demirtaş Ceyhun, Folklor, Sanata Düşman mi?, Nesin Vakfı Edebiyat Yılliği '84, p. 173 ff.
MA - Prof. Dr. Mehmet Akalin, Modern Lengüstige Giriş, Izmir 1983.
OP - Orhan Pamuk, Beyaz Kale, Istanbul 1985, p. 11 ff.
NM - Nezihe Meriç, Calgıcı, in: Nezihe Meriç, Bozbulanık, Öyküler, Istanbul 1981.
YK - Yaşar Kemal, Teneke, 4. baski, Istanbul.

References

Halliday, M.A.K., Hasan, R. (1976), Cohesion in English. London, Longman.
Hřebíček, L. (1971), Turkish grammar as a graph. Prague, Academia.
Hřebíček, L. (1985), Text as a unit and co-references. In: Ballmer, Th.T. (ed.), Linguistic dynamics. Berlin - New York, de Gruyter, 190-198.
Johanson, L. (1975), Some remarks on Turkic "Hypotaxis". Ural-Altaische Jahrbücher 47, 104-118.

Annotations

Alekseev, P.M., Metodika kvantitativnoj tipologii teksta. Učebnoe posobie (Methods of quantitative typology of text. A text-book). Leningrad, Leningradskij Gosudarstvennyj Pedagogičeskij Institut im. A.I. Gercena 1983, 75 pp.

The components of the term "quantitative typology of text" mean that: (1) this branch of linguistics uses *quantitative* ideas and methods; (2) its aim is to get quantitative descriptions of the *typological* properties of linguistic systems presented in a text; (3) text (generally and individually) is the primary object of linguistic observation, so that one can and really does come to typologies of languages, sublanguages, styles, etc., through the typology of texts or, to be more exact, through the typology of *text*.

The statistical-probabilistic approach is the major tool in quantitative text-typological studies. The core of this approach is formed by the problems of linguistic sampling and those of linguistic distributions. The former are the subject of Ch. I: "Sampling observation in QTT", the latter are the subject of Ch. II: "Linguistic distributions in QTT". Ch. III: "Linearity and non-linearity of rank distributions" gives a rather detailed review of rank distributions using data from more than 30 frequency dictionaries. The dependency of rank-distribution form on different factors is shown. A non-linear form of Zipf's law is offered.

Appendices I-II contain algorithms of linear and non-linear fitting of rank-frequency logarithmic graphs.

Alekseev, P.M. Kvantitativnaja tipologija teksta. Učebnoe posobie k speckursu (Quantitative typology of text. A text-book for students and post-graduate students). Leningrad, Leningradskij Gosudarstvennyj Pedagogičeskij Institut im. A.I. Gercena 1987, 80 pp.

Quantitative typology of text (QTT) is defined as the branch of linguistics that aims at modelling complex linguistic objects such as language (system and norm), speech (usage and speech "proper"), sublanguage, functional style and idiolect which are presented in an averaged and an individual text. QTT is based on linguistic, system-theoretical, probabi-

listic, semiotic and informational concepts of language and speech and uses respective methodology and techniques.

Central to QTT is the notion of probability and frequency distribution. Distribution is taken as a quantitative model of a linguistic system; each time a linguist tries to count something he deals with a distribution.

The book is composed of an introduction and two Chapters, Conclusion and Bibliography (34 works are cited and 19 frequency dictionaries used to illustrate the analysis).

Chapter I: "The linguistic basis of QTT" shows the role and place of quantitative ideas and methods in contemporary linguistics and gives a description of some fundamental linguistic notions in terms of system-theoretical, probabilistic, semiotic and information theoretical approaches.

Chapter II: "The statistical-probabilistic basis of QTT" provides a linguistic explanation for notions of probability, frequency and distribution, and outlines the levels of statistical-probabilistic descriptions of a linguistic system object. Considerable attention is paid to the most popular case of distribution analysis in linguistics - to that of the length-frequency distributions of a lexical text unit and its vocabulary.

Lesochin, M.M., Luk'janenkov, K.F., Piotrowski, R.G., Introduction to mathematical linguistics. An application of elements of mathematics to linguistics. Minsk, Nauka i tehnika 1982, 263 pp.

This introductory book covers all the major areas of mathematical linguistics and semiotics with its applicatiuon to text-processing and information retrieval systems. The volume is divided into six chapters:

1. Mathematical models of natural language (fuzziness in language and text, the language sign, the sign in the language of mathematics.
2. Applications of set theory and realtional algebra to modelling lexical systems.
3. Modelling semantic translation by means of mapping (semantic pattern recognition via thesaurus and frame).
4. A theory of artificial language.
5. Probability in linguistics.
6. Natural language message and the measurement of its information.

The book would be highly suitable as a text for undergraduate and graduate courses in linguistics and computer science, and is the only

book of its kind currently available. An English translation is in preparation.

Muchamedov, S.A., Piotrowski, R.G., Inženernaja lingvistika i opyt sistemno-statističeskogo issledovanija tjurkskich tekstov (Engineering linguistics and the systemic-statistic investigation of Turkic texts). Taškent, Fan 1986, 163 pp.

This book can be read as a kind of compendium of techniques devised by computational ("engineering") and statistical linguistics for the automated description, analysis and processing of a Turkic text. The volume is divided into four chapters:

1. Semiotics and communication
2. Linguistic models (computer generative models of Kirghiz, Azerbaijan and Uzbek words)
3. A quantitative model of Uzbek texts
4. Machine translation of Turkish texts.

CURRENT BIBLIOGRAPHY

GENERAL

1. BLUHME, H. (Hrsg.): Beiträge zur quantitativen Linguistik. Gedächtniskolloquium für Eberhard Zwirner. Tübingen, Narr 1988, 252 pp.

MORPHOLOGY

2. ANCHEEV, S.N., KUZ'MIN, L.A., LUGOVSKIJ, V.P., SIL'NICKIJ, G.G.: Issledovanie svjazi derivacionnych i sintaksičeskich charakteristik anglijskich glagolov metodom korreljacionnogo analiza [Examination of the relation of English verbs by means of correlation analysis]. Bartkov 1983, 108-124.

3. BARTKOV, B.I. (Ed.): Issledovanie derivacionnoj podsistemy količestvennym metodom [Quantitative study of the derivational subsystem]. Vladivostok, DVNC AN SSSR 1983.

4. BARTKOV, B.I.: Količestvennye metody v derivatologii (na materiale nemeckogo jazyka) [Quantitative methods in the study of derivation (using German data)]. Bartkov 1983, 3-40.

5. BARTKOV, B.I.: Slovoobrazovatel'naja nominacija v termino-systemach i norme [Labeling by word-formation in terminological systems and in the linguistic norm]. Vladivostok, DVO AN SSSR 1987.

6. KUZ'MIN, L.A.: Fonetičeskie i morfematičeskie faktory affiksal'noj sočetaemosti otglagol'nych prilagatel'nych v sovremennom anglijskom jazyke [Phonetical and morphological factors of the combinability of deverbative adjectives with affixes in present-day English]. Bartkov 1987, 105-131.

7. TICHONOV, A.N.: Sistema russkogo slovoobrazovanija v sorte količestvennych dannych [The system of Russian word formation from the quantitative point of view]. Bartkov 1983, 61-73.

8. TOTTIE, G.: Is there an adverbal in this text? (And if so, what is it doing there?) Sankoff 1986, 139-152

HISTORICAL LINGIUISTICS

9. BARTKOV, B.I., FEDUKINA, A.V.: Vozniknovenie, struktura i funkcija derivacionnoj modeli s suffiksom -(o)logy v anglijskom jazyke [Origin, structure and function of the derivation model with the suffix -(o)logy in English]. Bartkov 1987, 20-48.

10. KYTÖ, M.: On the use of the modal auxiliaries /can/ and /may/ in Early American English. Sankoff 1986, 123-138.

11. NEVALEINEN, T.: The development of preverbal /only/ in Early Modern English. Sankoff 1986, 111-121.

12. POPLOCK, Sh., Walker, D.: Going through (*L*) in Canadian French. Sankoff 1986, 173-198.

13. SANKOFF, D. (Ed.): Diversity and Diachrony. Amsterdam, Benjamins 1986.

LANGUAGE AQUISITION

14. BROEDER, P., EXTRA G., HOUT, R.v., STRÖMQVIST, S., VOIONMAA, K.: Process in the developing lexicon. Tilburg-Göteborg, University, Dep. of Linguistics. IV+160 pp.

15. HESSE, H., HESSE, B.: Wortschätze der Grundschule. Probleme ihrer Beschreibung. Wagner 1987, 82-101.

16. RICHARDS, B.: Type/token ratios: what do they really tell us? Journal of Child Language 14, 1987, 201-209.

17. WAGNER, K.R., ALTMANN, G., KÖHLER, R.: Zum Gesamtwortschatz der Kinder. Wagner 1987, 128-142.

18. WAGNER, K.R. (Hrsg.): Wortschatz-Erwerb Bern, Lang 1987.

PHONOLOGY

19. ALTMANN, G.: Ein Test für tendenzielle Vokalharmonie. Bluhme 1988, 167-170.

20. RICHTER, H.: Der multivariate Zusammenhang intonatorischer Merkmale. Bluhme 1988, 206-222.

21. TOBIN, V.: Two quantitative approaches to phonology: A contrastive analysis. Bluhme 1988, 71-112.

SOCIOLINGUISTICS

22. ASH, Sh., MYHILL, J.: Linguistic correlates of inter-ethnic contact. Sankoff 1986, 33-44.

23. GRAFF, D., LABOV, L., HARRIS, L.A.: Testing listener's reactions to phonological markers of ethnic identity: a new method for sociolinguistic research. Sankoff 1986, 45-58.

TEXT ANALYSIS

24. BERNET, Ch.: Faits lecicaux. Richesse du vocabulaire. Resultats. Thoiron 1988,1-11.

25. BRAINERD, B.: Two models for the type-token relation with time dependant vocabulary reservoir. Thoiron 1988, 13-22.

26. BRUNET, E.: La structure lexicale dans l'oeuvre de Hugo. Thoiron 1988, 23-42.

27. DUBROCARD, M.: Evaluation de l'étendue du lexique. Quelques essais de simulation. Thoiron 1988, 43-66.

28. ATEŞMAN, E.: Phonometrie und Textrezeption. Bluhme 1988, 171-182.

29. AUGST, O.: Ist die degressive Struktur des Wortgebrauchs ein Argument für den Rechtschreibgrundwortschatz (RGW). Wagner 1987, 115-127.

30. HOLMES, D.I.: The analysis of literary style. A review. Thoiron 1988, 67-76.

31. HUBERT, P., LABBE, D.: Note sur l'approximation du loi hypergeometrique par la formule de Muller. Thoiron 1988, 77-91.

32. HUBERT, P., LABBE, D.: Une modèle de partition du vocabulaire. Thoiron 1988, 93-114.

33. SCHACH, E. :Empirische Eigenschaften der TTR bei ausgewählten Texten. Wagner 1987, 102-114.

34. SCHWIBBE, M.H., RÄDER K., RICHTER, Th.: Kontentanalytische Untersuchung für Sprache Schizophrener. Bluhme 1988, 183-196.

35. SERANT, D.: A propos des modèles de reccourcissement des textes. Thoiron 1988, 115-123.

36. SERANT, D., THOIRON, Ph.: Richesse lexicale et topograpie des formes répétés. Thoiron 1988, 125-139.

37. THOIRON, Ph.: Richesse lexicale et classement des textes. Thoiron 1988, 141-163.

LEXICOLOGY

38. KÖHLER, R.: Selbstregulation der Lexik. Bluhme 1988, 156-166.

39. MENARD, N.: Mesure de la richesse lexicale. Paris, Slatkine 1983.

40. THOIRON, Ph., LABBE, D., SERANT, D. (Eds.): Études sur la richesse et la structure lexicale. Paris-Genève, Champion- Slatkine 1988, X + 172 pp.

DIALECTOLOGY

41. STEHL, Th.: Kommunikative Dialektologie oder Dialektometrie? Bluhme 1988, 238-247.

B B S

Bochumer Beiträge zur Semiotik

Ziele: Interdisziplinäre Beiträge zu praktischen und theoretischen Themen der Semiotik.

Erscheinungsweise: Unregelmäßige Abstände, ca. 5 - 10 Bände pro Jahr: Mono- graphien, Aufsatzsammlungen zu festgesetzten Themen, Kolloquiumsakten usw.

Herausgeber: Walter A. Koch (Bochum)

Herausgeberbeirat: Bernard Bichakjian (Nijmegen), Karl Eimermacher (Bochum), Achim Eschbach (Essen), Udo L. Figge (Bochum), Roland Harweg (Bochum), Elmar Holenstein (Bochum), Werner Hüllen (Essen), Frithjof Rodi (Bochum), Klaus Städtke (Berlin) .

Bände: lieferbar (*) und in Vorbereitung (bis 1989):

***Bd. 1:** HOLENSTEIN, Elmar, *Sprachliche Universalien*. xix + 250 S., pb (paperback) DM 49.80, ISBN 3-88339-419-X (12/85).

***Bd. 2:** ZHOU, Hengxiang, *Determination und Determinantien: Eine Untersuchung am Beispiel neuhochdeutscher Nominalsyntagmen*. xii + 267 S., pb DM 49.80, ISBN 3-88339-412-2 (3/85).

***Bd. 3:** KOCH, Walter A., *Philosophie der Philologie und Semiotik*. xv + 269 S., illus., pb DM 49.80, ISBN 3-88339-413-0 (1/87).

Bd. 4: KOCH, Walter A. (ed.), *For a Semiotics of Emotion*. Ca. 180 S., pb ca. DM 29.80, hc ca. DM 44.80, ISBN 3-88339-415-7

***Bd. 5:** ESCHBACH, Achim (ed.), *Perspektiven des Verstehens*. xix + 157 S., pb DM 34.80, ISBN 3-88339-414-9 (10/86).

***Bd. 6:** CANISIUS, Peter (ed.), *Perspektivität in Sprache und Text*. Ca. 230 S., pb ca. DM 34.80, ISBN 3-88339-416-5 (7/87).

***Bd. 7:** EISMANN, Wolfgang, GRZYBEK, Peter (eds.), *Semiotische Studien zum Rätsel*. Ca. 280 S., pb ca. DM 69.80, ISBN 3-88339-417-3 (6/87).

Bd. 8: KOCH, Walter A. (ed.), *Semiotik in den Einzelwissenschaften*. Ca. 1000 S., hc ca. DM 194.80, ISBN 3-88339-418-1

***Bd. 9:** SENNHOLZ, Klaus, *Grundzüge der Deixis*. xxvi + 314 S., pb DM 64.80, ISBN 3-88339-462-9 (10/85).

***Bd. 10:** KOCH, Walter A., *Evolutionäre Kultursemiotik*. xxii + 321 S., illus., pb DM 64.80, ISBN 3-88339-463-7 (3/86).

***Bd. 11:** CANISIUS, Peter, *Monolog und Dialog*. xxvi + 366 S., pb DM 69.80, ISBN 3-88339-464-5 (2/87).

***Bd. 12:** JOB, Ulrike, *Regulative Verben im Französischen: Ein Beitrag zur semantischen Rekonstruktion des internen Lexikons*. Ca. 230 S., pb ca. DM 59.80, ISBN 3-88339-487-4

***Bd. 13:** SCHMIDT, Ulrich, *Impersonalia, Diathesen und die deutsche Satzgliedstellung*. xxx + 368 S., pb DM 74.80, ISBN 3-88339-494-7 (3/87).

Bd. 14: KOCH, Walter A., POSNER, Roland (eds.), *Semiotik und Wissenschaftstheorie*. Ca. 350 S., pb ca. DM 59.80, ISBN 3-88339-554-4

Bd. 15: FIGGE, Udo L. (ed.), *Semiotik: Interdisziplinäre und historische Aspekte (BSC-Annalen II)*. Ca. 250 S., pb ca. DM 44.80, ISBN 3-88339-555-2

Bd. 16: KOCH,Walter A. (ed.), *Vom Gen zum Gedicht: Zur Geschichte der Stereotypie in der Zeichenverwendung (BSC-Annalen III)*. Ca. 250 S., pb ca. DM 44.80, ISBN 3-88339-596-X

Bd. 17: KOCH, Walter A. (ed.), *Aspekte einer Kultursemiotik*. Ca. 250 S., pb ca. DM 44.80, ISBN 3-88339-611-7

Bd. 18: KOCH, Walter A. (ed.), *Workshop in Evolutionary Cultural Semiotics. (DGS-Annalen)* .

Bd. 19: KOCH, Walter A. (ed.), *Semiosen und ihre Geschichte (BSC-Annalen IV)*. Ca. 200 S., pb ca. DM 38.80, ISBN

***Bd. 20:** KUGLER-KRUSE, Marianne, *Die Entwicklung visueller Zeichensysteme. Von der Geste zur Gebärdensprache*. Ca. 270 S., pb ca. DM 49.80, ISBN 3-88339-662-1 (8/88)

***Bd. 21:** FLEISCHER, Michael, SAPPOK, Christian, *Die populäre Literatur. Analysen literarischer Randbereiche an slavischem und deutschem Material*. Ca. S., pb ca. DM 69.80, ISBN 3-88339-647-8 (8/88)

Neuere und detailliertere Informationen zur Reihe (z.B. aktuelle Preisliste) sowie Bestellungen (Reihe oder Einzelbände) beim Verlag:
Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630 Bochum-Querenburg. Tel. (0234) 701360 oder 701383.

September 1988

BPX

BOCHUM PUBLICATIONS IN
EVOLUTIONARY CULTURAL SEMIOTICS

Aim and Scope: Transdisciplinary contributions to the analysis of sign processes and accompanying events from the perspective of the evolution of culture.
Modes of Publication: Irregular intervals, circa 5 to 10 volumes per year. Monographs, collections of papers on topical issues, proceedings of colloquies etc.
General Editor: Walter A. Koch (Bochum).
Advisory Editors: Karl Eimermacher (Bochum), Achim Eschbach (Essen).
Advisory Board: Paul Bouissac (Toronto) Yoshihiko Ikegami (Tokyo), Vjačeslav Vs. Ivanov (Moscow), Rolf Kloepfer (Mannheim), Roland Posner (Berlin), Thomas A. Sebeok (Bloomington), Vladimir N. Toporov (Moscow), Jan Wind (Amsterdam), Irene Portis Winner (Cambridge, Mass.), Thomas G. Winner (Cambridge, Mass.).

Volumes: Available (*) and in preparation (up to 1989):

***Vol. 1:** YAMADA-BOCHYNEK, Yoriko, *Haiku East and West: A Semiogenetic Approach*. xiv + 591 pp., illus., pb DM 94.80, ISBN 3-88339-404-1 (5/85).
***Vol. 2:** ESCHBACH, Achim, KOCH, Walter A (eds.), *A Plea for Cultural Semiotics*. Ca. 210 pp., pb DM 44.80, ISBN 3-88339-405-X (9/87)
Vol. 3: KOCH, Walter A., *Cultures: Universals and Specifics*. Ca. 170 pp., pb ca. DM 34.80, ISBN 3-88339-407-6
Vol. 4: KOCH, Walter A. (ed.), *Simple Forms: An Encyclopaedia of Simple Text-Types in Lore and Literature*. Ca. 700 pp., pb (paperback) ca. DM 129.80, hc (hardcover) ca. DM 144.80, ISBN 3-88339-406-8
Vol. 5: WINNER, Irene P., *Cultural Semiotics: A State of the Art*. Ca. 130 pp., pb ca. DM 24.80, ISBN 3-88339-408-4
***Vol. 6:** KOCH, Walter A., *Evolutionary Cultural Semiotics*. xxiii + 313 pp., illus., pb DM 59.80, ISBN 3-88339-409-2 (10/86).
Vol. 7: KOCH, Walter A. (ed.), *Culture and Semiotics*. Ca. 220 pp., pb ca. DM 44.80, hc ca. DM 59.80, ISBN 3-88339-421-1
Vol. 8: EIMERMACHER, Karl, GRZYBEK, Peter (eds.), *Sprache - Text - Kultur*. Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-410-6
Vol. 9: VOGEL, Susan, *Children's Humour: A Semiogenetic Approach*. Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-411-4
Vol. 10: KOCH, Walter A. (ed.), *Semiotics in the Individual Sciences*. Ca. 1000 pp., hc ca. DM 199.80, ISBN 3-88339-484-X
Vol. 11: KOCH, Walter A. (ed.), *Geneses of Language. Acta Colloquii*. Ca. 400 pp., pb ca. DM 69.80, ISBN 3-88339-485-8
Vol. 12: KOCH, Walter A. (ed.), *The Nature of Culture. Proceedings of the International and Interdisciplinary Symposium, October 7-11, 1986, Ruhr-University Bochum*. 2 vols., each ca. 500 pp., pb each ca. DM 84.80, ISBN 3-88339-553-6
***Vol. 13:** KOCH, Walter A., *Genes vs. Memes*. xvii + 97 pp., illus., pb. DM 29.80, ISBN 3-88339-551-X (12/87).
Vol. 14: KOCH, Walter A., *The Biology of Literature*. Ca. 150 pp., pb ca. DM 34.80, ISBN 3-88339-
Vol. 15: ARLANDI, Gian Franco (ed.)., *Ferruccio Rossi-Landi Probatio*. Ca. 150 pp., pb. ca. DM 34.80, ISBN 3-88339-
Vol. 16: KOCH, Walter A., *The Dawn of Language: Design Schemes in the Evolution of Communication Systems*. Ca. 150 pp., pb. ca. DM 34.80, ISBN 3-88339-
Vol. 17: KOCH, Walter A., *Stereotypy, Ritual, Myth: Towards Cultural Stratification*. Ca. 150 pp., pb. ca. DM 34.80, ISBN 3-88339-609-5
***Vol. 18:** KOCH, Walter A., *Hodos and Kosmos: Ways Towards a Holistic Concept of Nature and Culture*. Ca. 100 pp., pb. ca. DM 29.80, ISBN 3-88339-610-9 (12/87)
Vol. 19: KOCH, Walter A. (ed.),*The Whole and its Parts - Das Ganze und seine Teile. Approaches towards a Holistic Worldview*. Ca. 270 pp., pb. ca. DM 49.80, ISBN
Vol. 20: SHEVOROSHKIN, Vitaly (ed.), *Proto-Languages and their Contacts*. Ca. 200 pp., pb ca. DM 44.80, ISBN
Vol. 21: EIMERMACHER, Karl, WITTE, Georg (eds.), *Issues in Slavic Literary Theory*. Ca. 300 pp., pb ca. DM 64.80, ISBN
Vol. 22: KOCH, Walter A. (ed.), *Evolution of Culture - Evolution der Kultur. Paradigms of Future Interdisciplinary Semiotics*. Ca. 220 pp., pb ca. DM 44.80, ISBN
Vol. 23: SHEVOROSHKIN, Vitaly, KOCH, Walter A., (eds.), *The Language of the Ice Age. Attempts at Reconstructing Proto-Proto-Language*. Ca. 150 pp., pb ca. DM 34.80, ISBN

For more recent and more detailed information on the series (e.g. the current pricelist) and for orders for the whole series or individual volumes please contact the publisher: **Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630 Bochum, Fed. Rep. Germany.** Tel. (0234) 701360 or 701383.

September 1988

BBS

Bochumer Beiträge zur Semiotik

Ziele: Interdisziplinäre Beiträge zu praktischen und theoretischen Themen der Semiotik.
Erscheinungsweise: Unregelmäßige Abstände, ca. 5 - 10 Bände pro Jahr: Monographien, Aufsatzsammlungen zu festgesetzten Themen, Kolloquiumsakten usw.
Herausgeber: Walter A. Koch (Bochum)
Herausgeberbeirat: Karl Eimermacher (Bochum), Achim Eschbach (Essen), Udo L. Figge (Bochum), Roland Harweg (Bochum), Elmar Holenstein (Bochum), Werner Hüllen (Essen), Frithjof Rodi (Bochum).

Bände: lieferbar (*) und in Vorbereitung (bis 1987):

***Bd. 1:** HOLENSTEIN, Elmar, *Sprachliche Universalien.* xix + 250 S., paperback (pb) DM 44.80, ISBN 3-88339-419-X
***Bd. 2:** ZHOU, Hengxiang, *Determination und Determinantien: Eine Untersuchung am Beispiel neuhochdeutscher Nominalsyntagmen.* xii + 267 S., pb DM 44.80, ISBN 3-88339-412-2
***Bd. 3:** KOCH, Walter A., *Philosophie der Philologie und Semiotik.* Ca. 270 S., pb ca. DM 44.80, hardcover (hc) ca. DM 59.80, ISBN 3-88339-413-0
Bd. 4: KOCH, Walter A. (ed.), *For a Semiotics of Emotion.* Ca. 180 S., pb ca. DM 29.80, hc ca. DM 44.80, ISBN 3-88339-415-7
***Bd. 5:** ESCHBACH, Achim (ed.), *Perspektiven des Verstehens.* Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-414-9
Bd. 6: CANISIUS, Peter (ed.), *Perspektivität in Sprache und Text.* Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-416-5
Bd. 7: EISMANN, Wolfgang, GRZYBEK, Peter (eds.), *Semiotische Studien zum Rätsel.* Ca. 280 S., pb ca. DM 44.80, ISBN 3-88339-417-3
Bd. 8: KOCH, Walter A. (ed.), *Semiotik in den Einzelwissenschaften.* Ca. 1000 S., hc ca. DM 194.80, ISBN 3-88339-418-1
***Bd. 9:** SENNHOLZ, Klaus, *Grundzüge der Deixis.* xxvi + 314 S., pb DM 59.80, ISBN 3-88339-462-9
***Bd. 10:** KOCH, Walter A., *Evolutionäre Kultursemiotik.* xxii + 321 S., pb DM 59.80, ISBN 3-88339-463-7
***Bd. 11:** CANISIUS, Peter, *Monolog und Dialog.* Ca. 380 S., pb ca. DM 64.80, ISBN 3-88339-464-5
Bd. 12: JOB, Ulrike, *Regulative Verben im Französischen: Ein Beitrag zur semantischen Rekonstruktion des internen Lexikons.* Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-487-4
***Bd. 13:** SCHMIDT, Ulrich, *Impersonalia, Diathesen und die deutsche Satzgliedstellung.* Ca. 370 S., pb ca. DM 64.80, ISBN 3-88339-494-7
Bd. 14: KOCH, Walter A., POSNER, Roland (eds.), *Semiotik und Wissenschaftstheorie.* Ca. 350 S., pb ca. DM 59.80, ISBN 3-88339-554-4
Bd. 15: FIGGE, Udo L. (ed.), *Semiotik: Interdisziplinäre und historische Aspekte.* Ca. 250 S., pb ca. DM 44.80, ISBN 3-88339-555-2

Neuere und detailliertere Informationen zur Reihe (z.B. aktuelle Preisliste) sowie Bestellungen (Reihe oder Einzelbände) beim Verlag:
Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630-Bochum-Querenburg. Tel. (0234) 701360 oder 701383.

BPX

Bochum Publications in Evolutionary Cultural Semiotics

Aim and Scope: Transdisciplinary contributions to the analysis of sign processes and accompanying events from the perspective of the evolution of culture.
Modes of Publication: Irregular intervals, circa 5 to 10 volumes per year. Monographs, collections of papers on topical issues, proceedings of colloquies etc.
General Editor: Walter A. Koch (Bochum).
Advisory Editors: Karl Eimermacher (Bochum), Achim Eschbach (Essen).
Advisory Board: Yoshihiko Ikegami (Tokyo), Vjačeslav Vs. Ivanov (Moscow), Rolf Kloepfer (Mannheim), Roland Posner (Berlin), Thomas A. Sebeok (Bloomington), Vladimir N. Toporov (Moscow), Jan Wind (Amsterdam), Irene P. Winner (Cambridge, Mass.), Thomas G. Winner (Cambridge, Mass.).

Volumes: Available (*) and in preparation (up to 1987):

***Vol. 1:** YAMADA-BOCHYNEK, Yoriko, *Haiku East and West: A Semiogenetic Approach.* xiv + 591 pp., pb DM 94.80, ISBN 3-88339-404-1
Vol. 2: ESCHBACH, Achim, KOCH, Walter A. (eds.), *A Plea for Cultural Semiotics.* Ca. 320 pp., pb ca. DM 59.80, ISBN 3-88339-405-X
Vol. 3: KOCH, Walter A., *Cultures: Universals and Specifics.* Ca. 170 pp., pb ca. DM 34.80, ISBN 3-88339-407-6
Vol. 4: KOCH, Walter A. (ed.), *Simple Forms: An Encyclopaedia of Simple Text-Types in Lore and Literature.* Ca. 700 pp., pb (paperback) ca. DM 129.80, hc (hardcover) ca. DM 144.80, ISBN 3-88339-406-8
Vol. 5: WINNER, Irene P., *Cultural Semiotics: A State of the Art.* Ca. 130 pp., pb ca. DM 24.80, ISBN 3-88339-408-4
***Vol. 6:** KOCH, Walter A., *Evolutionary Cultural Semiotics.* Ca. 370 pp., pb ca. DM 69.80, hc ca. DM 84.80, ISBN 3-88339-409-2
Vol. 7: KOCH, Walter A. (ed.), *Culture and Semiotics.* Ca. 220 pp., pb ca. DM 44.80, hc ca. DM 59.80, ISBN 3-88339-421-1
Vol. 8: EIMERMACHER, Karl, GRZYBEK, Peter (eds.), *Cultural Semiotics in the Soviet Union.* Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-410-6
Vol. 9: VOGEL, Susan, *Children's Humour: A Semiogenetic Approach.* Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-411-4
Vol. 10: KOCH, Walter A. (ed.), *Semiotics in the Individual Sciences.* Ca. 1000 pp., hc ca. DM 199.80, ISBN 3-88339-484-X
Vol. 11: KOCH, Walter A. (ed.), *Geneses of Language. Acta Colloquii.* Ca.400 pp., pb ca. DM 69.80, ISBN 3-88339-485-8
Vol. 12: KOCH, Walter A. (ed.), *The Nature of Culture. Proceedings of the International and Interdisciplinary Symposium, October 7-11, 1986, Ruhr-University Bochum.* 2 vols., each ca. 500 pp., pb each ca. DM 84.80, ISBN 3-88339-553-6
***Vol. 13:** KOCH, Walter A., *Genes vs. Memes.* Ca. 100 pp., pb ca. DM 24.80, ISBN 3-88339-551-X

For more recent and more detailed information on the series (e.g. the current price-list) and for orders for the whole series or individual volumes please contact the publisher: **Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630-Bochum, Fed. Rep. Germany.** Tel. (0234) 701360 or 701383.