QUANTITATIVE LINGUISTICS

Vol. 42

Glottometrika 11

edited
by
Ludek Hřebíček

Universitätsverlag Dr. N. Brockmeyer
Bochum 1990

QUANTITATIVE LINGUISTICS

CIP-Titelaufnahme der Deutschen Bibliothek

Glottometrika 11 / by edited Ludek Hřebíček – Bochum : Universitätsverl. Brockmeyer.
ISBN 0932-7991

11 (1990)
(Quantitative linguistics ; Vol. 42)
ISBN 3-88339-777-6

NE: GT

ISBN 3-88339-777-6

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

Contents

Köhler, R., Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation 1

Hammerl, R., Zum Aufbau eines dynamischen Lexikmodells – dynamische Mikro- und Makroprozesse der Lexik 19

Köhler, R., Zur Charakteristik dynamischer Modelle 41

Hřebíček, L., Menzerath-Altmann's law on the semantic level 47

Schweiger, F., On Dixon's model of lexical diffusion in Australia 57

Kristophson, J., Gliederung einer Sprachfamilie (hier der Romania) mit Hilfe eines numerischen Kalküls 68

Rothe, U., Semantische Motivation der Genuszuweisung 95

Best, K.-H., Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen 107

Rothe, U., Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina 111

Kimura, M., Ogino, T., Kinukawa, H., Kurabayashi, Y., A study on the possibility of discriminating between homonyms by a semantic correlation approach considering their preceding and succeeding words 122

Hammerl, R., Untersuchungen zur Verteilung der Wortarten im Text 142

Saukkonen, P., Interpreting textual dimensions through factor analysis: Grammatical structures as indicators of textual dimensions 157

Tanaka, A., Stochastic models of sentence structure Japanese literature 172

Rothe, U., Zörnig, P., The entropy of phoneme frequencies. German and French 198

Hrebicek, L. (ed.), Glottometrika 11, 1989

Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation

Reinhard Köhler, Bochum

Einführung

Die Darstellung des Modells der sprachlichen Selbstregulation in Köhler (1986) beschränkte sich zunächst auf das Subsystem der Lexik; die Einbeziehung weiterer linguistischer Untersuchungsebenen wurde jedoch in Aussicht gestellt. Eine Schnittstelle zur ***Phonologie*** (über die Systemgröße ***Phonemanzahl***, die Systembedürfnisse nach Minimierung des Kodierungs- bzw. Dekodierungsaufwands auf der artikulatorisch/auditiven Ebene und die Auswirkung der ***Phonemanzahl*** auf die durchschnittliche ***Wortlänge***) wurden explizit angegeben und weitere, wie die zum *Text*, über das Anwendungsbedürfnis, die Bedürfnisse nach Kontextökonomie und -spezifizität sowie die Systemgrößen ***Frequenz*** und ***Polytextie*** sind bereits implizit vorhanden. Da die Eigenschaften der Sprachelemente – unabhängig von ihrer Zuordnung zu verschiedenen Analyseebenen durch die Linguisten – und ihre Wechselwirkungen untereinander stets gleichzeitig präsent und wirksam sind, ist die Isolierung solcher Aspekte, etwa in phonologische, morphologische, syntaktische, lexikalische, semantische oder pragmatische, gerade in einem dynamischen Modell mit explanativer Zielrichtung, nur sehr bedingt möglich. Zwar muß man sich aus praktischen Gründen stets auf einen kleinen Teilbereich des Untersuchungsobjekts – insbesondere bei der empirischen Arbeit – konzentrieren, die Zusammenhänge mit den anderen, gerade nicht betrachteten Bereichen dürfen jedoch nicht aus dem Auge verloren werden

Das vorliegende Papier hat die Aufgabe, an einigen Beispielen zu zeigen, wie im systemtheoretischen Modell der sprachlichen Selbstregulation solche verschiedenen Bereiche als – trotz gewisser Selbständigkeit, die die Subsysteme eines Systems immer anstreben – miteinander untrennbar verbunden darstellbar sind. Es wird gezeigt, wie bei der Modellierung funktional oder entwicklungsgeschichtlich eigenständiger Subsysteme der Sprache hierarchische Modellebenen eingeführt werden können.

Zur Illustration der integrativen Wirkung dieses Modellansatzes wird außerdem auf eine sich auf dem Modell ergebende Verbindung zwischen

dem diskutierten Model und dem Menzerathschen Gesetz (Altmann 1980; Altmann, Schwibbe, Kaumanns, Köhler, Wilde 1989) hingewiesen, das empirisch bereits hinreichend bestätigt wurde.

Schließlich wird anhand der zur Modellerweiterung aufgestellten Hypothesen und der daraus resultierenden Modellstruktur auf die dem Ansatz zugrundeliegende funktionalanalytische Erklärungslogik eingegangen.

Hierarchiebildung im Modell

Aus Köhler (1986) betrachten wir nur den Systemteil, der zur Anknüpfung der uns hier interessierenden neuen Größen nötig ist (vgl. Abb. 1). In diesem Teil sind als sprachexterne "Systembedürfnisse" folgende vorausgesetzt:

1. Das Kodierungsbedürfnis (***Kod***), welches für die Notwendigkeit steht, zur sprachlichen Erfassung von Bedeutungen lexikalische Einheiten zur Verfügung zu stellen. Dieses Bedürfnis wirkt sich über den Operator V, der im bisherigen Modell als Indentitätsoperator, also numerisch gleich 1, angesetzt war, auf die ***Lexikongröße*** aus: Eine Sprache benötigte, gäbe es nur die Wirkung dieses Bedürfnisses, genausoviele neue Wörter, wie neue Bedeutungen zu kodieren sind.

2. Das Bedürfnis nach Inventarminimierung (***minI***), das dem Kod-Bedürfnis entgegengesetzt wirkt. Diesem Bedürfnis entspricht die Sprache durch Belastung der lexikalischen Einheiten mit mehr als einer Bedeutung; letztere Eigenschaft, die ***Polylexie***, beeinflußt die ***Lexikongröße*** im Maße -L (wobei die Größe L die durchschnittliche Anzahl von Bedeutungen einer lexikalischen Einheit angibt) und realisiert dadurch die Konkurrenz von ***MinI*** zu ***Kod*** in bezog auf die ***Lexikongröße***.

3. Das Bedürfnis nach Übertragungssicherung (***Red***), welches zur Verminderung von Verwechslungsgefahr durch Redundanz (Nichtausnutzung aller möglichen Kombinationen von Phonemen zu Morphemen/Wörtern) eine zu große Ähnlichkeit zu vieler lexikalischer Einheiten verhindert - und sich dadurch vergrößernd auf die ***Wortlänge*** auswirkt.

4. Die Bedürfnisse nach Minimierung des Kodierungsaufwands (***minK***) beim Sprecher und nach Minimierung des Dekodierungsaufwands (***minD***) beim Hörer, die in direkter Konkurrenz stehen und - auf die Ebene der Phonologie soll hier nicht eingegangen werden (vgl. dazu Köhler, Altmann 1983 und Job, Altmann 1985) - sich über einen dynamischen Kompromiß (ein Fließgleichgewicht) auf die ***Polylexie*** auswirken.

5. Das Spezifikationsbedürfnis (***Spz***) zur Repräsentation der Notwendigkeit, gegebene Mehrdeutigkeiten und Vagheiten einer lexikalischen Einheit zu verringern.

Bei der Einführung des Modells wurde kurz auf die verschiedenen Möglichkeiten eingegangen, die einer Sprache zur Berücksichtigung des Spezifikationsbedürfnisses zur Verfügung stehen. Explizit in das Modell aufgenommen worden sind zunächst nur die Konsequenzen der morphologischen Methode: in dem Maße T, in dem eine Sprache morphologische Mittel (Komposition, Derivation, Flexion) dazu verwendet, die Menge der Bedeutungsalternativen zu verringern, steht die Länge der lexikalischen Einheit mit der Anzahl der möglichen Bedeutungen, der Polylexie, in Verbindung. Die Hypothese

> *Die Veränderungsrate der Polylexie ist umgekehrt proportional zur Länge , wobei der Proportionalitätsfaktor durch die Synthetizität der Sprache bestimmt ist*

wurde in Form der Differentialgleichung

$$\frac{y'}{y} = \frac{-T}{x}$$

formuliert. Die Lösung dieser Differentialgleichung ist

$$y = Cx^{-T}$$

wobei C zunächst eine beliebige Konstante sein kann. Durch Einbeziehung der Bedürfnisse minK und minD, die ebenfalls auf die Polylexie wirken, konnte C als Einfluß des Fließgleichgewichts zwischen diesen Bedürfnissen bestimmt werden, woraus sich schließlich für die Beschreibung der Abhängigkeit der Polylexie lexikalischer Einheiten einer Sprache von anderen Systemgrößen die Gleichung

$$PL = minK^{Q2}\ minD^{-Q1}\ L^{-T},$$

bzw. die logarithmierte Form (*)

$$PL^* = Q2\ minK^* - Q1\ minD^* - T\ L^*$$

ergibt die linearisierte Darstellung in Abbildung 1.[1)]

Es liegt nun nahe, zu untersuchen, welche Konsequenzen die Verwendung anderer als morphologischer Mittel zur Spezifikation nach sich ziehen, und wie die verschiedenen sprachlichen Mittel (dies sind im funktionalanalytischen Sinn funktionale Äquivalente) zueinander in Beziehung stehen.

Grundsätzlich stehen zur Reduktion von Mehrdeutigkeit die selben sprachlichen Mittel zur Verfügung wie zum Kodieren von Bedeutungen überhaupt: lexikalische, morphologische, syntaktische und prosodische. Das Ausmaß, in welchem eines der funktionalen Äquivalente zur Spezifikation herangezogen wird, beeinflußt die Wirkung des Spz-Bedürfnisses auf eine oder mehrere Systemgrößen, die mit dem jeweiligen Äquivalent gekoppelt sind.

Die lexikalische Lösung besteht darin, daß ein besonderer Ausdruck für eine spezifische, weniger vage bzw. weniger mehrdeutige Bedeutungskonstellation in das Lexikon aufgenommen wird. Dann aber entspricht die Anzahl von zu bildenden lexikalischen Einheiten nicht mehr im Verhältnis 1:1 der Anzahl der hinzukommenden Bedeutungen, wie es durch den Proportionalitätsoperator V = 1 zunächst festgelegt wurde. Der Wert von V muß vielmehr als Funktion des Spz-Bedürfnisses und einer Größe Φ_L, die den Anteil der lexikalischen Mittel an der Gesamtheit aller verwendeten Spezifiktionsmittel angibt, modelliert werden. Gleichzeitig hängt V von den anderen Systembedürfnissen ab, die ebenfalls durch Bildung von bedeutungsverwandten, synonymen Ausdrücken bedient werden, ab; diese sollen daher an dieser Stelle auch explizit in das Modell aufgenommen und als Bedürfnis nach Mitteln für Originalität bzw. Variabilität im Ausdruck (Var) postuliert werden. In Abbildung 2 ist eine Hypothese, nach der V, Φ_L und Spz linear zusammenhängen, graphisch dargestellt. In gleicher Weise wurde der Proportionalitätsfaktor T als Funktion des Spz-Bedürfnisses und der Größe Φ_M (für den Anteil der morphologischen Lösungen) präzisiert.

1) Zur Notation und zu den verwendeten mathematischen Hilfsmitteln s. Köhler (1986:43-49).

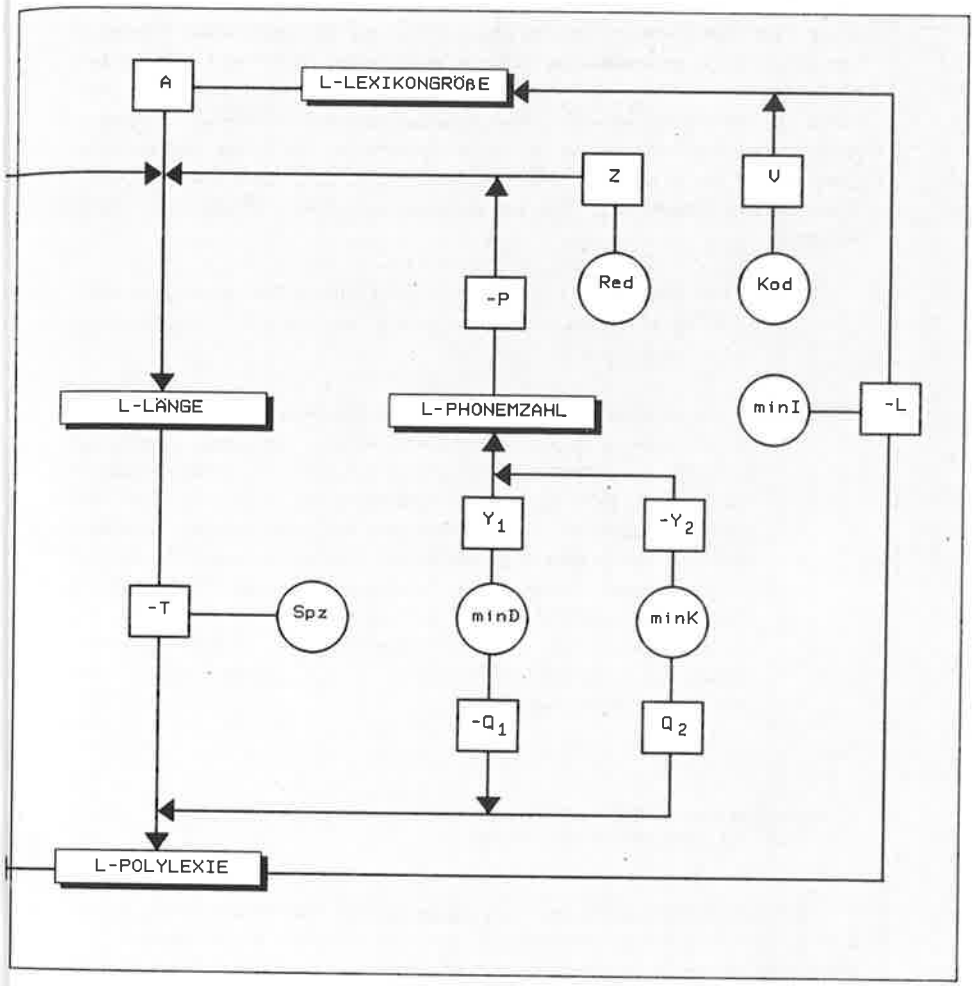


Abbildung 1. Die Abhängigkeit der Polylexie von weiteren Systemgrößen (linearisiert)

In dieser Abbildung haben wir eine neue Notation eingeführt; die funktionale Abhängigkeit eines Proportionalitätsfaktors von einer anderen Größe wird durch einen Pfeil dargestellt, der bis in das Operatorkästchen hineinragt. Dies soll andeuten daß die numerische Größe, nicht der stukturelle Zusammenhang des Operators von der bestimmenden Größe abhängt. Es wäre selbstverständlich möglich, anstelle einer solchen Notation eine Strukturveränderung im Modell vorzunehmen, die durch algebraische Eliminierung des Operators und Berechnung der Gesamtfunktion eine einfache Lösung ermöglichte. Dies ist aber aus folgenden Gründen nicht angebracht:

1. Die gewählte Darstellung ist zwar strukturell komplexer, ihre einfachere Alternative hätte jedoch den Nachteil konzeptueller Undurchsichtigkeit.

2. Die so entstehende hierarchische Struktur gibt nicht nur die funktionalanalytische Hypothese wieder, sondern entspricht auch der möglichen Annahme, daß die jetzt im Modell dazugekommene Schicht auch sprachgeschichtlich jünger ist und die ältere voraussetzt: Die Bedürfnisse nach Bedeutungsspezifikation und nach Ausdrucksvariabilität sowie die zugehörigen Systemfunktionen setzen das fundamentale Kodierungsbedürfnis und die Systemfunktion der Lexikalisierung voraus. Durch die Hierarchisierung ist der vorläufig nur als Identitätsoperator behandelte Proportionalitätsfaktor V zur eingenständigen Systemgröße aufgestiegen.

Kooperierende Systemfunktionen und linguistische Ebenen

Mit den Überlegungen des vorhergegangenen Abschnitts ist auch eine zuvor nicht behandelte Eigenschaft der lexikalischen Einheiten zur Systemgröße geworden: die **Synonymie**. Mit dieser Größe soll erfaßt werden, wieviele lexikalische Einheiten es in einer Sprache zu einer gegebenen lexikalischen Einheit gibt, von deren Bedeutungen wenigstens eine jeweils

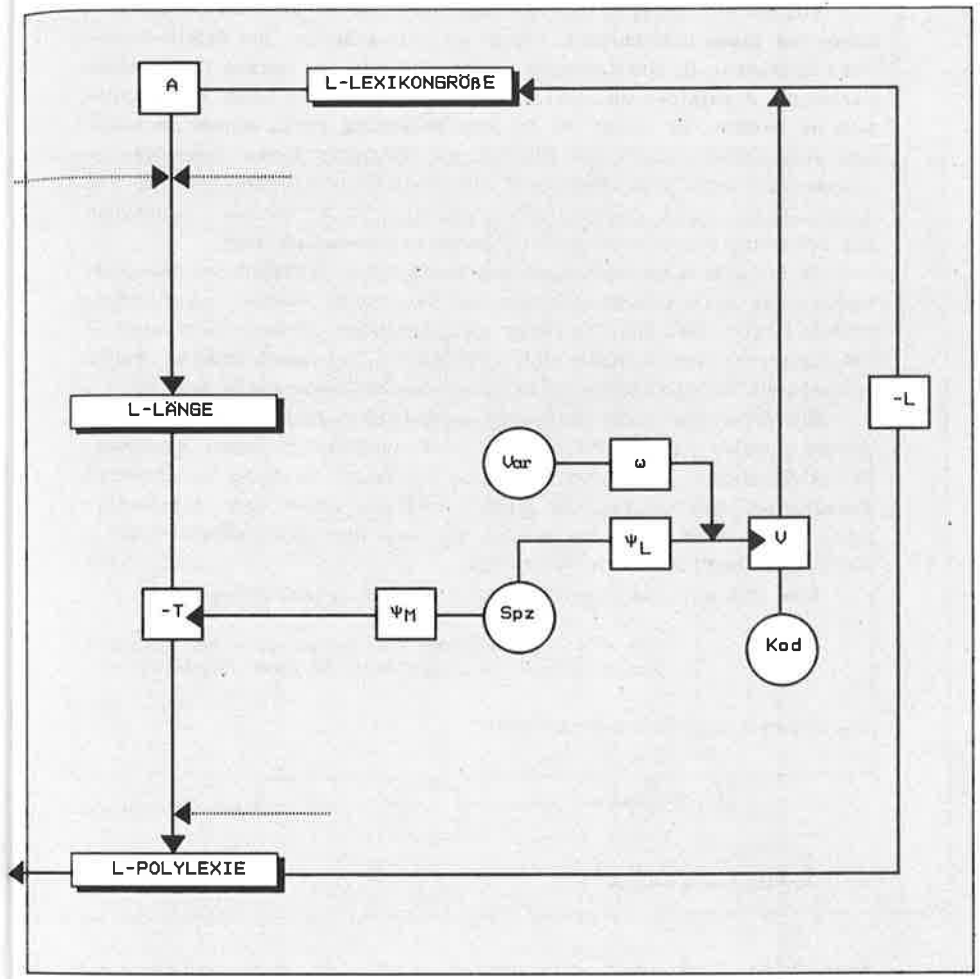


Abbildung 2. Proportionalitätsfaktoren als Funktionen von Systembedürfnissen

gleich (bzw. ausreichend verwandt) mit wenigstens einer ihrer eigenen Bedeutungen ist.[2)]

Wie bei den anderen Systemgrößen auch, ist es sinnvoll, einen globalen und einen individuellen Aspekt zu unterscheiden. Die durchschnittliche Synonymie in einer Sprache ergibt sich aus dem gerade diskutierten Verhältnis V zwischen zu kodierenden Bedeutungen und ihrer Repräsentation im Lexikon. Im Falle V=1 ist jede Bedeutung genau einmal im Lexikon ausgedrückt (durch die Wirkung der Polylexie kommt allerdings im Allgemeinen nicht jeder Bedeutung ein eindeutiger Asudurck zu. Bei V>1 werden Bedeutungen auch mehrfach lexikalisiert, es entsteht Synonymie zur Bedienung des Variations- und Spezifikationsbedürfnisses.

Es ist aber auch hierzu ein konkurrierendes Bedürfnis zu beachten, nämlich das nach Vereinheitlichung der Benennung, welches Invarianzbedürfnis (Ivz) heißen soll. Es kommt ganz besonders in fach- und wissenschaftssprachlichem Kontext zur Geltung. Wie in vielen anderen Fällen entsteht ein Fließgleichgewicht zwischen den konkurrierenden Kräften.

Die Synonymie einer gegebenen lexikalischen Einheit setzt sich aus diesem globalen Aspekt und der für jeden Ausdruck gesondert bestehenden Abhängigkeit von seiner Polylexie zusammen: Je mehr verschiedene Bedeutungen eine lexikalische Einheit trägt (je größer ihre Mehrdeutigkeit), desto mehr Synonyme werden für den Spezifikationsbedarfsfall - bei konstantem[3)] Anteil Φ_L - benötigt.

Dies läßt sich wie folgt zu einer Hypothese präzisieren:

> *Die relative Änderung der Synonymie einer lexikalischen Einheit ist proportional zu ihrer Polylexie.*

Als Differentialgleichung geschrieben:

$$\frac{Sy'}{Sy} = \frac{M}{PL};$$

2) Wir verzichten auf den Versuch, eine Operationalisierung der Relation "bedeutungsgleich" oder "bedeutungsverwandt" zu geben. Es wird lediglich vorausgesetzt, daß diese für jeden konkreten Untersuchungszweck möglich ist.

3) Bei dynamischer Betrachtung ist dies natürlich nicht der Fall; diese ceteris-paribus-Redeweise dient nur dem leichteren Verständnis der Zusammenhänge.

Hier stehen Sy für die Synonymie, PL für die Polylexie, Sy' für die erste Ableitung von Sy und M für den Proportionalitätsfaktor. Die Lösung der Differentialgleichung ist

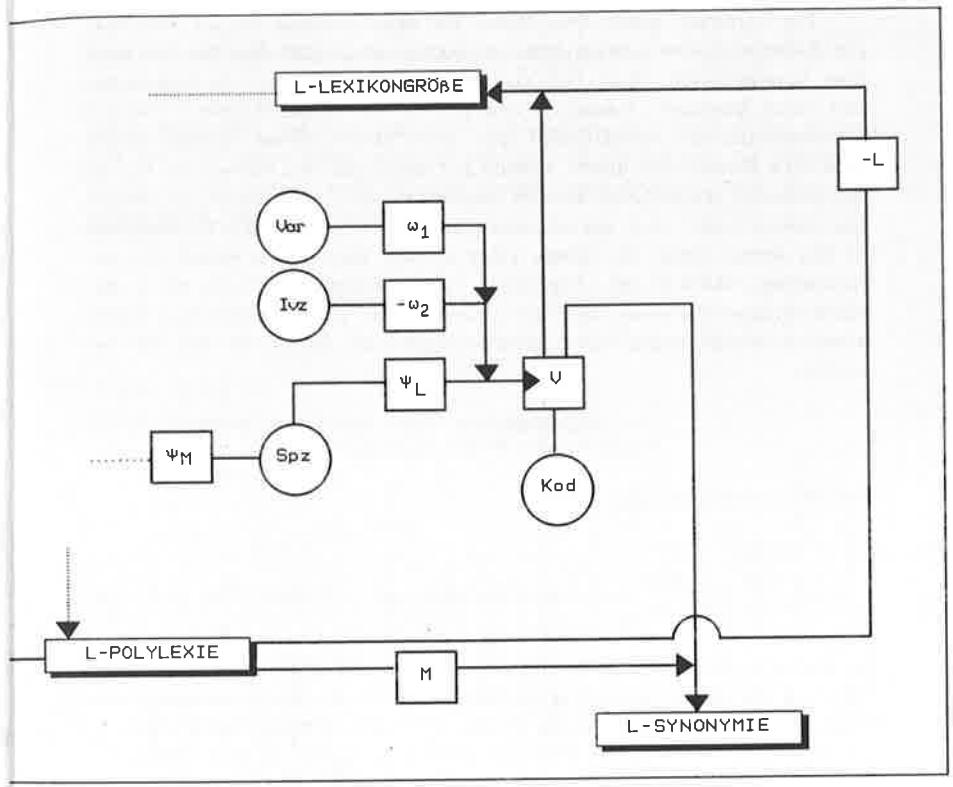


Abbildung 3. Die Synonymie bestimmt sich aus den globalen Wirkungen der Bedürfnisse nach Ausdrucksvariabilität, Ausdrucksvarianz und Spezifikation sowie aus der Polylexie der lexikalischen Einheit

$$Sy = C\ PL^{M}$$

Die Interpretation der Integrationskonstanten C ist leicht möglich; es handelt sich um den Globalanteil KodV, so daß

$$Sy = Kod^{V}\ PL^{M}.$$

Ein weiteres sprachliches Mittel zur Spezifizierung der zu kodierenden Bedeutung einer verwendeten lexikalischen Einheit ist die syntaktische Determination. Ohne daß die zahlreichen Varianten wie Determination durch Quantor-, Possesiv- und andere Pronomina, Phrasen- und Relativsatzattribute, Appositionen etc. hier unterschieden werden sollen, liegt eine Konsequenz dieser Mehode auf der Hand: Die Phrase, in der die determinierte lexikalische Einheit verwendet wird, wird durch sie länger. Aus diesem Grund wird die Aufnahme der neuen Systemgröße **Phrasenlänge** in das Modell nötig. Die Länge einer Phrase hänge also, außer von semantischen, stilistischen, kognitiven und anderen hier noch nicht berücksichtigten Faktoren, von der Notwendigkeit der syntaktischen Determination infolge lexikalischen Mehrdeutigkeit ab. Formuliert man die Hypothese

> *Die Längenzunahme einer Phrase ist proportional zur Polylexie der Wörter in dieser Phrase*

als Differentialgleichung

$$\frac{PhL'}{PhL} = \frac{\Phi}{PL},$$

wo PhL für die Länge einer Phrase, PhL' für die erste Ableitung dieser Größe, Φ für den Proportionalitätsfaktor und PL für die durchschnittliche Polylexie der Wörter der Phrase stehen, so ergibt sich mit der Lösung

$$PhL = D\ PL^{\Phi}$$

eine hyperbolische Funktion zur Bestimmung der Phrasenlänge aus der Polylexie. Der Parameter gibt die Stärke des numerischen Zusammenhangs an und kann daher leicht linguistisch interpretiert werden: es handelt

sich um nichts anderes als den Grad, in dem die betrachtete Sprache von syntaktischen Mitteln zur Bedeutungsspezifikation Gebrauch macht. Wir schreiben daher

$$\Phi = \mu_S Spz \quad .$$

Anschluß des Menzerathschen Gesetzes

Unter Benutzung des uns bekannten Zusammenhangs zwischen der Polylexie und der Länge einer lexikalischen Einheit können wir die erhaltenen Funktionsgleichungen noch in eine andere Form bringen. Nach Einsetzen der rechten Seite der Gleichung

$$PL = C\ L^{-T}$$

für PL haben wir

$$PhL = D\ (CL^{-T})^{\mu S}$$

$$= D\ C^{\mu S} L^{-T\mu S}$$

Mit $a = DC^{\mu_S}$ und $b = T\mu_S$ können wir

$$PhL = a\ L^{-b}$$

schreiben, was in Inhalt und Form mit dem Menzerathschen Gesetz auf Phrasenebene (s.a. Köhler 1982; 1984) identisch ist. Abbildung 4 verdeutlicht, wieder linearisiert, den Zusammenhang graphisch.

Die übrigen Ebenen, für die das Menzerathsche Gesetz gilt, müssen natürlich separat durch Identifizierung mit Zusammenhängen zwischen Systemgrößen des Modells integriert werden. Dies soll an dieser Stelle jedoch nicht durchgeführt werden.

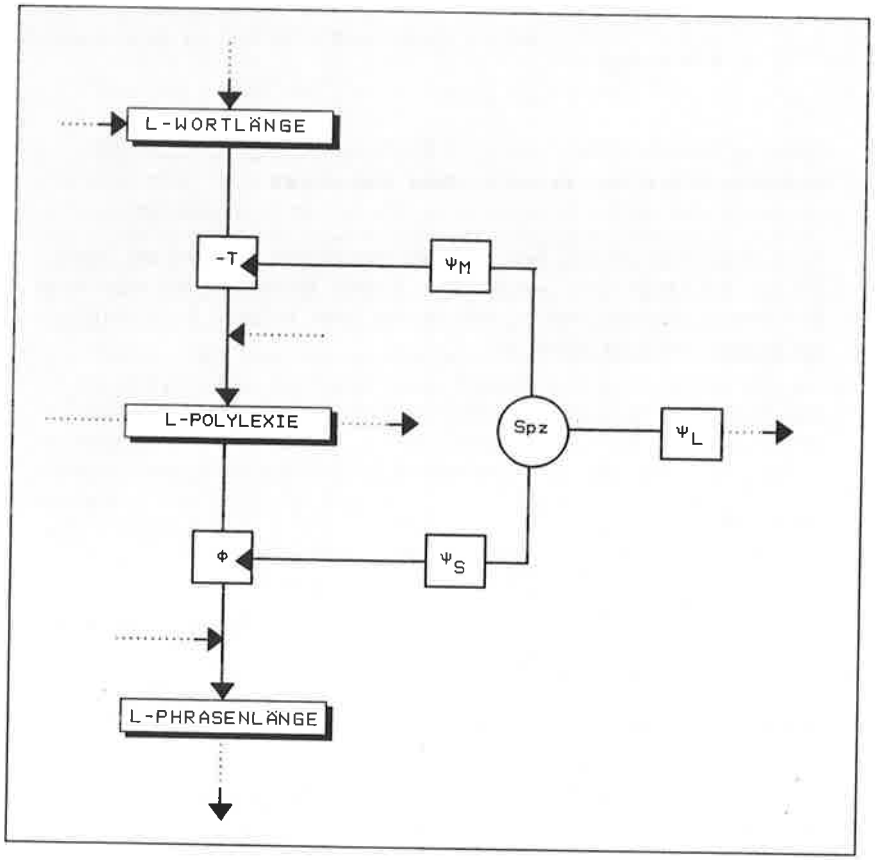


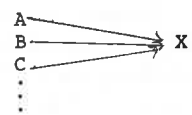
Abbildung 4. Das Menzerathsche Gesetz auf Phrasenebene als Zusammenhang zwischen Wortlänge, Polylexie und Phrasenlänge

Zur funktionalanalytischen Erklärungslogik

Eine Modellierung, wie sie oben durchgeführt wurde, geht über eine begrifflich-deskriptive Erfassung sprachlicher Zusammenhänge hinaus; es ist vielmehr charakteristisch für den verwendeten Modellansatz, daß mit ihm der Anspruch auf erklärende Kraft verbunden wird. Dabei liegt das wissenschaftliche Konzept der ontischen Erklärung zugrunde, nach dem ein Ereignis zu erklären heißt, es in ein verständliches Muster einzubetten (vgl. Salmon 1982; 1984:18). Das deduktiv-nomologische Erklärungsschema von Hempel-Oppenheim (vgl. Hempel 1965), das sogar in deterministischen Fällen seine Schwächen hat (vgl. Salmon, McLaughlin 1982; Cartwright 1983) und heftige Kontroversen und Lösungsversuche in probabilistischen Fällen hervorbrachte (vgl. z.B. Humphreys 1981; Fetzer 1981; Rogers 1981; Fetzer, Nute 1979; van Fraassen 1980; Railton 1978, 1981; Sayre 1977; Grünbaum, Salmon 1988, um nur einige wenige zu nennen), kann nur cum grano salis benutzt werden.

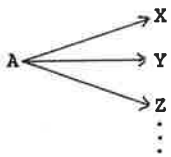
In der Sprache können wir nämlich mit folgenden Situationen konfrontiert werden:

(a) Erscheinung A ist eine hinreichende, aber nicht notwendige Ursache der Erscheinung X, wenn X auch von anderen Erscheinungen verursacht werden kann:



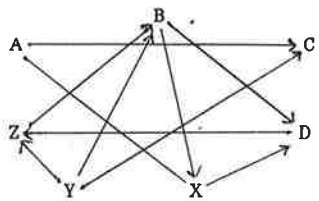
In diesem Fall liefert A eine Erklärung, aber keine notwendige Ursache. X kann gleichzeitig von mehreren Mechanismen erzeugt werden. Im probabilistischen Fall - der für die Sprache immer zutrifft - gibt es noch weitere Probleme.

(b) A ist eine notwendige Ursache von X, erzeugt jedoch auch Y, Z,...:



In diesem Fall haben wir zwar die notwendige Ursache, aber keine Erklärung, denn es bliebe ja zu erklären, ***warum*** in gegebenem Fall gerade X auftrat, wenn A auch andere Wirkungen haben kann (vgl. Salmon 1982).

(c) Die typische Situation in der Sprache ist jedoch noch komplizierter, nämlich etwa wie folgt



so daß uns im Grunde nur eine einzige Möglichkeit bleibt, nämlich die Funktionalanalyse, bei der man die kausale und probabilistische Vernetzung der Sprachentitäten berücksichtigt. In Unterschied zu der deduktiv-nomologischen Erklärung liegt bei der funktionalen Erklärung keine Ursache mit zwingend eintretender Wirkung vor; die Existenz und die Eigenschaften eines Explanandums werden statt dessen durch Angabe der Funktion des Explanandums im System erklärt. Die logische Vollständigkeit der Erklärung wird durch die Annahme von Selbstregulationsmechanismen des Systems erreicht, die den Systemzustand ständig an einen optimalen Zustand anzunähern suchen. Dieser Sollzustand entspricht dem Zustand, in dem alle Systembedürfnisse bedient sind. Eine Erscheinung E_f, deren Funktion es ist, ein Systembedürfnis zu bedienen, kann dann unter Berufung auf den Selbstregulationsmechanismus erklärt werden.

Ein allgemeines Problem von funktionalen Erklärungsversuchen ist das der möglichen funktionalen Äquivalente. Eine Erklärung, wie sie oben skizziert wurde, ist noch immer nicht vollständig, wenn die Funktion, mit deren Notwendigkeit eine E_f erklärt werden sollte, auch auf andere

Weise, etwa von anderen E_i oder anderen Ausprägungen von E_f, ausgeübt werden könnte. Eine logisch einwandfreie funktionale Erklärung bedürfte also der Angabe der Beziehungen zwischen allen denkbaren funktionalen Äquivalenten zu einer Funktion. In (1981) fordert Altmann die Ableitung dieser Beziehung aus Axiomen, theoretischen Annahmen oder Gesetzen. In (Köhler 1986:25-33) findet man eine etwas ausführlichere Diskussion dieses Problems, als es hier möglich ist. Dort wird auch ein Schema für eine logisch vollständige Funktionalerklärung für selbstregulierende Systeme aufgestellt:

a) Das System S ist selbstregulierend: Für jedes Bedürfnis besteht ein Mechanismus, der den Systemzustand so verändert, daß es bedient wird.

b) An das System S sind die Bedürfnisse B_1, B_2,...,B_k gestellt.

c) Das Bedürfnis B_j kann durch die funktionalen Äquivalente E_1,..., E_f,...,E_m bedient werden.

d) Zwischen den funktionalen Äquivalenten besteht die Relation $R(E_1,...,E_f,...,E_m)$.

e) Aufgrund der Systemstruktur besteht zwischen den Elementen s_1,...,s_z des Systems S die Relation $Q(s_1,...,s_z)$.

E_f ist Element des Systems S mit der Ausprägung A_f

Am Beispiel der in den vorherigen Abschnitten entwickelten Hypothesen soll nun gezeigt werden, wie den angeführten Forderungen nachgekommen werden kann. Oben wurde gesagt, daß Sprachen zur Bedienung des Spezifikationsbedürfnisses lexikalische, morphologische, systaktische und prosodische Hilfsmittel entwickeln können. Diese Aussage entspricht der Angabe der möglichen funktionalen Äquivalente für die Funktion der Bedeutungsspezifikation (Punkt b) und der Annahme, daß diese Aufzählung vollständig ist (Punkt c). Für die ersten drei Hilfsmittel wurden oben strukturelle Annahmen gemacht; der Zusammenhang zwischen dem Grad ihres jeweiligen Einsatzes in einer Sprache und je einem Funktionsparameter wurde explizit modelliert. So ist der Parameter T in der Funktionsgleichung

$$PL = \min K^{Q2} \; \min D^{-Q1} L^{-T}$$

für die Abhängigkeit der Polylexie von der Wortlänge einer lexikalischen Einheit eine Funktion des Ausmaßes μM, in dem eine Sprache morphologische Mittel zur Spezifikation heranzieht, usw. Die Systemgröße Länge,

Polylexie, Polytetie, Frequenz, Lexikongröße, Phonemanzahl, Synonymie und Phrasenlänge (beim augenblicklichen Stand der Modellierung) hängen ebenfalls funktional zusammen. Damit ist die Relation Q im Punkt e des Erklärungsschemas durch Auflisten aller Funktionsgleichungen des Modells gegeben.

Um die Ausprägungen A_i der funktionalen Äquivalente E_i logisch korrekt folgern zu dürfen, müssen wir noch die Relation R zwischen den Äquivalenten bestimmen. Bisher sind uns dazu erst wenige Aussagen möglich:

1. Lexikalische, morphologische, syntaktische und prosodische Mittel zur Spezifikation schließen sich gegenseitig nicht aus.

2. Die Summe aller Spezifikationsvorgänge ist die Summe der Verwendungen der vier Mittel.

Daraus ergibt sich R(L,M,S,P) als

$$\mu_L + \mu_M + \mu_S + \mu_P = 1,$$

und wir können die Ausprägung eines jeden funktionalen Äquivalents, z.B. der Spezifikation durch prosodische Mittel, dem Erklärungsschema folgend, als

P ist Element des Systems S mit der Auspägung

$$\mu_P = 1 - \mu_S - \mu_M - \mu_L$$

ableiten. Zur numerischen Bestimmung der vier Unbekannten wären allerdings weitere Gleichungen erforderlich, die uns leider noch nicht zur Verfügung stehen. Linguistisch ausgedrückt stellt sich die Frage danach, wie die Sprachen die Verwendungsgrade der möglichen Hilfmittel steuern – eine Frage, die heute noch nicht beantwortet werden kann. Empirsche Vergleiche typologisch unterschiedlicher Sprachen können hierzu zwar Ideen beitragen, aber zur Erklärung können nur theoretisch abgeleitete Hypothesen und Gesetze führen.

Literatur

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

Altman, G. (1981), Zur Funktionalanalyse in der Linguistik. In: Esser, J., Hübler, A. (Eds.), *Forms and Functions.* Tübingen, Narr: 25–32.

Altmann, G. (1989), The levels of linguistic investigation". *Theoretical Linguistics 14, 227–240.*

Cartwright, N. (1983), How the laws of physics lie. Oxford, Clarendon

Fetzer, J.H. (1981), Probability and explanation. *Synthese 48, 371–408.*

Fetzer, J.H., Nute, D.E. (1979), Syntax, semantics, and ontology: A probabilistic causal calculus. *Synthese 40, 453–495.*

Grünbaum, A., Salmon, W.C. (eds.) (1988), The limitations of deductivism. Berkeley, University of California Press.

Hempel, C.G. (1965), The logic of functional analysis. In: Hempel, C.G., *Aspects of scientific explanation.* New York, The Free Press: 297–330.

Humphreys, P. (1981), Aleatory explanations. *Synthese 48, 225–232.*

Job, U., Altmann, G. (1985), Ein Modell für die anstrengungsbedingte Lautveränderung. *Folia Linguistica Historica 6, 401–407.*

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

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

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

Köhler, R. (1989), Systems theoretical linguistics. *Theoretical Linguistics 14, 241–258.*

Köhler, R., Altmann, G. (1983), Systemtheorie und Semiotik. *Zeitschrift für Semiotik 5/4, 424–431.*

Railton, P. (1978) A deductive–nomological model of probabilistic explanation. *Philosophy of Science 45, 206–226.*

Railton, P. (1981), Probability, explanation, and information. *Synthese 48, 233–256.*

Rogers, B. (1981), Probabilistic causation, explanation, and detection. *Synthese 48, 201–223.*

Salmon, W.C. (1982), Comets, pollen and dreams: Some reflections on scientific explanation. In: Salmon, McLaughlin 1982: 155–178.

Salmon, W.C., McLaughlin, R. (eds.) (1982), What? Where? When? Why?. Essays on induction, space and time, explanation. Dordrecht, Reidel 1982.

Sayre, K.M. (1977), Statistical models of causal relations. *Philosophy of Science 44, 203-214.*

van Fraassen, B.C. (1980), The scientific image. Oxford, Clarendon.

Hřebíček,L.(ed.), Glottometrika 11, 1989

Zum Aufbau eines dynamischen Lexikmodells — dynamische Mikro- und Makroprozesse der Lexik

R. Hammerl, Bochum

1. Vorwort

In der Arbeit "Zur sprachlichen Synergetik: Struktur und Dynamik der Lexik" (1986) beschreibt Köhler den Aufbau eines Modells für ein lexikalisches Subsystem natürlicher Sprachen, das die gegenseitigen Abhängigkeiten einer begrenzten Zahl quantitativer lexikalischer Eigenschaften betrifft. Köhler stützt sich dabei auf Erkenntnisse der von Haken (1978) im Bereich der Physik ausgearbeiteten Theorie der kooperativen Phänomene, der Synergetik. Sprache wird als ein selbstregulierendes System aufgefaßt, wo bestimmte menschliche Bedürfnisse (z.B. das Bedürfnis nach Minimierung des Produktionsaufwandes in der sprachlichen Kommunikation, das Anwendungsbedürfnis sprachlicher Einheiten usw.) bei Berücksichtigung bestimmter sprachlicher Parameter (wie z.B. der Phonemzahl, der Lexikongröβe) als die "Kräfte" angesehen werden, welche die gegenseitigen Abhängigkeiten zwischen sprachlichen Eigenschaften gestalten. Gestützt auf bestimmte linguistische Vorüberlegungen (die z.B. zur Festlegung der kausalen Abhängigkeitsrichtungen führen) wurde dann die mathematische Beschreibung dieser Abhängigkeiten vorgenommen, wobei sich die entsprechenden Funktionen als Lösungen von Differentialgleichungen erster Ordnung ergaben. Die gefundenen Abhängigkeiten wurden dann einer ersten empirischen Überprüfung unterzogen. Besonders diskutiert Köhler in seiner Arbeit die Abhängigkeit der (mittleren) Länge L lexikalischer Einheiten von deren Frequenz F, da dort eine "Oszillation der Lexik" festgestellt wurde, die mit dem von Köhler abgeleiteten Modell nicht beschrieben werden konnte (es bleibt zunächst dahingestellt, ob die Lösung einer Differentialgleichung 2. Ordnung, wie sie Köhler (1986, 144 ff.) vorschlägt, zu einer zufriedenstellenden Lösung dieses Problems führen kann).

Im Aufsatz "Ein Beitrag zu Köhler's Modell der sprachlichen Selbstregulation" (Hammerl, Maj 1989) haben wir unter Anwendung von Wissen aus der Systemtheorie und Kybernetik versucht, eine für sprachwissenschaftliche Untersuchungen neue Methode zur Untersuchung der Abhängigkeiten zwischen sprachlichen Eigenschaften (dort illustriert an der Relation zwischen L und F) abzuleiten, obwohl uns bewußt war, daß die Anwendung dieser Untersuchungsmethode aufgrund der ungünstigen Datenlage und des Fehlens empirisch und theoretisch validierter diachronischer Sprachgesetze zur Zeit nicht möglich ist.

Im Artikel "Zur Charakteristik dynamischer Systeme" (Köhler, in diesem Band) nimmt Köhler Stellung zu mehreren von Hammerl, Maj (1989) angesprochenen Problemen.

Die Lektüre dieses Artikels von Köhler überzeugte mich davon, daß die von Hammerl, Maj vorgestellte Untersuchungsmethode und deren methodologischer Hintergrund mehr und sicher auch bessere Erklärung bedarf, daß Köhler selbst in einigen Fällen eine solche Präzisierung in der Darstellung seines Lexikmodells vornimmt, die unseren Vorstellungen darüber sehr nahe kommen, und daß es notwendig ist, einige Grundprobleme der Modellierung lexikalischer Abhängigkeiten (des Aufbaus eines lexikalischen Modells) neu zu überdenken und an empirischem Material zu überprüfen (auf der Grundlage umfangreicher Daten zur polnischen Sprache, die im Rahmen des Projektes "Sprachliche Synergetik" vorbereitet wurden; die empirische Überprüfung muß jedoch Gegenstand einer gesonderten Arbeit sein, die bald vorgestellt werden kann). Hier soll zunächst eine erste Stellungnahme zu den von Köhler diskutierten Problemen unseres Aufsatzes (Hammerl, Maj 1989) vorgenommen und darüber hinaus sollen auch einige spezielle Probleme des Aufbaus eines lexikalischen Basismodells unter Berücksichtigung neuer Erkenntnisse diskutiert werden.

2. Grundprobleme des Aufbaus eines Basismodells der Lexik

Köhler wendet bei der Beschreibung seines Lexikmodells Erkenntnisse und Begriffe an, die aus der Physik bekannt sind (z.B. die Begriffe Dauer, Geschwindigkeit und Beschleunigung der Bewegung, Kraft; auch die Begriffe des Basismodells und der Kernelemente des Basismodells erinnern an das aus der klassischen Physik bekannte Basissystem und deren Basiselemente). Es ist sicher von Vorteil, bestimmte Erfahrungen der Physik (und anderer Naturwissenschaften) über die Gewinnung von Wissen von

der Realität auch beim Aufbau und der Überprüfung eines lexikalischen Basismodells heranzuziehen. Aus diesem Grunde wollen wir auch einige grundlegende methodologische Fragen unter diesem Gesichtspunkt diskutieren, um gleichzeitig einige grundlegende Kritikpunkte unseres Aufsatzes (Hammerl, Maj 1989) am Köhlerschen Modell und Kritikpunkte Köhlers (Köhler, in diesem Band) bezüglich unseres Aufsatzes klären zu können.

Die naturwissenschaftliche Forschung beruht auf der Überzeugung, daß die Erscheinungen der uns umgebenden Welt der Wirkung von Gesetzen unterliegen, die erkannt werden und diese Erscheinungen erklären können. Auch in der sprachlichen Realität (in der Sprache) wirken Gesetze, die recht kompliziert und den Sinnenswahrnehmungen nicht direkt zugänglich sind. Das heißt jedoch nicht, daß es nicht möglich oder sogar sinnlos ist, nach solchen Gesetzmäßigkeiten zu suchen. Hinsichtlich dieses Grundsatzes stimmen wir mit Köhler überein; es bestehen lediglich in einigen speziellen Fragen teilweise differierende Auffassungen über die Modellierung der jeweiligen Abhängigkeiten.

Da die zu untersuchende Realität in der Regel recht kompliziert und vielgestaltig ist, beschränkt man sich auf die Analyse kleiner Wirklichkeitsausschnitte und schafft bewußt Idealisierungen der Wirklichkeit, die mit Begriffen der jeweiligen Wissenschaft beschrieben werden sollen. Man versucht dann, im Rahmen der jeweiligen Idealisierung durch Anwendung induktiver, deduktiver oder kombinierter Verfahren die Beziehungen zwischen den untersuchten Erscheinungen zu erkennen, mathematisch zu beschreiben, empirisch zu überprüfen und in ein System anderer (schon validierter) Gesetze einzuordnen. Somit entstehen bestimmte Modelle der Realität, wobei die aus den jeweiligen Modellannahmen abgeleiteten Gesetze nur beschränkte Allgemeingültigkeit haben, d.h. nur in dem Bereich gelten, wo das Modell als ausreichendes Abbild der Wirklichkeit angesehen werden kann.

Zur Erklärung ein und derselben Erscheinung der Wirklichkeit können somit verschiedene Modelle dienen, was oft sogar unumgänglich ist, da nur so ein tieferes Wissen über die jeweilige Erscheinung erzielt werden kann. Die Wahl des entsprechenden Modells ist ausschließlich dem jeweiligen Wissenschaftler überlassen.

Aus diesem Grunde ist es auch nicht uneingeschränkt möglich, den Wert eines Modells aus der Sicht und mit Begriffen eines anderen Modells einzuschätzen. Prüfstein des Wertes solcher Modelle kann nur deren Erklärungsgehalt für die untersuchten Realitätsausschnitte sein; dazu sind aber umfangreiche empirische Untersuchungen notwendig; das ist nicht ausschließlich auf theoretischem Wege möglich.

Im Kapitel 5 werden wir einige Vorschläge zur Ableitung eines konkreten Basismodells der Lexik machen, das denselben sprachlichen Realitätsausschnitt betrifft wie das Basismodell Köhlers, sich jedoch von diesem in mehreren Punkten unterscheidet.

Das Köhlersche Basismodell stellt eine gewisse Idealisierung des sprachlichen Systems dar, da u.a. nur die Beziehungen (und zwar nur bestimmte der möglichen Paarbeziehungen) lexikalischer Eigenschaften untersucht werden und andere Relationen zwischen diesen und anderen sprachlichen und außersprachlichen Eigenschaften nicht oder nur über (zeitlich) konstante Anteile berücksichtigt werden. Außerdem handelt es sich bei den untersuchten Relationen nicht um Beziehungen zwischen den Eigenschaftswerten konkreter lexikalischer Einheiten, sondern zwischen Ausprägungen konkreter lexikalischer Einheiten hinsichtlich einer Eigenschaft und der Mittelwerte über bestimmte Gruppen lexikalischer Einheiten hinsichtlich einer zweiten Eigenschaft.

Diese Abhängigkeiten betreffen somit keine Relationen zwischen Ausprägungen konkreter lexikalischer Einheiten hinsichtlich der untersuchten Eigenschaften, sondern betreffen jeweils bestimmte Lexemgruppen (die hinsichtlich der ersten Eigenschaft die gleiche Ausprägung haben) mit recht unterschiedlichem Umfang (z.B. gibt es eine relativ große Zahl lexikalischer Einheiten mit der Frequenz F = 1, eine wesentlich kleinere Zahl mit der Frequenz F = 2 usw., was für die Untersuchung der Abhängigkeit zwischen der Frequenz und mittleren Länge von Bedeutung ist), d.h., sie betreffen bestimmte Makroabhängigkeiten bzw. Makroprozesse der Lexik, wenn man den dynamischen Aspekt dieser Abhängigkeiten betrachtet.

Als Mikroprozesse sollen dementsprechend dynamische Abhängigkeiten zwischen Ausprägungen konkreter lexikalischer Einheiten hinsichtlich der untersuchten Eigenschaften bezeichnet werden. Das in Hammerl, Maj (1989) vorgestellte Lexikmodell betrifft solche Mikroprozesse lexikalischer Einheiten, das Lexikmodell von Köhler (1986) betrifft Makroabhängigkeiten.

Der Mittelwert der Mikroabhängigkeiten kann in bestimmten Fällen mit dem entsprechenden Wert der Makroabhängigkeiten mehr oder weniger übereinstimmen, dies muß aber nicht in jedem Falle so sein. Ein solcher Fall liegt unserer Meinung nach bei der Beschreibung der Abhängigkeit zwischen der Häufigkeit F und der mittleren Länge L im Köhlerschen Modell vor (vgl. hierzu auch Kapitel 3 und 5).

Vor der Untersuchung der Zusammenhänge zwischen verschiedenen Eigenschaften (im Rahmen eines bestimmten Modells) muß man sich auch über den "allgemeinen Typ" dieser Abhängigkeiten Klarheit verschaffen,

d.h. entscheiden, ob diese Abhängigkeiten als kausale Abhängigkeiten in nur einer Abhängigkeitsrichtung interpretiert werden können (z.B. Veränderungen der Frequenz verursachen Veränderungen der mittleren Länge lexikalischer Einheiten), als kausale Abhängigkeiten in beiden Richtungen (z.B.: Veränderungen der Frequenz bewirken Veränderungen der mittleren Länge und entsprechend umgekehrt) und unter welchen Bedingungen oder als nichtkausale Abhängigkeiten.

Die klassische Physik z.B. beruht auf der Überzeugung, daß jeder Vorgang in der Natur durch ganz bestimmte Ursachen in gesetzmäßiger Weise bestimmt ist und daß gleiche Ursachen stets gleiche Wirkungen nach sich ziehen. Die Unterscheidung von Ursache und Wirkung ist aber in mehreren Fällen überhaupt nicht möglich: So kann man z.B. bei einem Kondensator nicht sagen, ob die Ladung die Ursache des elektrischen Feldes oder das Feld die Ursache der Ladung ist.

Wenn man von Kausalen oder "quasi-kausalen" (Köhler 1986, 70) Abhängigkeiten spricht, so muß auch gesagt werden, was unter diesen Begriffen zu verstehen ist. In der Physik z.B. kommt es "bei dem Begriff des Kausalzusammenhanges weder auf die zeitliche Aufeinanderfolge noch auf die Unterscheidung von Ursache und Wirkung an, sondern entscheidend ist allein der gesetzmäßige Zusammenhang zwischen verschiedenen physikalischen Größen" (Höfling 1971, 327), der experimentell nachgeprüft werden kann. Der experimentelle Nachweis einer so verstandenen Kausalbeziehung zwischen zwei sprachlichen Eigenschaften ist oft nur sehr beschränkt möglich. Aus den genannten Gründen muß der Kausalbegriff bei dessen Verwendung in linguistischen Untersuchungen immer klar definiert werden.

Köhler (1986:10) schreibt, daß über die Form der Abhängigkeit zwischen der Länge und Frequenz lexikalischer Einheiten keine Übereinstimmung herrscht und daß selbst die Abhängigkeitsrichtung umstritten ist. Die von Köhler vorgelegte Argumentation für die Wahl einer konkreten Abhängigkeitsrichtung (L hängt von F ab) als Resultat des Wirkens bestimmter Bedürfnisse ist natürlich kein Beweis für die Richtigkeit dieser Wahl. Es handelt sich hier um Annahmen, die es zu überprüfen gilt. Vor dieser Überprüfung ist es notwendig, zunächst einmal exakte Kriterien darüber auszuarbeiten, wie man eine Abhängigkeitsrichtung erkennt und wie man sie nachprüfen kann; es müssen aber auch andere Abhängigkeitsrichtungen ganz oder zum Teil ausgeschlossen werden. Solange dies aber nicht möglich ist, ist der Standpunkt Köhlers, daß die Frequenz F die unabhängige Variable in der Beziehung zwischen F und L ist, wenig überzeugend. Aus diesem Grunde weisen wir auch den Einwand von Köhler

(in diesem Band,....) zurück, daß wir (im Artikel Hammerl, Maj 1989) die Abhängigkeitsrichtungen zwischen den lexikalischen Eigenschaften der Frequenz F und der Länge L verdrehen und unsere eigenen Voraussetzungen negieren ("denn die Frequenz ist die unabhängige Variable" in der Relation zwischen F und L (Köhler, ebenda)), denn das muß erst einmal bewiesen werden.

Im Kapitel 5 werden wir an einem einfachen Beispiel die von Köhler (in diesem Band) geforderten linguistischen Vorüberlegungen darstellen, die durchaus andere als nur die von Köhler anerkannten Abhängigkeitsrichtungen zwischen den lexikalischen Eigenschaften zulassen.

Köhler (1986) geht bei der Modellierung der Abhängigkeiten zwischen den durch die unabhängige Variable x und der abhängigen Variablen y repräsentierten lexikalischen Eigenschaften von der Differentialgleichung

$$\frac{dy}{y} = a \frac{dx}{x} \qquad (1)$$

aus, deren Lösung zur Funktion

$$y = ax^b \qquad (2)$$

bzw. deren linearen Form

$$\ln y = \ln a + b \ln x \qquad (3)$$

führt.

Wenn nun diese Abhängigkeit für konkrete lexikalische Eigenschaften nicht bestätigt werden kann, so müssen die theoretischen Grundsätze, die zur Aufstellung der Differentialgleichung (1) führten, modifiziert werden.

Für die Modellierung von Makroabhängigkeiten lexikalischer Eigenschaften, die wir an polnischem Sprachmaterial überprüft haben, gehen wir von teilweise anderen theoretischen Grundsätzen (einer anderen Differentialgleichung) über die Abhängigkeiten zwischen den von Köhler (1986) untersuchten lexikalischen Eigenschaften aus.

Bei der Untersuchung komplexer Sachverhalte der Realität wird man von einfachen Modellen ausgehen, die wesentliche Inhalte der untersuchten Sachverhalte betreffen, von bestimmten Basismodellen, die die Beziehungen zwischen einer begrenzten Zahl von Basiselementen betreffen. Die Wahl der Elemente des Basissystems ist zwar grundsätzlich willkürlich (wie z.B. in der Physik die Wahl der Grundgrößen), man wird aber diejenigen wählen, die aus der Sicht des jeweiligen Forschers als die wesentlichsten und einfachsten angesehen werden können und die dann zur Kennzeichnung und Erklärung komplizierter Sachverhalte dienen können (so wird z.B. in der Physik aus den Grundgrößen ein ganzes System von weiteren Größen abgeleitet).

Köhler (1986) berücksichtigt in seinem Basismodell 4 (Grund)größen aus: die Länge, die Frequenz, die Polylexie und Polytextie lexikalischer Einheiten.

Ohne es besonders zu bemerken, bleibt die Größe "Zeit" (eine physikalische Grundgröße) außer acht, was vor allem deshalb von Bedeutung ist, da Köhler ein dynamisches Lexikmodell ableiten will. Den Aspekt der "Dynamik" eines Modells wollen wir im nächsten Kapitel etwas genauer darstellen.

3. Zum Begriff der Dynamik von Modellen - dynamische Mikro- und Makrorelationen der Lexik

Das von Köhler (1986) vorgestellte Lexikmodell wird als "dynamisches" Modell eingeführt, ohne den Begriff der "Dynamik" eines Modells explizit zu definieren. Als dynamische Eigenschaften lexikalischer Einheiten werden "Dauer, Geschwindigkeit und Beschleunigung" der Bewegungen der lexikalischen Einheiten im lexikalischen Raum dargestellt, wobei auch diese Begriffe ohne exakte Erklärung bleiben.

Unter dem Begriff Dynamik (von griech. "dynamis" - Kraft) wird in der klassischen Naturwissenschaft die Lehre von den Kräften der Bewegung (genauer: der Bewegungsveränderung) verstanden (Historisches Wörterbuch der Philosophie 1972:302); in der Mechanik z.B. wird Bewegung als Ortsveränderung gegenüber einem Punkt im Bezugssystem angesehen (vgl. z.B. Dorn 1972:12), Dauer, Geschwindigkeit und Beschleunigung der Bewegung werden als zeitabhängige Größen eingeführt.

Auch in speziellen Arbeiten, die sich mit dynamischen Systemen beschäftigen, wird der Begriff "Dynamik" mit zeitabhängigen Veränderungen bestimmter Größen verbunden; so schreibt z.B. Luenberger (1979:1): "The term dynamic refers to phenomena that produce time-changing pattern, the characteristics of the pattern at one time being interrelated with those at other times. The term is nearly synonymous with time-evolution or pattern of change ...".

Was versteht nun Köhler (1986) unter dem Begriff "Dynamik"?

In dem Kapitel zum Stichwort "Dynamik" (Köhler 1986, 85 ff.) wird der diachronische Aspekt von Modellgrößen diskutiert, also die Verände-

rungen dieser Größen in der Zeit; an anderer Stelle (Köhler 1986, 40) schreibt Köhler: "... das zu konstituierende Modell soll aber den dynamischen Aspekt, die Diachronie, einschließen. Die Abbildung von Veränderungen der lexikalischen Einheiten ist möglich, wenn das Achsensystem der strukturellen Eigenschaften zusätzlich eine Zeitachse erhält."

Die aus diesen Textfragmenten ableitbare Schlußfolgerung, daß unter der Dynamik dieses Modells Veränderungen lexikalischer Einheiten in der Zeit verstanden werden können, wird in Köhler (in diesem Band) abgelehnt. Köhler (ebenda) untersucht keine zeitlichen Veränderungen lexikalischer Eigenschaften, sondern gegenseitige Abhängigkeiten lexikalischer Eigenschaften ohne explizite Berücksichtigung der Zeit. Aus diesem Grunde haben wir in Hammerl, Maj (1989) das Köhlersche Submodell, welches die Abhängigkeit der Länge von deren Frequenz beschreiben soll, als "statisches" Modell bezeichnet.

In Antwort auf diese Charakterisierung präzisiert Köhler (in diesem Band), was er unter dem dynamischen Aspekt seines Modells versteht, nämlich die funktionalen Abhängigkeiten der Systemgrößen voneinander (d.h. für das genannte Submodell: die Abhängigkeit der mittleren Länge von der Frequenz) und zieht den Schluß, daß wir den Begriff "Dynamik" ausschließlich für Veränderungen der Systemgrößen auf der Zeitachse reservieren. Dem ist nicht so. Auch in Hammerl, Maj (1989) werden unter dynamischen Abhängigkeiten funktionale Abhängigkeiten zwischen Systemgrößen verstanden, ohne aber - wie bei Köhler - deren zeitlichen Veränderungen außer acht zu lassen, d.h., wir untersuchen die Relationen zwischen der Länge L(t) (in Abhängigkeit von der Zeit t) und der Frequenz F(t); erst dann, wenn als bewiesen gilt, daß der Einfluß der Variablen Zeit auf die funktionalen Abhängigkeiten zwischen beiden Größen konstant ist (was Köhler einfach voraussetzt), braucht diese bei der Modellierung nicht gesondert berücksichtigt zu werden.

In diesem Zusammenhang möchten wir auf einen anderen wichtigen Aspekt hinweisen: Köhler untersucht eigentlich nicht die Abhängigkeiten zwischen 2 Eigenschaften, die jeder einzelnen lexikalischen Einheit zukommen, sondern zwischen einer Eigenschaft, die allen einzelnen lexikalischen Einheiten zukommt, und einer Eigenschaft, die einer ganzen Lexemgruppe zukommt. So wird z.B. die Abhängigkeit der Frequenz einzelner lexikalischer Einheiten von der mittleren Länge all der lexikalischen Einheiten untersucht, die mit diesen konkreten Frequenzen auftreten (vgl. auch Hammerl, Maj 1989, Abb.6,7). Solche Relationen (Abhängigkeiten) werden im folgenden als Makrorelationen bezeichnet. Die entsprechende Mikrorelation würde die Abhängigkeit der Länge konkreter lexikalischer Einheiten von deren konkreten Frequenzen betreffen. Bei der

Beschreibung dieser (und auch anderer) Mikrorelationen ist die Berücksichtigung der Zeit von besonderer Bedeutung (untersucht werden muß also die Abhängigkeit L(t) von F(t)).

Auf die Ausprägung einer konkreten lexikalischen Einheit bezüglich bestimmter lexikalischer Eigenschaften (z.B. der Lexemlänge) hat eine sehr große Zahl von sprachlichen und außersprachlichen Faktoren Einfluß, wobei weder die Zahl dieser Faktoren noch die Spezifik deren Einflußnahme auf die untersuchte Eigenschaft bekannt ist. Somit ist auch deren Berücksichtigung bei der Modellierung und Beschreibung der gegenseitigen Abhängigkeiten zwischen bestimmten lexikalischen Eigenschaften unmöglich. In bestimmten Fällen können einige dieser Faktoren erkannt und deren Einfluß analysiert werden. So ist es z.B. bei der Untersuchung der Abhängigkeit der Lexemlänge von der Lexemhäufigkeit, wo z.B. auch der Einfluß der Polysemie des jeweiligen Lexems auf dessen Länge berücksichtigt werden kann.

In den Fällen, wo die gegenseitigen Abhängigkeiten bestimmter Eigenschaften konkreter Einheiten durch eine Vielzahl von bisher noch nicht erkannten Faktoren beeinflußt wird, kann die Berücksichtigung der Variablen Zeit, d.h. der zeitabhängigen Ausprägungen der konkreten Einheiten bezüglich der untersuchten Eigenschaften, mit einem erheblichen Erkenntnisfortschritt verbunden sein. So ist es z.B. bei der Untersuchung der Gesetzmäßigkeiten des radioaktiven Zerfalls, wo zunächst noch nicht erklärt werden kann, wann welches konkrete Atom zerfällt und warum nicht früher oder später, es kann aber angegeben werden, wie viele konkrete Atome pro Zeiteinheit im Durchschnitt zerfallen. Die Zeit stellt hier somit eine Bezugsgröße dar, und die Abhängigkeit der Zahl N der zerfallenen Atome von der Zeit t stellt keine kausale Abhängigkeit dar.

Bei der von Hammerl, Maj (1989) besprochenen Untersuchung der zeitlich veränderlichen Länge L(t) von der Frequenz F(t) stellt die Zeit genauso nur eine Bezugsgröße dar, keine unabhängige Variable im Sinne kausaler Abhängigkeiten.

Bei der Beschreibung der entsprechenden Makrorelation können sich unter Umständen die individuellen zeitabhängigen Einflüsse aufheben, so daß unter diesem Gesichtspunkt eine Elimination der Zeit als Variablen eventuell möglich wäre.

Dieses Vorgehen kann jedoch nur so lange gerechtfertigt werden, wie dadurch die Ergebnisse empirischer Untersuchungen erklärt werden können. Aber gerade ein gegensätzlicher Fall scheint bei der Untersuchung der Abhängigkeit zwischen der mittleren Länge und der Frequenz, die Köhler (1986) an deutschem Sprachmaterial durchführt, vorzuliegen. Bei der Untersuchung dieser Abhängigkeit stellt Köhler eine "Oszillation" der

Lexik an der Frequenzachse fest und versucht diese über eine Differentialgleichung zweiter Ordnung zu erklären. Die Feststellung Köhlers, es handle sich hierbei um eine Oszillation in der Frequenzdimension, nicht in der Zeitdimension, ist unbegründet, den wenn man die Variable Zeit nicht berücksichtigt, wie will man dann eine Oszillation der Länge lexikalischer Einheiten in der Zeit ausschließen?

Wir vermuten, daß diese Oszillation dadurch zustande kommt, daß die durch Frequenzveränderungen hervorgerufenen Längenveränderungen einzelner lexikalischer Einheiten erstens erst nach einer bestimmten Zeit auftreten, d.h. einer zeitlichen Verzögerung unterliegen, und gleichzeitig von anderen lexikalischen Eigenschaften (z.B. der Polylexie) und noch vieler anderer sprachlicher und außersprachlicher Eigenschaften abhängen. Der Einfluß dieser Größen kann bei jeder lexikalischer Einheit ein anderer sein und deren Globaleinfluß kann nicht als eine konstante Größe hinsichtlich der Variablen Zeit angesehen werden.

Eine Verkürzung einer lexikalischen Einheit L1 bei erhöhter Auftrittswahrscheinlichkeit muß z.B. gar nicht eintreten (und in den meisten Fällen ist es wohl auch so); dafür kann z.B. ein anderes, kürzeres Lexem L2 die "häufig" verwendete Bedeutung des Lexems L1 übernehmen. Dies sind jedoch Prozesse, die bestimmten zeitlichen Verzögerungen unterliegen und bei verschiedenen Einzellexemen anders aussehen können. Untersucht man dann - wie Köhler - die Abhängigkeit der mittleren Länge lexikalischer Einheiten von deren Frequenz zu einem konkreten (obwohl: beliebigen) Zeitpunkt t, so werden diese Mikroprozesse nicht berücksichtigt, was zwangsläufig zu einer bestimmten Streuung der jeweiligen Variablenmittelwerte führen muß. Ob es sich hier um periodische Schwankungen handelt (Köhler 1986, 196ff.), muß erst bestätigt werden.

Wir versuchen, ausgehend vom Köhlerschen Lexikmodell ein solches dynamisches Lexikmodell zu erarbeiten (einige wichtige Hinweise hierzu findet der Leser im Kapitel 5), welches diese Befunde zum großen Teil erklären kann. Dazu werden mehrdimensionale Abhängigkeiten lexikalischer Eigenschaften untersucht werden müssen.

Zusammenfassend soll noch einmal betont werden, daß die Berücksichtigung der Variablen Zeit bei der Untersuchung der Abhängigkeiten zwischen bestimmten Eigenschaften sehr nutzbringend sein kann, vor allem dort, wo die untersuchten Einheiten einer Vielzahl verschiedener, nur schwer oder überhaupt nicht erfaßbarer Faktoren ausgesetzt sind. Die Zeit stellt aber lediglich eine Bezugsgröße (Ordnungsgröße) dar und keine unabhängige Größe im Sinne kausaler Abhängigkeiten.

Für die Beschreibung von (Makro)relationen innerhalb bestimmter Modelle können in diesen Fällen z.T. die zeitveränderlichen Einflüsse als

konstante Anteile berücksichtigt werden. So gilt auch das Gravitationsgesetz (vgl. Köhler, in diesem Band) nur innerhalb eines bestimmten Modells von 2 Körpern (Massenpunkten), da z.B. der Einfluß deren konkreten Form, deren Eigenrotation, der Luftwiderstand und der Einfluß anderer sich bewegender Körper nicht berücksichtigt wird. Nur unter diesen Bedingungen ist es möglich gewesen, das Gravitationsgesetz ohne Bezug auf die Variable Zeit abzuleiten.

Wenn Köhler ein dynamisches Lexikmodell erarbeiten will, so muß der Begriff der Dynamik exakt erklärt werden, um Fehlinterpretationen auszuschließen.

Wenn Köhler bei der Untersuchung von Makrorelationen den dynamischen Aspekt als Untersuchung der funktionalen Abhängigkeiten zwischen den untersuchten Eigenschaften versteht, so ist das entsprechende Modell natürlich ein "dynamisches" Modell. Hammerl, Maj (1989) wenden keinen gegensätzlichen Dynamikbegriff an, sondern einen allgemeineren, indem die Dynamik in das Modell über funktionale, zeitabhängige Relationen eingeht.

Um auf dieser Grundlage eventuellen weiteren Mißverständnissen in der Interpretation des Artikels von Hammerl, Maj (1989) vorzubeugen, soll nun noch einmal kurz das Hauptanliegen dieses Artikels dargelegt werden, bevor im Kapitel 5 einige neue Aspekte der Untersuchung der Abhängigkeiten lexikalischer Eigenschaften (besonders der Relation zwischen der Länge und deren Frequenz) diskutiert werden.

4. Grundanliegen des Aufsatzes von Hammerl, Maj (1989)

Im Aufsatz von Hammerl, Maj (1989) sollte in erster Linie eine Untersuchungsmethode vorgestellt werden (die an Verfahren und Techniken, die aus der Systemtheorie und Kybernetik bekannt sind, anknüpft) zur Untersuchung dynamischer Mikro- und Makrorelationen zwischen Eigenschaften lexikalischer Einheiten, deren komplexe Anwendung zunächst noch - aufgrund einer unzureichenden Datenlage und fehlender oder noch nicht umfassend bestätigter diachronischer Sprachgesetze - nicht möglich ist. Die prinzipielle Möglichkeit der Anwendung der vorgestellten Untersuchungsmethode bei der Ableitung und Analyse eines lexikalischen Submodells wird am Beispiel der Abhängigkeit zwischen der Länge L und der Frequenz F lexikalischer Einheiten aufgezeigt.

Aus den genannten Gründen konnten die Autoren des genannten Aufsatzes in einigen Fällen nur von stark vereinfachten Modellen ausgehen.

Dieser Aufsatz stellt somit in erster Linie eine theoretische Arbeit dar; trotz der Tatsache der noch fehlenden empirischen Überprüfungsmöglichkeit der vorgestellten Untersuchungsmethode ist die Schlußfolgerung Köhlers, wir kämen zu keinem verwertbaren Ergebnis, sicher verfrüht. Diesbezüglich sei nur daran erinnert, daß "verwertbar" stets verwertbar im Rahmen eines bestimmten Modells, vor einem bestimmten methodologischen Hintergrund, von konkreten Forschern zu einem konkreten Zeitpunkt und zu konkreten Zwecken heißt.

Ein für die Autoren selbst wichtiges Ergebnis war die Möglichkeit der Modellierung mehrdimensionaler Abhängigkeiten zwischen Eigenschaften sprachlicher Einheiten (bei Berücksichtigung deren Veränderungen in der Zeit) unter Anwendung systemtheoretischer Methoden, der Möglichkeit der exakten mathematischen Beschreibung nicht nur der Relationen zwischen abhängigen und unabhängigen Größen (der Eingangs- und Ausgangsgrößen des Modells), sondern auch der komplizierten Prozesse im "Modellinnern", die bisher nur zum Teil beschrieben werden konnten, der exakten - und nicht intuitiven - Darstellung dieser Abhängigkeiten in einem Blockschaltbild und über die Computeranwendung die Möglichkeit der Simulation des Modellverhaltens. Somit wurde auch grundsätzlich eine reale Möglichkeit der empirischen Überprüfung der vom Modell vorausgesagten dynamischen Abhängigkeiten zwischen den jeweiligen Eigenschaften sprachlicher Einheiten geschaffen.

Diese Prognose wie auch die gesamte Modellierung der dynamischen Abhängigkeiten setzt die Kenntnis der diachronischen Gesetze über die Veränderungen der jeweiligen Eigenschaften voraus. Bei der Demonstration der Anwendung der vorgestellten Untersuchungsmethode am Beispiel der Modellierung der Abhängigkeit zwischen der Länge und der Frequenz lexikalischer Einheiten gingen wir vom verallgemeinerten Piotrovskigesetz aus (Altmann 1983), wobei jedoch nicht behauptet werden sollte, daß die auf dieser Grundlage von uns gewählten Funktionen L=f(t) und F=g(t) die realen zeitlichen Längen- und Frequenzänderungen lexikalischer Einheiten in allen Fällen exakt beschreiben. Somit ist die Kritik Köhlers (in diesem Band), daß es sich bei der zeitlichen Längenveränderung nicht um einen Verbreitungsvorgang handelt einerseits gerechtfertigt, andererseits ist aber auch der (bisher noch nicht mathematisch beschriebene und ebenso empirisch überprüfte) Vorschlag Köhlers (ebenda), daß es sich um die Entstehung neuer Formen in dem Maße handelt, wie vorher die entsprechenden alten bereits verbreitet waren, unserer Meinung nach eine

die Realität zu stark vereinfachende Vorstellung, denn Längenveränderungen hängen nicht nur von den beits vorhandenen Längen lexikalischer Einheiten ab, sondern in mindestens demselben Maße von den zeitlichen Polylexie- und Frequenzveränderungen (vgl. hierzu: Kapitel 5). Die von Hammerl, Maj (1989) vorgestellte Untersuchungsmethode kann für jedes bessere und adäquatere Modell für die zeitlichen Veränderungen der Häufigkeit und Länge sprachlicher Einheiten angewandt werden, sobald solche Modelle nur vorliegen.

Wenn nun von uns bewußt verschiedene Modellvarianten hinsichtlich ihres Verhaltens in Simulationsexperimenten untersucht wurden, so nur deshalb, weil aus den dann ableitbaren Konsequenzen wiederum Rückschlüsse über die Voraussetzungen der jeweiligen Modellvariante selbst möglich sind. Dann gehören eben "linguistisch unsinnige" Ergebnisse (vgl. Köhler, in diesem Band) ebenso zu positiven (verwertbaren) Ergebnissen. Wir verdrehen die von Köhler bei der Untersuchung der Abhängigkeiten zwischen der Länge und der Häufigkeit vorgeschlagenen Abhängigkeitsrichtungen bewußt, um eben auch aufzeigen zu können, wie man somit rückschließend wieder Informationen über die bisher stark intuitiv angenommenen Abhängigkeitsrichtungen gewinnen kann. Außerdem ist die Köhlersche Argumentation darüber, daß die Frequenz die unabhängige Variable sei, wenig überzeugend, hängen doch Länge, Häufigkeit, Polylexie und Polytextie und andere Eigenschaften lexikalischer Einheiten auf recht komplizierte Weise voneinander ab (vgl. hierzu: Kapitel 5).

Inwieweit die Anwendung der von Hammerl, Maj (1989) vorgestellten und sicher noch völlig unzureichend an konkretem sprachlichen Material erklärten Methode in sprachwissenschaftlichen Untersuchungen nutzbringend ist, können erst weiterführende Untersuchungen zeigen. Zum Einwand Köhlers aber, daß es sich um eine bloße Übernahme einer anderswo bewährten Methode handelt, muß gesagt werden, daß nur ein schon in vielen verschiedenen Wissenschaften angewandtes allgemeines methodisches Instrumentarium übernommen wurde (und vor allem darüber wurde im Aufsatz von Hammerl, Maj (1989) gesprochen). Diese allgemeine methodische Vorgehensweisen und mathematischen Beschreibungsmethoden müssen dann unter Berücksichtigung von Wissen der jeweiligen konkreten Wissenschaft, der konkreten Fragestellung auf spezielle Sachverhalte angewandt werden. Um dies aber exakt und umfassend realisieren zu können, müssen umfangreiche empirische und theoretische linguistische Untersuchungen über das zeitliche Verhalten von Eigenschaften lexikalischer Einheiten durchgeführt werden.

Auch die Anmerkung Köhlers, daß bei der von uns angewandten Vorgehensweise kein inhaltlicher Bezug zwischen den Größen Länge und Frequenz hergestellt wird, ist muß zurückgewiesen werden.

Aus linguistischen Vorüberlegungen müssen zunächst die zeitlichen Veränderungen der untersuchten lexikalischen Eigenschaften (einer konkreten Gruppe lexikalischer Einheiten oder von Einzellexemen) modelliert und empirisch überprüft werden. Auf dieser Grundlage (die bisher ja noch nicht als gesichert gelten kann) kann dann mit der vorgelegten Methode (nach den angegebenen Regeln) ein Modell (ein Glied zwischen Eingangs- und Ausgangsgrößen) so abgeleitet werden, daß die Eingangsgrößen gerade die entsprechenden Ausgangsgrößen bewirken. Das "Modellinnere" wird somit nicht arbiträr oder axiomatisch vorgegeben, sondern aus dem Ergebnis der Untersuchungen rekonstruiert. Je genauer einwirkende Eingangs- und Ausgangsgrößen unterschieden und im Modellansatz berücksichtigt werden, desto adäquater beschreibt das Modell die Realität.

Nach diesen grundsätzlichen Überlegungen sollen nun noch einige - bisher noch nicht umfassend überprüfte - Vorschläge für die Ableitung eines dynamischen Lexikmodells zur Untersuchung von Makrorelationen besprochen werden, welches sich teilweise vom Lexikmodell Köhlers unterscheidet. Eine detailliertere Vorstellung dieses Modells erfolgt in einem späteren Aufsatz.

5. Vorschläge für die Ableitung eines dynamischen Lexikmodells (zur Beschreibung von ausgewählten Makrorelationen)

Die unten angeführten Vorschläge sind Resultat der Modellierung der Verteilung und der Makrorelationen zwischen den Eigenschaften Häufigkeit, Länge (in Phonemen, Silben, Buchstaben) und Polylexie von Lexemen und deren Überprüfung an polnischem Sprachmaterial (teilweise auch an entsprechendem Sprachmaterial anderer Sprachen).

Köhler (1986) hat die Abhängigkeit zwischen je 2 von insgesamt 4 lexikalischen Eigenschaften mathematisch beschrieben und an deutschem Sprachmaterial überprüft. Es fällt jedoch in vielen Fällen auf, daß die Anpassung der theoretischen Kurve an die empirischen Werte nach dem F-Test relativ gut ist, obwohl dies aus der graphischen Darstellung (wo empirische und theoretische Werte eingetragen wurden) nicht hervorgeht; dies liegt wohl auch daran, daß die Bedingungen für die Anwendung des F-Testes (z.B. daß die Streuungen der Einzelwerte $y_{i,j}$ für jedes x_i um

deren Mittelwert y_i (die ja von Köhler berücksichtigt werden) gleich sein und daß die Werte für $y_{i,j}$ für alle i normalverteilt sein müssen) nicht in allen Fällen als erfüllt gelten können. Wenn dann außerdem die Werte für y_i fast nur auf bestimmte (und relativ kleine) Intervalle für die entsprechende Variable x beschränkt sind, so kann auch der F-Test nicht uneingeschränkt zur Begründung der guten Anpassung der theoretischen und empirischen Werte über den gesamten Bereich der Variablen x dienen.

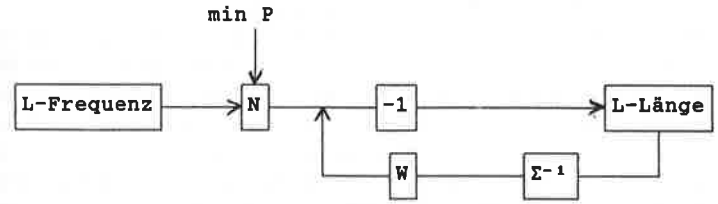
Im folgenden wollen wir die Abhängigkeit der Länge von der Frequenz, die wir ja auch in Hammerl, Maj (1989) zur Illustration herangezogen haben, etwas näher untersuchen. Dabei gehen wir von den Annahmen Köhlers zur Begründung dieser Abhängigkeit aus:

In Köhlers Modell ist die Frequenz die unabhängige Größe, die die mittlere Länge lexikalischer Einheiten als abhängige Größe bestimmt. Eine Vergrößerung der Frequenz z.B. soll somit eine Verkleinerung der mittleren Länge der Einheiten mit der jeweiligen Frequenz hervorrufen (die Veränderungsrate der Länge ist umgekehrt proportional zur Frequenz einer lexikalischen Einheit). Da jedoch eine unbegrenzte Längenverkleinerung (L-->0) nicht möglich ist, wurde noch eine zweite Annahme gemacht: Eine lexikalische Einheit wird um so stärker gekürzt, je länger sie ist (und umgekehrt).

Diese Annahmen führten Köhler auf eine Differentialgleichung 2. Ordnung, die die Oszillation der Lexik (siehe unten) beschreiben soll.

Zur Abhängigkeit der mittleren Länge von der Frequenz führt Köhler folgendes Blockschaltbild an (Abbildung 1):

Abbildung 1
Abhängigkeit der Länge von der Frequenz
nach Köhler (1986, 72)

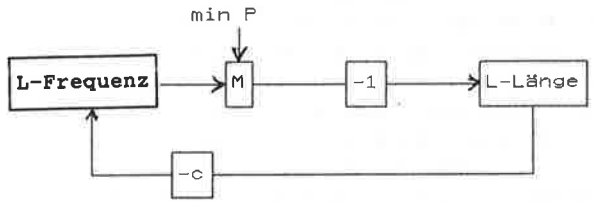


Hier steht min P für das Bedürfnis nach Minimierung des Produktionsaufwandes und wirkt über den Proportionalitätsfaktor N und -1 auf die L-Länge ein, diese wirkt wiederum zeitlich verzögert über den Proportionalitätsfaktor W und -1 wieder auf sich selbst zurück. Das würde bedeuten, daß eine durch Frequenzerhöhung hervorgerufene Längenverkleinerung

nach einer bestimmten Zeit selbst noch einmal eine Längenverkleinerung hervorrufen würde usw. Dies ist jedoch in der Wirklichkeit nicht möglich.

Köhler (1986:73) stellt auch noch eine andere Modellvariante vor, wo die Länge auf die Frequenz zurückwirkt. Das entsprechende Blockschaltbild sieht dann so aus (Abbildung 2):

Abbildung 2
Abhängigkeit der Länge von der Frequenz
nach Köhler (1986, 73)



In unserem Aufsatz (Hammerl, Maj 1989:13 ff) sind wir vor allem auf diese Modellvariante eingegangen und haben die Rückführung der Länge auf die Frequenz als zeitlich versetzte Rückkopplung bezeichnet (es sollte nicht ausgedrückt werden, daß auch Köhler diese Rückkopplung als zeitlich versetzt ansieht).

Köhler (in diesem Band) verbindet diese Rückkopplung ausschließlich mit der Vermutung, "daß ein Sprecher unter zwei gleicherweise möglichen (also gleichzeitig existierenden), aber verschieden langen Ausdrücken häufiger den kürzeren wählt" (und negiert die zeitliche Verzögerung dieser Rückkopplung). In Wirklichkeit sind aber die Wechselbeziehungen zwischen der Häufigkeit und der Frequenz weit komplizierter und nicht nur auf den von Köhler beschriebenen Aspekt beschränkt.

Nehmen wir an, ein Lexem hat zum Zeitpunkt t_0 die Länge L_0 und in der Zeit von t_0 bis t_1 wächst die Frequenz dieser lexikalischen Einheit um $F(t_1 - t_0)$ auf $F(t_1)$ an. Diese Frequenzerhöhung kann dann eine Längenverkleinerung $L(t_2 - t_0)$ $(t_0 < t_1 < t_2)$ hervorrufen, so daß die lexikalische Einheit zu einer bestimmten Zeit die Länge $L(t_2)$ hat. Diese Längenveränderung läuft nicht gleichzeitig mit der Frequenzerhöhung ab, sondern mit einer bestimmten zeitlichen Verzögerung. In anderen Fällen braucht eine solche Längenveränderung des Ausgangslexems gar nicht einzutreten z.B. überall dort, wo entweder die Länge des Ausgangslexems schon relativ klein ist (und keine kleineren Längenveränderungen als um ±1 Einheit, z.B. einen Buchstaben, ein Phonem, möglich sind) oder dort, wo das z.B. aus morphologischen, stilistischen, phonetischen Gründen

kaum möglich ist. Beispiele für Wörter, die in 2 Formen mit unterschiedlicher Länge auftreten (z.B. Universität - Uni, Diskothek - Disko, Eisenbahn - Bahn) gibt es innerhalb des gesamten Lexikons einer Sprache nur sehr wenige. Das deutet auch darauf hin, daß ein Lexem, dessen Frequenz sich relativ stark verändert, nur in sehr wenigen Fällen tatsächlich gekürzt wird, so daß dann 2 Formen (die längere und die kürzere) gleichzeitig existieren.

Bestimmte Veränderungen in der Anwendungshäufigkeit lexikalischer Einheiten treten ständig auf; es ist sicher nicht so, daß jede (auch noch so kleine) Veränderung der Frequenz der Wörter Längenveränderungen hervorrufen kann. Die Frequenzveränderungen F müssen erst einen gewissen Schwellenwert übersteigen, um zu einer möglichen Längenveränderung führen zu können (man bedenke auch, daß Länge und Frequenz diskrete, keine stetigen Variablen sind) und die neue Frequenz $F(t_1)$ muß dazu auch mindestens einen bestimmten Zeitraum δt aufrechterhalten werden, um als solche wirksam werden zu können.

Bei einer Häufigkeitsveränderung eines konkreten Lexems muß somit neben der relativ selten auftretenden Längenveränderung dieses Lexems noch eine andere Möglichkeit häufiger genutzt werden:

Wenn ein polysemes Lexem häufiger gebraucht wird als vorher, so ist das meistens auf die häufigere Verwendung nur einer Bedeutung des polysemen Lexems zurückzuführen. Trägt ein Lexem A z.B. drei selten verwendete Bedeutungen und eine immer häufiger werdende Bedeutung, so ist die Anwendung dieses Lexems in der häufigeren Bedeutung und der ursprünglichen Länge L nicht effektiv. Es ist wohl möglich, daß die Länge dieses Lexems gekürzt wird (wie wir das schon oben beschrieben haben), aber auf die Längenreduzierung dieses Lexems haben auch alle anderen Bedeutungen dieses Lexems (deren Anwendungshäufigkeit) Einfluß und auch aus diesem Grunde kann eine Längenreduzierung unterdrückt oder abgeschwächt werden. In diesen Fällen ist es möglich, daß ein anderes, kürzeres Lexem B (ein neu gebildetes oder ein schon existierendes) die häufig verwendete Bedeutung des Lexems A übernimmt. Diese Bedeutungsübertragung hängt aber nicht nur von der Länge des Lexems B ab, sondern auch von der Spezifik der eventuell schon vorhandenen Bedeutungen dieses Lexems, deren Anzahl (der Polylexie) und natürlich auch deren Anwendungshäufigkeiten ab. Somit können auch auf diesem Wege Synonyme entstehen (natürlich spielen bei der Bildung von Synonymen gleichzeitig noch viele andere Faktoren eine Rolle).

Diesen Prozeß versuchen wir nun noch etwas exakter und umfassender darzustellen:

Zeitpunkt t_0 : Gegeben ist ein Lexem A mit einer

Polylexie $Pl_a(t_0)$,
Frequenz $F_a(t_0)$ und
Länge $L_a(t_0)$.

In der Zeit vom Zeitpunkt t_0 bis t_1 nimmt die Häufigkeit dieses Lexems zu, alle anderen Eigenschaften bleiben unverändert, d.h.:

$$PL_a(t_0) = PL_a(t_1),$$
$$L_a(t_0) = L_a(t_1),$$
$$F_a(t_0) > F_a(t_1).$$

Zum Zeitpunkt t_1 existiert gleichzeitig ein Lexem B mit

$$PL_b(t_1) \geq PL_a(t_1),$$
$$L_b(t_1) < L_a(t_1),$$
$$F_b(t_1) > F_a(t_1).$$

Während der Zeit von t_1 bis t_2 wird die häufigste Bedeutung von Lexem A auf Lexem B übertragen, welches nun in der Regel für die Realisation dieser Bedeutung in der sprachlichen Kommunikation angewandt wird. Somit kann z.B. gelten:

$$Pl_a(t_2) \leq PL_a(t_1),$$
$$L_a(t_2) = L_a(t_1),$$
$$F_a(t_2) \geq F_a(t_1)$$

und

$$PL_b(t_2) > PL_b(t_1),$$
$$L_b(t_2) = L_b(t_1),$$
$$F_b(t_2) > F_b(t_1).$$

Nach der Übernahme der häufigeren Bedeutung durch Lexem B hat

a) die Frequenz von Lexem A abgenommen , d.h. $F_a(t_2) = F_a(t_1)$ und eventuell sogar $F_a(t_2) \leq F_a(t_0)$, die Polylexie dieses Lexems (wenn es z.B. nicht mehr oder nur sehr selten in der Bedeutung gebraucht wird, die auf Lexem B übertragen wurde) hat abgenommen (bzw. wird abnehmen), d.h. $PL_a(t_2) \leq PL_a(t_1) = PL_a(t_0)$, die Länge ist dagegen gleich geblieben, d.h. $L_a(t_2) = L_a(t_1) = L_a(t_0)$.

Lexem A hat zwar noch die ursprüngliche Länge, aber auch eine kleinere Frequenz und vor allem eine "kleinere" Polylexie (da ja Lexem A in der auf B übertragenen Bedeutung nur noch sehr selten bzw. gar nicht mehr auftritt). Dieses Lexem hat demzufolge bezüglich dieser beiden Eigenschaften ein Defizit; für dieses Lexem kann das z.B. bedeuten, daß es wesentlich kürzer ist als andere Lexeme mit einer so niedrigen Polylexie und Frequenz.

b) Die Frequenz von Lexem B hat zugenommen, d.h. $F_b(t_2) > F_b(t_1)$, die Polylexie von Lexem B hat zugenommen, d.h. $PL_b(t_2) > PL_b(t_1)$, die Länge von Lexem B hat sich nicht geändert, d.h. $L_b(t_2) = L_b(t_1)$.

Lexem B hat bezüglich der Eigenschaften Frequenz und Polylexie einen Überschuß im Vergleich zu den früheren Werten; das kann z.B. bedeuten, daß dieses Lexem länger ist als andere Lexeme mit einer so großen Polylexie und Frequenz.

Infolge der dargestellten Veränderungen kann z.B. Lexem A eine andere Bedeutung aufnehmen, um so das Defizit bezüglich der Frequenz und Polylexie möglichst auszugleichen. Lexem B kann z.B. eine seiner seltenen Bedeutungen an ein anderes Lexem abgeben (oder verkürzt werden), um den Überschuß bezüglich der Frequenz und Polylexie auszugleichen.

Wir nehmen an, daß das lexikalische System durch solche Defizite und Überschüsse gekennzeichnet ist und daß ein Bestreben (entsprechend dem Prinzip der geringsten Anstrengung bzw. entsprechender Ökonomiebedürfnisse) zum möglichst besten und schnellsten Ausgleich dieser Defizite und Überschüsse besteht, d.h. daß ein Bestreben zur Annäherung an einen "optimalen Gleichgewichtszustand" (Normzustand) besteht, der dann erreicht wäre, wenn für alle Lexeme die Defizite und Überschüsse ausgeglichen wären.

Frequenzveränderungen können wohl in bestimmten Fällen zeitlich verzögerte Längenveränderungen hervorrufen, aber nicht in allen Fällen muß die Länge geändert werden, sondern es können auch Polylexieveränderungen auftreten und somit auch wieder rückwirkend Veränderungen der Frequenz des jeweiligen Lexems. D.h., daß eine zeitlich verzögerte Rückwirkung der Länge (über die Polylexie) auf die Frequenz auftreten kann.

Möglich wäre auch ein anderer Fall: Die Defizite von Lexem A können auch dadurch (zumindest teilweise) ausgeglichen werden, daß unter

dem Wirken des Spezifikationsbedürfnisses z.B. durch morphologische Mittel die Länge vergrößert wird; die Länge kann somit sogar so stark vergrößert werden, daß sie - im Verhältnis zur Frequenz und Polylexie des Lexems - über dem Normwert liegt, d.h. einen Überschuß aufweist. Entsprechend dem Ökonomiebedürfnis könnte dann zur Eliminierung dieses Überschusses auch die Anwendungshäufigkeit dieses Lexems reduziert werden (eventuell kann sogar eine der Bedeutungen dieses Lexems auf ein anderes Lexem übertragen werden); somit kann die ursprüngliche Frequenzerhöhung wohl einen Einfluß auf die Länge des Lexems ausüben, was aber selbst wieder rückwirkend Frequenzveränderungen bewirken kann.

Oben wurden nur zwei mögliche Fälle der durch Frequenzerhöhung eines Lexems bewirkten Veränderungen im lexikalischen System besprochen. Diese stellen nur einen kleinen Teil der insgesamt zulässigen Veränderungsmöglichkeiten dar.

Diese Schlußfolgerung wird auch noch durch eine andere Tatsache, die bisher nicht genügend Beachtung fand, gestützt:

Sprachliche Veränderungen werden durch äußere Bedürfnisse hervorgerufen und diese wirken wieder auf die Bedürfnisse zurück. Die Bedürfnisse selbst stehen in einer bestimmten Hierarchie, d.h., daß solch allgemeine Bedürfnisse wie das der Sprachökonomie durch verschiedene speziellere Bedürfnisse bedient werden können, die untereinander - in funktionaler Hinsicht - äquivalent sind (z.B. das Bedürfnis der Kontextspezifierung, Minimierung des Produktions- und Kodierungsaufwandes u.a.) und selbst wiederum durch Veränderungen verschiedener sprachlicher Eigenschaften (z.B. Frequenzerhöhung, Polylexieabhnahme, Längenvergrößerung von Lexemen usw.) in Erscheinung treten. Das Ökonomiebedürfnis kann für ein konkretes Lexem z.B. durch eine Längenzunahme (und einer späteren Frequenzabnahme und Polylexieabnahme) bedient werden oder durch Frequenzzunahme und einer möglichen späteren Längenreduzierung.

Aus diesen Gründen ist für die Untersuchung dieser Zusammenhänge eine starre Festlegung der Richtungen der Abhängigkeiten zwischen den sprachlichen Einheiten von untergeordneter Bedeutung; alle möglichen Relationen zwischen den einzelnen Eigenschaften sind prinzipiell zulässig, wenn sie nur der Erfüllung des entsprechenden Bedürfnisses dienen, aber nicht jedes konkrete Lexem läßt alle Relationen zu.

Da z.B. sowohl eine Längenveränderung eine Frequenzveränderung bewirken kann als auch umgekehrt, wird vermutet, daß auch die diese Abhängigkeiten beschreibenden Funktionen demselben Funktionstyp an-

gehören (das betrifft auch alle anderen Abhängigkeiten, z.B. PL = f(L), PL = f(F), F = f(L), L = f(F), L = f(PL), F = F(PL), L = f(PL,F) usw.

Auch die Verteilungen der Lexeme hinsichtlich der untersuchten Eigenschaften müssen aus den allgemeinen Annahmen, die zur Modellierung der Abhängigkeit zwischen je zwei Eigenschaften herangezogen wurden, abgeleitet werden können. Eine Überprüfung dieser Hypothesen soll Gegenstand einer gesonderten Arbeit sein.

Abschließend soll noch einmal kurz auf das Problem der Oszillation der Lexik (vgl. Köhler 1986, 137 ff.) eingegangen werden.

Wenn - wie wir oben gezeigt haben - innerhalb des lexikalischen Systems ein ständiges Bestreben zur Eliminierung von bestehenden Defiziten und Überschüssen besteht und als Resultat der dadurch hervorgerufenen dynamischen Veränderungen neue Defizite und Überschüsse entstehen, so resultiert daraus, daß sich zu keinem einzigen Zeitpunkt alle Lexeme im "Normzustand" befinden werden. Es kann somit auch nicht ohne weiteres angenommen werden, daß sich die individuellen Abweichungen der Eigenschaftsausprägungen der Lexeme vom jeweiligen Normwert bei der Mittelwertbildung über alle Lexeme aufheben.

Bei der Untersuchung der Makrorelation zwischen der Frequenz F und der mittleren Länge L und der Beschreibung dieser Abhängigkeit mit der Regressionsfunktion

$$L = aF^b , \qquad (4)$$

wo a und b Konstanten sind, stellt Köhler fest, daß die theoretischen Werte für die Länge um die entsprechenden empirischen Werte schwanken.

Im Rahmen des oben von uns skizzierten Modellansatzes können diese Schwankungen erstens darauf zurückgeführt werden, daß hier nur die Abhängigkeit zwischen der Länge und Frequenz (ohne Ausschaltung des Einflusses der Polylexie) und nicht die Abhängigkeit der Länge von der Frequenz bei Ausschaltung des Einflusses der Polylexie bzw. die gegenseitige Abhängigkeit zwischen der Länge, Frequenz und Polylexie untersucht wurde und zweitens Einflüsse, die durch die Defizite und Überschüsse konkreter Lexeme bezüglich dieser Eigenschaften hervorgerufen werden (Längenveränderungen und die ihnen entsprechenden Frequenzveränderungen treten nicht gleichzeitig auf, sondern mit einer bestimmten zeitlichen Verzögerung, die von Lexem zu Lexem recht verschieden sein kann), bei der Modellierung der untersuchten Abhängigkeit vernachlässigt wurden.

In diesem Kapitel sollten nur einige einfache Vorschläge zur Erarbeitung eines dynamischen Lexikmodells angeführt werden, die im Zusammenhang mit der von Köhler (in diesem Band) zum Artikel von Hammerl, Maj (1989) geäußerten Kritik stehen. Das dynamische Lexikmodell und die aus ihr resultierenden mehrdimensionalen Abhängigkeiten sind zur Zeit Gegenstand umfangreicher empirischer Untersuchungen und werden in einer gesonderten Arbeit vorgestellt.

Literatur

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

Dorn (1972), Physik. Oberstufe. Ausgabe A. Hannover-Berlin- Darmstadt-Dortmund, Schroedel.

Haken, H. (1978), Synergetics. Berlin-Heidelberg-New York, Springer.

Hammerl, R., Maj, J. (1989), Ein Beitrag zu Köhler's Modell der sprachlichen Selbstregulation. Glottometrika 10, 1-31.

Höfling, O. (1971), Lehrbuch der Physik. Oberstufe. Ausgabe A. Bonn, Dümmler.

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

Luenberger, D.G. (1979), Introduction to Dynamic Systems. Theory, Models and Applications. New York-Chichester-Brisbane-Toronto.

Historisches Wörterbuch der Philosophie (1972), Hrsg. von J.Ritter, Band 2. Basel-Stuttgart, Schwabe.

L.Hřebíček (ed.), Glottometrika 11, 1989

Zur Charakteristik dynamischer Modelle.

Anmerkungen zu einem Beitrag von R. Hammerl und J. Maj

R. Köhler, Bochum

Im Band 10 der *Glottometrika* (1988:1-31) veröffentlichten R. Hammerl und J. Maj einen "Beitrag zu Köhlers Modell der sprachlichen Selbstregulation". Nach einer Einführung in Grundbegriffe der Systemtheorie befassen sich die Autoren mit ausgewählten Problemen des lexikalischen Basissystems, wie es in Köhler (1986) dargestellt wurde. Der Aufsatz enthält interessante Kritik und wichtige Anregungen, die weiter verfolgt werden sollten. Die Ansichten der Autoren treffen sich z.T. mit meinen eigenen Wünschen nach Erweiterungen und Verbesserungen - z.T. aber entspringt ihre Kritik einer in den Ingenieurwissenschaften üblichen Sichtweise und einer auf deren Modellierungstradition eingeschränkten Auffassung zentraler Begriffe. Auf einige solche Punkte möchte ich hier eingehen.

1. Hammerl und Maj gründen ihre Hauptkritik an dem 1986 vorgestellten Basismodell auf ihre Behauptung, es handele sich um ein *statisches* Modell eines dynamischen Systems, da bei der Modellierung die Zeit nicht berücksichtigt worden sei.

Da ich dieses Modell nun ausdrücklich als dynamisches eingeführt hatte, differieren entweder unsere Meinungen darüber, was es bedeutet, die Zeit in einem Modellansatz zu berücksichtigen, oder aber darüber, worin in einem systemtheoretischen Modell die Dynamik besteht.

In der Darstellung des Basismodells von 1986 nehme ich zur Frage nach dem Zeitverhalten Bezug auf die Arbeiten von Piotrovski und vor allem Altmann/v.Buttlar/Rott/Strauß. Die 1985 von Altmann et al. theoretisch aus einem interaktionistischen Ansatz abgeleitete logistische Kurve wurde bisher durch alle empirischen Untersuchungen zum zeitlichen Verlauf von Eigenschaftsänderungen lexikalischer Einheiten bestätigt. Auf die *explizite* Formulierung dieses Zeitverhaltens im Modell wurde jedoch aus zwei Gründen verzichtet:

a) Die unzureichende Datenlage hätte keine gültigen empirischen Überprüfungen im Rahmen des modellierten Sprachausschnitts zugelassen.

b) Aus Gründen der methodischen Beschränkung auf die *funktionalen* dynamischen Beziehungen zwischen den Systemgrößen fand die Zeitachse keine Berücksichtigung in den Gleichungen. Das Modell blieb also in dieser (wie auch in mancher anderer) Hinsicht absichtlich unvollständig.

Die zweite Begründung muß angesichts der Kritik von Hammerl und Maj näher erläutert werden. Selbstverständlich existieren sprachliche Systeme ebenso wie physikalisch-technische nicht unabhängig von der Zeit, d.h. Veränderungen der Eigenschaften von einzelnen Elementen und von ganzen Systemen finden in der zeitlichen Dimension statt. Darin aber beschränkt sich die Dynamik nicht: Viel aufschlußreicher als die Betrachtung von Variablen zu verschiedenen Zeitpunkten ist die Untersuchung der *funktionalen* Abhängigkeiten der Systemgrößen voneinander. So werden zwar sprachliche Ausdrücke mit der Zeit kürzer oder länger; die Veränderung ihrer Länge ist aber eben keine Funktion der Zeit, sondern eine der Frequenz (deren Veränderung natürlich auch auf der Zeitachse zu beobachten ist). Die Zeit spielt also in funktionaler Hinsicht eine untergeordnete (d.h. keine eigenständige) Rolle. Entsprechend wird der Zusammenhang modelliert: die unabhängige Variable, nach der auch differenziert wird, ist eine andere als die Zeit.

Das diskutierte Basismodell ist also ein dynamisches Modell, weil es Veränderungen im Sprachsystem erfaßt – im Gegensatz etwa zur strukturalistischen Auffassung vom Sprach*system*, die vollkommen statisch ist. Der Standpunkt von Hammerl und Maj dagegen bedeutet eine nicht leicht nachzuvollziehende und vor allem wenig nützliche Reservierung des Terminus *dynamisch* für solche Modelle, die die Veränderung von Systemgrößen auf der Zeitachse beschreiben.

Das Ausklammern der Zeit als untergeordnete Größe ist übrigens keine Besonderheit der Linguistik. Auch das Gravitationsgesetz, um nur ein Beispiel aus der Physik zu nennen, das ein Modell für ein System aus zwei Körpern mit den Massen m_1 und m_2 darstellt und die zwischen diesen Körpern wirkende Kraft mit

$$F = gm_1m_2r^{-2}$$

angibt, verzichtet auf die Zeitdimension ohne im Verdacht zu stehen, bloß statisch zu sein.

Hammerl und Maj schreiben (S. 11) "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 [...] aus". Gerade dies ist aber vielleicht der Hauptirrtum der Autoren! Genau wie das Gravitationsgesetz unterliegen die aus Differentialgleichungen gewonnenen Gesetzeshypothesen, die im Basismodell enthalten sind, *keinen* zeitlichen Gültigkeitsbeschränkungen! Während sich die Systemvariablen in verschiedenen Dimensionen verändern, sind gerade die *Abhängigkeiten* als solche dauerhaft.

2. Ein weiteres Beispiel für den Grundfehler, den Hammerl und Maj begehen, bietet folgende Textpassage (S. 11): "Köhler [...] 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 [...]". Die Befangenheit der Autoren in den Gepflogenheiten der technischen Anwendung von systemtheoretischen Modellen verführte sie neben der Mißinterpretation (von zeitlicher Verzögerung war nur in bezug auf eine Modellvariante die Rede, wo die Länge als auf sich selbst rückgekoppelt darstellbar diskutiert wurde, nicht aber im Zusammenhang mit der Rückwirkung der Länge auf die Frequenz) zu der Behauptung, daß das Modell deshalb keine Möglichkeit zur Selbstregulation besitze.

Tatsächlich aber besteht der Selbstregulationsmechanismus aus dem explizit im Modell ausgedrückten Zusammenwirken von kooperierenden und konkurrierenden Kräften (wie dem Anwendungsbedürfnis, dem Bedürfnis nach Minimierung des Produktionsaufwands, dem Spezifizierungsbedürfnis usw.); die angesprochene Rückwirkung der Länge auf die Frequenz beeinflußt die Selbstregulation nur in ihrer Form. Es handelt sich um die linguistisch begründete Vermutung, daß ein Sprecher – ceteris paribus – unter zwei gleicherweise möglichen (also gleichzeitig existierenden), aber verschieden langen Ausdrücken häufiger den kürzeren wählt (also eine keineswegs *zeitlich verzögerte* Rückkopplung).

Schließlich aber habe ich in der Arbeit von 1986 als Modell für den Zusammenhang zwischen Länge und Frequenz nach Einführung der zusätzlichen Größe *Kürzung* eine Differentialgleichung zweiter Ordnung vorgeschlagen (S. 137–146):

$$L'' + aL' + bL = f(F)$$

die auch die am untersuchten Sprachmaterial entdeckte Oszillation der Länge (in der Frequenz-Dimension - nicht in der Zeit!) erklären kann. Hierzu ist eine gesonderte Untersuchung in Vorbereitung.

3. Einen beträchtlichen Teil ihres Beitrags widmen Hammerl und Maj dem Versuch, ein ihren Vorstellungen von systemtheoretischer Modellierung besser entsprechendes Modell für die beiden Größen Länge und Frequenz aufzustellen - ein Modell, das in erster Linie die Zeitabhängigkeit der beiden Variablen abbildet und auch für rechnerische Simulationsexperimente geeignet ist.

Nahellegenderweise legen sie das aus dem von Altmann verallgemeinerten Piotrovski-Gesetz folgende Zeitverhalten für die Variablen zugrunde. Sie kommen jedoch zu keinem verwertbaren Ergebnis; die Konsequenzen aus den aufgestellten Modellvarianten sind entweder linguistisch unsinnig (die Verminderung der Frequenz ruft eine Verkleinerung der Länge hervor), verdrehen die Abhängigkeitsrichtung oder negieren ihre eigenen Voraussetzungen (denn die Frequenz *ist* die unabhängige Variable, zur Begründung vgl. Köhler 1986).

Wenn wir ein linguistisch interpretierbares Modell des Prozesses aufstellen wollen, das die konkreten Mechanismen und Zeitabläufe wiedergibt, die in einer Sprachgemeinschaft zur Kürzung von Audrücken bei erhöhter Frequenz führen, so können wir nicht einfach den beiden Größen a priori unabhängig je eine Zeitfunktion zuordnen und hoffen, daß uns die Laplace-Transformation Aufschluß über die Selbstregulation gibt. Das Piotrovski-Gesetz als Zeitfunktion ist außerdem das Resultat eines interaktionistischen Modellansatzes, eines Modells für die Verbreitung neuer Formen. Faßt man - wie Hammerl und Maj, indem sie die Gültigkeit dieses Gesetzes voraussetzten - die Längenveränderung als *Verbreitungsvorgang* auf, so ist dies ein Fehler: es handelt sich um die *Entstehung* neuer (kürzerer oder längerer) Formen in dem Maße, wie vorher die entsprechenden alten bereits verbreitet waren. Vor allem aber wird bei dieser Vorgehensweise kein inhaltlicher Bezug zwischen den Größen Länge und Frequenz hergestellt. Wie soll sich ein solcher dann aus dem Modell ergeben? Ein korrektes Modell müßte eben die Abhängigkeit zwischen den beiden Variablen aus einem linguistisch begründeten Mechanismus ableiten, wobei das Zeitverhalten der abhängigen Variablen als Funktion des Zeitverhaltens der unabhängigen aus der Modellstruktur ablesbar würde.

Ich möchte also keineswegs behaupten, daß Untersuchungen des Zeitverhaltens sprachlicher Systeme grundsätzlich unfruchtbar sind -

aber ein Ansatz, der uns zu einer tieferen Einsicht in die Mechanismen sprachlicher Selbstregulation verhelfen soll, muß von linguistischen Vorüberlegungen ausgehen und darf sich nicht auf das Übernehmen einer anderswo bewährten Methode beschränken.

Die Zeit spielt auch in linguistischen Modellen eine zentrale Rolle, nämlich immer dann, wenn der Wert einer Variablen zu einem Zeitpunkt von einem Variablenwert zu einem *anderen Zeitpunkt* abhängt - besonders wenn es sich um *dieselbe* Variable handelt (z.B. im Fall des Piotrovski-Gesetzes). Modelle von Abhängigkeiten mit synchroner Wirkung sind nicht weniger dynamisch als jene; aber die Zeitachse 'protokolliert' hier lediglich ein Geschehen, welches nicht von der Zeit, sondern von anderen Größen bestimmt wird (wie die Gravitation, die nicht mit der Zeit abnimmt, sondern mit der Entfernung). In wieder anderen Fällen muß man die Vernachlässigung der Zeit als Vereinfachung betrachten. Die funktionale Abhängigkeit der Länge von der Frequenz unterliegt notwendigerweise einer (nicht im Basismodell erfaßten) zeitlichen Verzögerung (Ausbreitungsgeschwindigkeit, Wahrnehmungsschwelle etc.). Sollte es einen Grund für die Annahme geben, daß diese mehr darstellt als eine Konstante, und daß sie eine Wirkung im Sprachsystem ausübt, so müßte auch dieser zeitliche Aspekt unter funktionalen Gesichtspunkten analysiert werden.

4. Über die funktionale Abhängigkeit hinaus sollten Zusammenhänge zwischen Größen wie Länge und Frequenz auch im Hinblick auf die lokalen Mechanismen untersucht werden. Das Hauptanliegen des Beitrags von Hammerl und Maj besteht offensichtlich darin, solche Forschungen anzuregen, und darin stimme ich mit ihnen völlig überein. Mir erscheint für diesen Zweck der Ansatz fruchtbar, die Veränderung von Länge und Frequenz als stochastischen Prozeß zu modellieren:

Zu jedem Zeitpunkt t besteht eine Wahrscheinlichkeit dafür, daß zu einer gegebenen lexikalischen Einheit mit der Länge L und der Frequenz F eine neue Variante mit der Länge L' (kürzer oder länger als L) zum Zeitpunkt t+1 entsteht. Diese Wahrscheinlichkeit hängt von F und L zu t ab.

Die Frequenz der neuen Variante ist beim erstmaligen Auftreten natürlich 1; ihre weitere Entwicklung unterliegt einem Geburts- und Todes-Prozeß: die Variante kann sich entweder durchsetzen oder nicht. Die Wahrscheinlichkeit ihrer Ausbreitung steigt mit schon erreichter Frequenz.

In dieser Situation existieren zwei oder mehr verschieden lange Varianten der betrachteten lexikalischen Einheit, die im Hinblick auf ihre Anwendung miteinander in Konkurrenz stehen. Die Anwendungswahr-

scheinlichkeit jeder Variante hängt von der Länge und der zum Zeitpunkt t erreichten Verbreitung (=Frequenz) ab.

Der Mittelwert des Gesamtprozesses sollte mit der Differentialgleichung, wie sie in Köhler (1986) aufgestellt wurde, übereinstimmen.

Eine nach meiner Ansicht auch sehr gut für die Modellierung im Zeitbereich geeignete Schnittstelle des Basismodells bieten die Gleichgewichtsprozesse zwischen den konkurrierenden Bedürfnissen. Wie sehen die Mechanismen, die für die Fließgleichgewichte z.B. zwischen den sich widersprechenden Sprecher- und Höreranforderungen (minD und minK) verantwortlich sind, konkret aus? Welche Mutations- und Selektionsprozesse sind wirksam, und wie sind sie gekoppelt? Wir dürfen allerdings nicht hoffen, entsprechende Modelle leicht empirisch überprüfen zu können, da die Operationalisierungs- und Meßschwierigkeiten gerade hier besonders groß sein dürften.

Literatur

Altmann, G., v.Buttlar, H., Rott, W., Strauß, U. (1983), A law of language change. In: Brainerd, B. [Hrsg.]: Historical Linguistics. Bochum, Brockmeyer 1983, 104–115.

Hammerl, R., Maj, J. (1988), Ein Beitrag zu Köhler's Modell der sprachlichen Selbstregulation. Glottometrika 10, 1–31.

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

Hřebíček, L. (ed.), Glottometrika 11, 1989

Menzerath–Altmann's Law on the Semantic Level

L. Hřebíček, Prague

Words are connected into sentences by grammatical means. The way sentences are linked to form texts is one of the main problems of modern textology.

The present paper starts with a modeling image operating with concepts borrowed from the theory of social communication. The problem arises whether the resulting semantic formants really represent the semantic structure sought. We tried to use the Menzerath–Altmann's law as a criterion for the solution of this question. It will be indicated below that this law holds for certain semantic formants; thus the law achieves the status of a linguistic universal. At the same time, this general validity is used as a background for establishing a new linguistic level and its units.

It is clear that this paper, which proves the validity of the law on two Turkish texts only, is to be regarded as preliminary.

1. Text in Communication

Communication is a transfer of information. The change in the amount of information proper to a system is accompanied by a change in its state. Social communication is communication in human social systems. There exists a large number of works concerning social communication, cf. for example, S.W. King (1975); a set of quantitative characteristics is discussed by L. Hřebíček (1986).

Human beings, the basic units of social systems, are provided with memories that enable them to process communicative stimuli with a time delay. Memory is a constituent of social communication systems; we suppose that it is provided with a semantic structure and with the ability to analyze stimuli. Memory attributes meanings to stimuli of different

sorts. It participates in the reception as well as the emission of communication stimuli.

Text in a natural language is a linguistic unit which has a particular structure and dynamics. We suppose that the structure of a text represents an imprint of the instant (or the approximately instant) state of the speaker's (author's, producer's) memory. A text is a formation of stimuli directed to a hearer's (reader's, recipient's) memory with the intention of supplying it with information and changing its state.

A text is semantically analyzed or interpreted by the recipient's memory as a complex of linguistic signs; for the notion of sign cf. for example W.A. Koch (1974, p. 313).

Linguistic units are signs in communication. They can be tentatively divided into three classes:

1. ***Elementary signs***; words or word forms together with certain morphemes added to them.
2. ***Sentences***; these signs consist of elementary signs connected into sentences by grammatical means.
3. ***Sign aggregations***; with the help of semantically identical lexical units and/or with the help of text references, sentences are connected into sign aggregations.

Semantically identical lexical units occurring in one and the same text are elementary signs with an identical denotatum.

Text references are formed by different lexical units or morphemes denotating identical objects or events. Text references can be understood as semantic equivalences. Our treatment of text references agrees with their description in M.A.K. Halliday and R. Hasan (1976) or in B. Palek (1988); with respect to their dynamics, cf. L. Hřebíček (1985) and G. Altmann (1988: 81-85).

Both of the means mentioned, identical lexical units with identical denotata and text references, are sub-sets of the set of ememtary signs.

2. Semantic Structure of a Text

Each sign aggregation consists of sentences in which semantic identities or equivalences occur. Thus a text consists of several sets of sentences;

in each sentence of one and the same aggregation, a unit having one and the same meaning occurs, regardless of whether it is expressed by an identical lexical unit or morpheme, or by a text reference.

A text must be interpreted by a human recipient; he uses his memory and states identities and equivalences on the level of elementary signs and thus finds the sign aggregations. This process depends on the ability of the individual memory to find identities and equivalences. Consequently, there are variations in different interpretations of one and the same text. Linguistic competence is the concept on which the possibility of ascertaining mutually similar or even identical semantic structures in one and the same text by different recipients can be based.

Now we want to know whether the semantic structure described, consisting of elementary signs, sentences and aggregations is a real linguistic structure or only a product of the wishful thinking of one linguist.

In our opinion, Menzerath-Altmann's law can serve as such a criterion. This law was formulated by P. Menzerath (1928) and (1954), and its mathematical form discovered by G. Altmann (1980) and corroborated on the length of morphemes and words, measured in the number of syllables, and on the length of syllables as their constituents.

The short expression of the law is as follows:

> "The longer a language construct - the shorter its components (constituents)."

In accordance with this expression, Altmann set up the following differential equation:

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

x - construct,
y - component (constituent),
b,c - constants.

He derives three variants of its solution, from which the following one is used in the present paper:

$$y = Ax^{b},$$

where A is a coefficient.

Our original intention was to indicate that Menzerath-Altmann's law cannot be valid on the text level. Intuitively, a text consists of sentences and it would be ridiculous for example to assume that Thomas Mann wrote shorter sentences in his ***Josef und seine Brüder*** than in ***Der Erwählte***. The result of our experiment turned out to be surprising: a text does not consist immediately of sentences.

3. Analyses of Two Texts

In this preliminary paper we present results obtained from investigating two Turkish texts.

The first one is the chapter ***Halka Doğru*** from the book by Ziya Gökalp (1970, pp. 46-61). This text challenges Turkish intellectuals to go to villages; two reasons are presented: to bring European civilization to the villagers and to learn the Turkish national culture there which, at the time of the composition of the text, was supposed to replace the Ottoman culture based on Islamic civilization.

This text contains 105 sentences and its total length is 915 words. We determined 82 sign aggregations, of which we present the longest ones: (x indicates the number of sentences and $y = n_x/x$ is the mean sentence length of the aggregation in the number of words.)

halk "people" (y = 396/45);
seçkin "member of élite" (y = 264/31);
millî (and derivated lexemes) "national" (y = 265/23);
Türk (and derivatives) "Turk" (y = 209/22);
Osmanlı (and derivatives) "Ottoman" (y = 147/15);
kültür "culture" (y = 128/13);
biz "we" (in the sense of "our community") (y = 155/11);
terbiye "education" (and similar notions) (y = 67/9).

As a short example of how the aggregations were obtained we present a meticulous translation of the first four sentences of the text:

1. "One of the first foundations of Turkism is the principle 'Toward People'.
2. Once, in order to apply this principle, we published a review in Istanbul under the name 'Toward People'.
3. After that, in Izmir, a review of the same name was published.
4. What does 'to go to the people' mean?"

The semantic identities and equivalences of this fragment are represented by the units halk "people", ad - isim "name", esas - prensip "principle". Sentences 1., 2., and 4.belong to the aggregation 'halk', the first sentences of aggregation 'esas - prensip' are 1., 2., and the first sentences of aggregation 'ad - isim 'are sentences 2. and 3.

If in one sentence the unit on which a given aggregation is based occurs more than once, its analysis is the same as if it has occurred only once.

From the hypothesis formulated by Menzerath, it follows that there is an inverse correlation between x (the length of aggregation measured in the number of sentences) and y (the sentence length computed as the mean sentence length of the observed values belonging to the aggregation).

In the text analyzed, from its 82 pairs of values of x and y, the observed Spearman's rank correlation coefficient $R = -0.2198$ with the degrees of freedom $f = n - 2 = 80$ and $t = -2.0148 < -t_{0.05} = -2.0003$ has been ascertained; this coefficient was computed with corrections for the identical rank numbers. Consequently, the coefficient is significant and it has the expected negative value.

In Table 1 the observed values of x and y are presented, together with the computed y_c. The significance of the difference between the variances of y and y_c was tested by F-test.

As $F = 4.975 > F_{0.025}(13,13)$, $H_0: s_1^2 = s_2^2$ must be rejected.

The means are : $y = 11.001$ and $y_c = 10.841$;

the testing criterion is:

$$t = 0.244 < t_{0.05}(13) = 2.159.$$

There is no reason to reject

$$H_0: m_1 = m_2.$$

In this way only two parameters of the variables were tested. However, we want to prove the correspondence of the variables in a broader sense. We want to analyze their total distributions in order to know whether the theoretical sample represented by y_c can serve as a model for the observed y. For this purpose the Wilcoxon test for paired values was used. This test is based on differences $d = y - y_c$ and their rank values. The observed value of the test criterion is

$$T = 50 > T_{0.05}(14) = 21;$$

there is no reason to reject the hypothesis that the two distributions do not differ significantly.

Table 1
Observed and computed values from the text by Z. Gökalp

x	g	w	y=w/(xg)	$y_c=Ax^b$	$d=y-y_c$	Rank of d
2	48	992	10.33	12.66	-2.33	11
3	10	355	11.83	12.17	-0.34	1
4	6	343	14.29	11.84	2.45	12
5	4	274	13.70	11.58	2.12	9
6	4	283	11.79	11.38	0.41	3
9	1	67	7.44	10.95	-3.51	14
11	1	155	14.09	10.74	3.35	13
13	1	128	9.85	10.56	-0.71	6
15	1	147	9.80	10.42	-0.62	5
16	2	402	12.56	10.35	2.21	10
22	1	209	9.50	10.04	-0.54	4
23	1	265	11.52	10.00	1.52	8
31	1	264	8.52	9.71	-1.19	7
45	1	396	8.80	9.37	-0.37	2

A = 13.5320; b = -0.0965; T = 50
x - length of aggregation (in number of sentences)
g - number of aggregations
w - the sum of words in g aggregations
y - mean sentence length (in number of words)

Let us support this result with the following two imaginary experiments: The sentence length of the analyzed text is always between 1 and 40 words. For each observed aggregation each observed sentence length was substituted for by a value not greater than 40, taken from the table of random sampling numbers. We obtained values of y which only slightly fluctuate around the mean (= 20). It is evident that they are not correlated with x.

105 cards with the numbers corresponding to the length of the observed sentence were put into a polling urn. Sets of cards were drawn from the urn, each set corresponding to the observed length of the aggregation (in the number of sentences); of course, each drawing of cards began with the full number of 105 cards. For each "aggregation" a value was obtained which only slightly differs from the mean sentence length of the entire text (= 915/105 = 8.71). Again, no correlation between x and y was ascertained.

These are of course very simple experiments. They can only confirm the basic principles of mathematical statistics. Nevertheless, they vividly indicate that the sentence length and the arrangements of sentences into sign aggregations follow Menzerath-Altmann's law.

The second text analyzed is a poem by Yunus Emre, a Turkish mystical poet of the 13th century. This poem was published by A. Gölpinarli (1965, pp. 81-82). It begins with the verse:

Ata belinden bir zaman anasına düşdi gönül
Hak'dan bize destur oldı hazineye düşdi gönül

"From Father's waist once to his mother heart fell,
from the God to us was a permission, to the treasure heart fell."

The number of references of the type of poetic images in poetry is usually higher than in prose. They occur in the verse quoted: ata "father" - Hak "God", ana "mother" - hazine "treasure", gönül "heart" - destur "permission", etc. The reader's competence to ascertain all such relations between or among elementary signs requires training in reading such texts.

The poem written in an isosyllabic metre and having a sentence structure correlated with this metre was selected with the intention of indicating that there are texts for which Menzerath-Altmann's law does not hold. However, we found that the contrary was the case.

The correlation between x and y is expressed by Spearman's R = -0.326 which is significant with n = 28 on α = 0.05. Table 2 contains values analogous to those in Table 1.

Table 2
Observed and computed variables from the text by Yunus Emre

x	g	w	y=w/(xg)	$y_c = Ax^b$	d=y-y_c	Rank of d	
2	17	121	3.56	3.35	0.21	2	
3	5	46	3.07	3.29	-0.22		3
6	1	25	4.17	3.18	0.99	7	
7	1	17	2.43	3.15	-0.72		6
8	2	46	2.88	3.13	-0.25		4
9	1	27	3.00	3.12	-0.12		1
14	1	48	3.43	3.05	0.38	5	

A = 3.4640 b = -0.0483 T = 14

The results of testing y and y_c are similar to those obtained from the preceding text:

$F = 28.863 > F_{0.025}(6, 6) = 5.82$; $H_0: s_1^2 = s_2^2$ is to be rejected.

$t = 0.359 < t_{0.025}(6) = 2.447$; there is no reason to reject $H_0: m_1 = m_2$.

The observed values are

$$\bar{y} = 3.22 \text{ and } \bar{y}_c = 3.18.$$

Wilcoxon test: $T = 14 > T_{0.05}(7) = 2$;

there is no reason to reject the hypothesis on the grounds that there is no significant difference between the distributions of y and y_c.

4. Conclusions

We cannot expect the entire variation of y to be explained by Menzerath-Altmann's law; this is due to the fact that y is affected by random factors encompassed in interpretations, i.e. in sign identities and sign equivalences. Thus we obtained the above indicated results of F-tests. Nevertheless, the law affects the relation between x and y to the degree indicated by correlation coefficients. Only a part of the variation of y can be explained by the law. This is sufficient to allow us to some to the conclusion that the semantic structure of a text is in accordance with Menzerath-Altmann's law.

A text appears to be a unit or construct consisting not of sentences, but of sign aggregations. At the same time, Menzerath-Altmann's law can serve as a criterion for distinguishing between levels. By all accounts, this law can be taken as a principle of stability for linguistic formants and it deserves a deeper investigation.

Let us add that the observed phenomenon seems to be related to the hypothesis of polylexy formulated by R. Köhler (1986: 100) and to the principle of the information threshold.

References

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

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

Gökalp, Z. (1970), Türkçülüğün esasları. Istanbul: Millî Eğitim.

Gölpinarli, A. (1965), Yunus Emre, Risâlat al-nushiyya ve dîvân. Istanbul: Sulhi Garan.

Halliday, M.A.K., Hasan, R. (1976), Cohesion in English, London: Longman.

Hřebíček, L. (1985), Text as a unit and co-references. In: T. Ballmer (ed.), Linguistic dynamics. Berlin - New York: de Gruyter, 190-198.

Hřebíček, L. (1986), Quantities of social communication. Prague: Oriental Institute.

Kind, S.W. (1975), Communication and social influence. Reading (Mass.): Addison-Wesley.

Koch, W.A. (1974), Semiotik und Sprachgenese. In: W.A. Koch (ed.), Perspektiven der Linguistik II. Stuttgart: Kröner: 312-346.

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

Menzerath, P. (1928), Über einige phonetische Probleme. Actes du premier congrès international de linguistes. Leiden: Sijthhoff: 104–105;

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

Palek, B. (1988), Referenční výstavba textu (The structure of reference in a text). Praha: Univerzita Karlova.

Hřebíček, L. (ed.), Glottometrika 11, 1989

On Dixon's model of lexical diffusion in Australia

F. Schweiger, Salzburg

Abstract: Dixon has proposed a mathematical model for lexical diffusion within Australian languages. A fairly rigorous mathematical treatment for the associated non-linear dynamical system is given.

1. Introduction

In the book "The Dyirbal Language of North Queensland" (which has become famous especially due to the treatment of ergativity) Dixon discusses the finding that two or more languages which have been contiguous for a long enough time have about 50% vocabulary in common. In his own words (loc.cit. p.331/332): "That is, if two dialects move into contiguity and, at the beginning have no (or very little) vocabulary in common – through borrowing from each other to replace proscribed items – the percentage of common vocabulary will build up until it levels off at about 50%. On the other hand, if a tribe splits into two and the two new tribes remain in contiguity, then they will at first have almost identical vocabularies; as different words become taboo at different times in the two sister dialects, and are replaced from neighbouring dialects, the percentage of common vocabulary will gradually decrease until it levels off at about 50%." Dixon also proposes a mathematical model for the cases of 3 and even more contiguous languages which in principle gives some explanation for the value 50%.

In this paper we will give a rigorous mathematical treatment of Dixon's model for 3 contiguous languages. it is shown that in fact (1/2, 1/2, 1/2) is the unique fixed point of a non-linear dynamical system which will be constructed in section 2. In section 3 its fixed points are calculated. In section 4 we discuss the domain of attraction for the relevant fixed point.

2. The Dixon model

We suppose that there are three languages A, B and C which are contiguous to each other. Let $p(X,Y;t)$ denote the fraction of vocabulary shared by languages X and Y at (discrete) time t. Clearly

$$p(X,Y;t) = p(Y,X;t).$$

We assume that there is coefficient α, $0 < \alpha < 1$ which remains unchanged over the time span considered and can be interpreted as a parameter of replacement of vocabulary. To obtain a dynamical system on the 3-dimensional unit cube we will impose the restriction $\alpha < 1/2$. Later on we also need $\alpha < 3/7$. This seems to be very natural, since clearly α is considered a small number. Within a time interval of length 1 language B will lose $\alpha p(B,C;t)$ of the vocabulary it has in common with C, and C will lose an identical amount. Thus the total loss is $-2\alpha p(B,C;t)$. Here we assume that the loss works independently in both languages (clearly the loss may be caused by several reasons not only - as Dixon assumes - by tabooing).

On the other hand B will borrow some new vocabulary from C. Now C has $1 - p(B,C;t)$ of its vocabulary different from B, and $1 - p(C,A;t)$ different from A. If B borrows some vocabulary from C which C has in common with A we cannot distinguish such a borrowing from a borrowing by B directly from A. Therefore we will take into account only the amount of vocabulary which B borrows from C which is not common with A. Therefore we can expect the gain by B borrowed from C as

$$\alpha \frac{1 - p(B,C;t)}{1 - p(B,C;t) + 1 - p(C,A;t)}$$

Similary the gain by C borrowed from B will be

$$\alpha \frac{1 - p(B,C;t)}{1 - p(B,C;t) + 1 - p(B,A;t)}.$$

Hence

$$p(B,C;t+1) = p(B,C;t) - 2\alpha p(B,C;t) + \alpha \frac{1 - p(B,C;t)}{2 - p(B,C;t) - p(C,A;t)}$$

$$+ \alpha \frac{1 - p(B,C;t)}{2 - p(B,C;t) - p(B,A;t)}.$$

To simplify notation we now put

$$p(A,B;t) = p, \quad p(B,C;t) = q, \quad p(C,A;t) = s$$

$$p(A,B;t+1) = P, \quad p(B,C;t+1) = Q, \quad p(C,A;t+1) = S.$$

Then the replacement after a time interval of lenght 1 leads to the equations

$$P = p + \alpha\left(\frac{1 - p}{2 - p - q} + \frac{1 - p}{2 - p - s} - 2p\right)$$

$$Q = q + \alpha\left(\frac{1 - q}{2 - q - s} + \frac{1 - q}{2 - q - p} - 2q\right)$$

$$S = s + \alpha\left(\frac{1 - s}{2 - s - p} + \frac{1 - s}{2 - s - q} - 2s\right).$$

This is a non-linear dynamical system with evolution map

$$T(p,q,s) = (P,Q,S).$$

Clearly

$$-2p \le \frac{1 - p}{2 - p - q} + \frac{1 - p}{2 - p - s} - 2p \le 2 - 2p.$$

Therefore

$$p + \alpha(-2p) \le P \le p + \alpha(2 - 2p)$$

The condition $0 \le p \le 1$ gives $0 \le P \le 1$ as long as $\alpha < 1/2$. From this we get

Theorem 1: If $\alpha < 1/2$ the map T as defined above maps the 3-dimensional unit cube $0 \le p \le 1$, $0 \le q \le 1$, $0 \le s \le 1$ into itself.

3. Calculation of the fixed points

It is easily seen that the point (1/2, 1/2, 1/2) is in fact a fixed point of our system. This result does not depend on the value of α.

Theorem 2: For $0 < \alpha < 1$ the point (1/2, 1/2, 1/2) is the unique fixed point in the unit cube.

Proof: The equation

$$T(p,q,s) = (p,q,s)$$

leads to

$$\frac{1-p}{2-p-q} + \frac{1-p}{2-p-s} = 2p$$

$$\frac{1-q}{2-q-s} + \frac{1-q}{2-q-p} = 2q$$

$$\frac{1-s}{2-s-p} + \frac{1-s}{2-s-q} = 2s.$$

Addition of these equations leads to

$$p + q + s = \frac{3}{2}.$$

We substitute $s = 3/2 - p - q$ into the first two equations and get

$$\frac{1-p}{2-p-q} + \frac{2-2p}{1+2q} - 2p = 0$$

$$\frac{2-2q}{1+2p} + \frac{1-q}{2-p-q} - 2q = 0.$$

This leads to

$$5 - 11p - 6pq + 4p^2 + 4p^2q + 4pq^2 = 0$$

$$5 - 11q - 6pq + 4q^2 + 4pq^2 + 4p^2q = 0.$$

Subtraction of these equations gives

$$11q - 11p + 4p^2 - 4q^2 = 0.$$

This is equivalent to

$$p - q = 0$$

respectively

$$11 - 4p - 4q = 0.$$

Substitution of $p = q$ into one of the former equations gives

$$8p^3 - 2p^2 - 11p + 5 = 0.$$

This cubic equation has the roots

$$p = \frac{1}{2}, -\frac{5}{4}, 1.$$

From this we obtain the fixed points

$$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$$

$$(-\frac{5}{4}, -\frac{5}{4}, 4).$$

If (p,q,s) is a fixed point clearly any permutation of the coordinates also gives a fixed point. Therefore we additionally find the fixed points:

$$(-\frac{5}{4}, 4, -\frac{5}{4})$$

$$(4, -\frac{5}{4}, -\frac{5}{4}).$$

These two fixed points also correspond to the case $11 - 4p - 4q = 0$. The case $p = q = 1$ does not correspond to a fixed point (note that in this case $2 - p - q = 0$).

3. The nature of the fixed point

We go back to the equations of evolution:

$$P = p + \alpha\left(\frac{1 - p}{2 - p - q} + \frac{1 - p}{2 - p - s} - 2p\right)$$

$$Q = q + \alpha\left(\frac{1 - q}{2 - q - s} + \frac{1 - q}{2 - q - p} - 2q\right)$$

$$S = s + \alpha\left(\frac{1 - s}{2 - s - p} + \frac{1 - s}{2 - s - q} - 2s\right).$$

Then a calculation show that the Jacobian matrix of the fixed point (1/2,1/2,1/2) is given as

$$\begin{bmatrix} 1 - 3\alpha & \frac{\alpha}{2} & \frac{\alpha}{2} \\ \frac{\alpha}{2} & 1 - 3\alpha & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \frac{\alpha}{2} & 1 - 3\alpha \end{bmatrix}$$

The eigenvalues are

$$\lambda = 1 - \frac{7\alpha}{2}$$

which is a double root, and

$$\mu = 1 - 2\alpha.$$

Since $0 < \alpha < 1$ clearly $|\mu| < 1$. To secure $|\lambda| < 1$ one has to suppose that $\alpha < 4/7$. Since we already suppose $\alpha < 1/2$ this is no restriction. So we get

Theorem 3: If $\alpha < 4/7$ the unique fixed point (1/2,1/2,1/2) in the unit cube is attractive.

Remarks: (1) The vector (1,1,1) is an eigenvector for $\mu = 1 - 2\alpha$. The line joining (0,0,0) and (1,1,1) is invariant under the action of T. In fact all points (p,q,s) which satisfy p = q = s are strongly attracted to the fixed point. Put p = q = s then

$$P = p + \alpha(1 - 2p) = p(1 - 2\alpha) + \alpha.$$

The recursion

$$p_{n+1} = p_n(1 - 2\alpha) + \alpha$$

has the explicit solution

$$p_n = \frac{1}{2} + (p_0 - \frac{1}{2})(1 - 2\alpha)^n.$$

Let us recall the following important special cases

(1.1) $p_0 = q_0 = s_0 = 0.$

This means that three languages with totally different vocabulary become contiguous at time t = 0 and replace their vocabulary according to the present model.

(1.2) $p_0 = q_0 = s_0 = 1.$

This means that one language splits into three dialects at time t = 0. These three dialects undergo some independent development but also exchange their vocabulary.

In both cases we end at p = q = s = 1/2.

(2) The eigenvectors belonging to $\lambda = 1 - (7\alpha/2)$ span the plane p + q + s = 3/2 which is orthogonal to (1,1,1) and contains the point (1/2,1/2,1/2).

Strangely enough also this plane is invariant under the action of T. Addition of the defining equations show

$$P + Q + S = (p + q + s)(1 - 2\alpha) + 3\alpha.$$

Therefore p + q + s = 3/2 is a fixed point of this iteration. Furthermore it shows that the plane p + q + s = 3/2 attracts any point in the 3-dimensional space.

4. The domain of attraction

Since T is a continuously differentiable map there is a neighbourhood N of (1/2,1/2,1/2) with the property that

$$\lim_{n\to\infty} T^n(p,q,s) = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2})$$

for any point (p,q,s) ε N. If we can show that for any point (p,q,s) lying in the unit cube there is a number m = m(p,q,s) such that

$$T^m(p,q,s) \;\varepsilon\; N$$

then clearly

$$\lim_{n\to\infty} T^n(p,q,s) = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}).$$

Since the plane p + q + s = 3/2 is strongly attractive any accumulation point of the sequence $(T^n(p,q,s))$, n = 1,2,..., lies in this plane. Let us suppose that we have already shown that any point of the intersection of the plane p + q + s = 3/2 with the unit cube is attracted by the fixed point (1/2,1/2,1/2). Let (p^*,q^*,s^*) be an accumulation point of the orbit $(T^n(p,q,s))$, n = 1,2,.... Then there is a subsequence (n(j)), j = 1,2,... such that

$$\lim_{j\to\infty} T^{n(j)}(p,q,s) = (p^*,q^*,s^*).$$

By continuity of T we obtain

$$\lim_{j\to\infty} T^{n(j)+k}(p,q,s) = T^k(p^*,q^*,s^*) \;\varepsilon\; N$$

for sufficiently large k.

We introduce the new coordinates

$$x = \tfrac{2}{3}p - \tfrac{1}{3}q - \tfrac{1}{3}s$$

$$y = -\tfrac{1}{3}p + \tfrac{2}{3}q - \tfrac{1}{3}s$$

$$z = \tfrac{1}{3}p + \tfrac{1}{3}q + \tfrac{1}{3}s - \tfrac{1}{2}.$$

This transforms the plane p + q + s = 3/2 into the plane z = 0. Then a calculation shows that the reduced dynamical system is given by the equations:

$$X = x(1 - 2\alpha) + \frac{\alpha}{2}\left(\frac{-x + y}{1 - x - y} - \frac{2x + y}{1 + y}\right)$$

$$Y = y(1 - 2\alpha) + \frac{\alpha}{2}\left(\frac{x - y}{1 - x - y} - \frac{x + 2y}{1 + x}\right)$$

or equivalently

$$X = x + \alpha\frac{-7x + 2y^2 + 6x^2 + 2xy + 4x^2y + 4xy^2}{2(1 - x - y)(1 + y)}$$

$$Y = y + \alpha\frac{-7y + 2x^2 + 6y^2 + 2xy + 4x^2y + 4xy^2}{2(1 - x - y)(1 + x)}.$$

The intersection of the plane p + q + s = 3/2 with the unit cube transforms into the hexagon H with vertices

$$(\tfrac{1}{2}|0), (0|\tfrac{1}{2}), (-\tfrac{1}{2}|\tfrac{1}{2}), (-\tfrac{1}{2}|0), (0|-\tfrac{1}{2}), (\tfrac{1}{2}|-\tfrac{1}{2}).$$

The fixed points are given by

$$(0|0), (-\tfrac{7}{4}|\tfrac{7}{2}), (-\tfrac{7}{4}|-\tfrac{7}{4}), (\tfrac{7}{2}|-\tfrac{7}{4}).$$

Theorem 4: If $\alpha < 3/7$ then the unit cube is contained in the domain of attraction of the fixed point (1/2,1/2,1/2) for the given 3-dimensional dynamical system.

Proof: It is sufficient to show that the hexagon H as described above lies in the domain of attraction of the fixed point (0|0) for the reduced 2-dimensional system. A calculation shows that

$$X - Y = (x - y)\left(1 + \alpha\left(-2 - \frac{1}{1 - x - y} - \frac{\frac{1}{2} + x + y}{(1 + x)(1 + y)}\right)\right).$$

Clearly

$$\frac{1}{1 - x - y} \leq 2 \quad \text{for} \quad (x,y) \in H.$$

Since

$$(1 + x)(1 + y) = (1 + \frac{x + y}{2})^2 - (\frac{x - y}{2})^2$$

one sees that the function

$$\frac{\frac{1}{2} + x + y}{(1 + x)(1 + y)}$$

attains its maximum on the boundary of H. Then some calculations show that

$$\frac{\frac{1}{2} + x + y}{(1 + x)(1 + y)} \leq \frac{2}{3} .$$

Therefore

$$1 - \alpha\frac{14}{3} \leq 1 + \alpha(-2 - \frac{1}{1 - x - y} - \frac{\frac{1}{2} + x + y}{(1 + x)(1 + y)}) < 1.$$

Note that both maxima are attained at $(1/2|0)$ and $(0|1/2)$.

For $\alpha < 3/7$ we obtain

$$-1 < 1 - \alpha\frac{14}{3} < 0.$$

In this case the intersection of the line $x = y$ with the hexagonal domain H is attractive.

But for $x = y$ we obtain

$$X = x - \alpha\frac{x(4x + 7)}{2(x + 1)} = x(1 - 2\alpha) - \frac{\alpha}{2}\frac{3x}{1 + x}.$$

We have to consider this 1-dimensional dynamical system on the interval $[-1/4,1/4]$. A calculation shows that (for $\alpha < 3/7$):

$$X(-\frac{1}{4}) = -\frac{1}{4} + \alpha < \frac{1}{4}$$

$$|X'(-\frac{1}{4})| = |1 - \frac{14}{3}\alpha| < 1$$

$$X(\frac{1}{4}) = \frac{1}{4} - \frac{4}{5}\alpha > -\frac{1}{4}$$

$$|X'(\frac{1}{4})| = |1 - \frac{74}{25}\alpha| < 1 .$$

Since $x = 0$ is the unique fixed point with slope $1 - 7/2\alpha$ in the interval $[-1/4,1/4]$ this fixed point is attractive for the whole interval. Note that $X = X(x)$ has a unique minimum for $x > -1$ and no point of inflection.

Now we have shown that a point $(x|x)$, $-1/4 \leq x \leq 1/4$ satisfies

$$\lim_{n\to\infty} T^n(x|x) = (0|0) .$$

Since any point $(x|y) \in H$ is attracted to this segment a similar reasoning as outlined at the beginning of this section shows that in fact

$$\lim_{n\to\infty} T^n(x|y) = (0|0).$$

Remarks: (1) Numerical considerations suggest that theorem 4 is true for any $\alpha < 1/2$. Note that the parts of the boundary of the hexagon H which are formed by the lines $x = 1/2$ resp. $y = 1/2$ are mapped onto parts of the lines $X = 1/2 - 2\alpha$ and $Y = 1/2 - 2\alpha$ respectively. Therefore if $1/4 < \alpha < 1/2$ clearly $X < 0$ resp. $Y < 0$ which means that these segments are mapped into the "other side" of the plane.

(2) The 2-dimensional system has lines of discontinuity, namely the lines given by the equations

$$1 - x - y = 0$$
$$1 + x = 0$$
$$1 + y = 0 .$$

These lines form a triangle with vertices $(-1|-1)$, $(-1|2)$, $(2|-1)$ which contains the hexagon H. Therefore one expects the existence of an "invariant circle" (a closed curve which is invariant under the action of the system) which surrounds the hexagon H.

References

Dixon, R.M.W. (1972). ***The Dyirbal Language of North Queensland.*** Cambridge at the University Press.

Hřebíček, L. (ed.), Glottometrika 11, 1989

Gliederung einer Sprachfamilie (hier der Romania) mit Hilfe eines numerischen Kalküls

J. Kristophson, Bochum

0. Allgemeine Vorbemerkungen

0.1. Dieser Beitrag soll einen Versuch zur Gliederung von Sprachfamilien mit Hilfe eines numerischen Kalküls darstellen. Ihm liegt die Annahme zu Grunde, daß es Beziehungen zwischen Sprachen gibt, die mit dem Terminus "Sprachfamilie" benannt werden können. Dieser Terminus, eine Metapher aus dem menschlichen Leben, besagt, daß zu einem Zeitpunkt t_2 existierende Sprachen $L_1, L_2, \ldots, L_n$ (es müssen mindestens zwei sein) auf eine zu einem früheren Zeitpunkt t_1 existierende Sprache L zurückzuführen sind. L selbst wird zum Zeitpunkt t_2 nicht mehr gesprochen, es "lebt" in L_1, L_2 weiter, wenn diese Metapher gestattet ist. Eine Sprachfamilie gilt als akzeptiert, wenn sich zwischen $L_1, L_2, \ldots, L_n$ und L Korrespondenzregeln finden lassen, die Ausdrücke (Laute) und Inhalte (Morphologie und Semantik) miteinander verbinden. Ein Beispiel aus zwei slavischen Sprachen soll dies verdeutlichen:

Russ. 'gorodov', serb. 'gradova' (beide gen.pl. = der Städte) sind verbindbar, weil

1. 'g...d'phonetisch und distributionell identisch oder sehr ähnlich

2. 'oro/ra' distributionell identisch, phonetisch ähnlich, auf eine dritte gemeinsame Vorform zurückführbar

3. 'ov' distributionell identisch, phonetisch ähnlich, rückführbar auf ein Element einer alten Deklinationsklasse, aber unterschiedliche Funktion, d.h. hier liegt funktionelle Trennung vor

4. Schaffung einer neuen, aber verschiedenen gen.pl.-Endung in beiden Sprachen

5. Bewahrung der Kategorien 'gen.'und 'pl.'in beiden Sprachen

6. gemeinsame Bedeutungsverschiebung von 'Burg'zu 'Stadt'.

Es liegen also zahlreiche, korrespondierende Gemeinsamkeiten zwischen den beiden gewählten Sprachen vor, die eine Verbindung zu einer Vorform, hier zum Urslavischen, herstellen lassen.

Es soll hier noch auf eine geniale, aber auch gefährliche Vereinfachung hingewiesen werden. Der Zusammenhang zwischen den Beispielsprachen wurde durch Buchstabenfolgen dargestellt, die größere Ähnlichkeiten vortäuschen, als die sprachliche Wirklichkeit sie bietet. Sprecher kommunizieren eben primär durch Schallwellen. Dieser Einwand soll aber nicht entmutigen, so ziemlich die gesamte Linguistik operiert mit Buchstabenfolgen, auch wenn die Buchstaben vom Standardalphabet oder von der Standardorthographie abweichen. Das Genialische an den Buchstaben besteht darin, daß sie eine endliche Anzahl distinkter Größen darstellen. Sie eignen sich daher als Datenmenge für die verschiedensten Untersuchungsziele und -methoden. Wenn also auch nur Buchstabenphilologie getrieben wird, so repräsentieren doch Buchstaben sprachliche Zustände und Prozesse und können sich letztlich immer lautlichen Erscheinungen annähern.

0.2. Anlaß für die folgenden Gedanken ist eine gewisse Kritik an Muljačić (1967), dessen Vorgehen aber als erweiterungsfähig und -würdig anzusehen ist. Konkretes Ziel ist es, die Stellung des Rumänischen innerhalb der romanischen Sprachen zu bestimmen. Die romanischen Sprachen bilden eine Sprachfamilie, wahrscheinlich sogar den Idealfall von Sprachfamilie, deren Ausgangspunkt, Zwischenstufen und Endstufen bekannt oder wenigstens wegen der Materialfülle gut nachzeichenbar sind. Aber gerade die Materialfülle macht es wiederum schwierig, Nähe und Ferne der einzelnen Sprachen abzuwägen und richtig zu beurteilen. Diese Fragestellung hat in der Romanistik zu einer unfangreichen Diskussion und zu unterschiedlichen Lösungsversuchen geführt (vgl. dazu Muljačić 1967, Iliescu 1969). Wie aber der Beitrag von Muljačić (1967) zeigt, lassen sich noch neue Argumente für das Abwägen und Beurteilen der Zusammenhänge innerhalb der Romania, letztlich aber auch in anderen Sprachfamilien, fin-

den. Die Grundfrage lautet also: "Welche romanische Sprache ist am meisten mit dem Rumänischen verwandt?" Die Antwort kann aber nicht nur der Name einer anderen romanischen Sprache sein, sondern sie muß auch einen Grad der Verwandtschaft angeben.

1. Linguistisches Konzept

Als theoretisch linguistisches Konzept für das weitere Procedere (vgl. Kristophson 1984) liegt die Annahme zu Grunde, daß Sprachverwandtschaft bzw. Sprachfamilie als Dreiecksverhältnis von gemeinsamer Ausgangssprache zu den verschiedenen Nachfolgersprachen und von den Nachfolgersprachen zueinander zu erfassen sei. Ausgangssprache und Nachfolgersprache unterscheiden sich dadurch, daß eine gewisse Anzahl von Elementen der Ausgangssprache sich verändert hat, eine gewisse Anzahl dagegen nicht. Die Nachfolgersprachen unterscheiden sich untereinander dadurch, daß für gleiche Elemente der Ausgangssprache eine Streuung von Erhaltung und Veränderung zu beobachten ist und daß bei eingetretenen Veränderungen ebenso eine Streuung von unterschiedlichen oder gemeinsamen Endergebnissen sich ergeben haben kann.

Beispiel: Die romanischen Sprachen unterscheiden sich vom Latein durch Einführung eines Artikels, also durch eine Veränderung in allen Sprachen. Die romanischen Sprachen unterscheiden sich voneinander durch Wahl verschiedener Elemente für den Artikel oder durch die Stellung des Artikels.

Die Entscheidung für Bewahrung oder Veränderung eines Elements ist a priori als Zufallsentscheidung zu bewerten und läßt sich durchaus als ja/nein-Frage formulieren. Die Veränderung eines Elements in zwei Nachfolgersprachen in gemeinsamer Richtung kann auch ein Zufall sein, wahrscheinlicher ist aber eher ein Zusammenhang, z.B. durch Sprachkontakt, kulturelle Beeinflussung, u.ä. Da die Zahl der Veränderungen hoch ist, sollten auch zahlreiche berücksichtigt werden, was fast zwangsläufig Metrisierung, Indexbildung u.a. zur Folge hat.

Als allgemeines Modell zur Darstellung der Prozesse, die für die Aufgliederung einer Sprachfamilie eine Rolle spielen, kann ein so zu nennendes Explosionsmodell dienen. Beim Explosionsmodell wird ein Ballon aufgeblasen, der schließlich platzt und dessen Teile sich in verschiedene Richtungen fortbewegen, einzelne in die gleiche, sogar mit möglichen Berührungen. Eine unterschiedliche Fluggeschwindigkeit und unterschiedliche Flugweiten sich ebenso einzukalkulieren. Dieses Modell erfaßt recht gut den Tatbestand des Lateinisch/Romanischen mit den Stadien: Rom, Latium, Imperium, Provinzen, neuromanische Sprachen und Dialekte. Die unterschiedlichen, aber empirisch faßbaren "Flugwege" von jeweils zwei Partikeln des Ballons im Explosionsmodell lassen sich als Ratio gemeinsamer Neuerungen, unterschiedlicher Neuerungen, Neuerungen gegen Archaismen, gemeinsamer Archaismen (hier die üblichen linguistischen Termini statt Veränderung und Bewahrung) formulieren, berechnen, graphisch darstellen und dann wieder linguistisch interpretieren.

2. Daten

Als Daten bieten sich die von Muljačić (1967) gesammelten 40 Merkmale für 12 Sprachen an. Die Merkmale sind beantwortete Fragen an die jeweiligen Sprachen. Eine Frage bedeutet eine Auswahl zwischen zwei genannten sprachlichen Eigenschaften, von denen eine gewählt werden muß. Die Antwort ist eine getroffene Wahl, die durch + oder - dargestellt wird, wobei + die erste Wahlmöglichkeit, - die zweite bedeutet. Dieses Verfahren soll grundsätzlich beibehalten werden, nur werden die Fragen so umgeordnet, daß die erste gewählte Möglichkeit immer Archismus (A), die zweite immer Neuerung (N) bedeuten soll. Dieses hat den Vorteil, daß qualitative Information erhalten bleibt, denn die Gliederung einer Sprachfamilie soll nicht nach dem Kriterium gemeinsamer Eigenschaften schlechthin, dieses wären nämlich auch gemeinsam bewahrte Archaismen, erfolgen, sondern nach Neuerungen. Durch diese neue Formulierung der Fragen ist eine Einzelsprache schon auf den ersten Blick als archaisch oder als neuerungssüchtig einschätzbar.

Die Neuerungen werden, wo erforderlich, nach ihren qualitativen Ergebnissen indiziert, also N_1, N_2 usw., so daß beim Vergleich von zwei Sprachen gemeinsame Neuerungen und unterschiedliche Neuerungen getrennt werden können. Diese Unterscheidung ist wichtig, da der Begriff "gemeinsame Neuerung" eine ausschlaggebende Rolle bei der Gliederung einer Sprachfamilie spielt. Jede der zwölf Sprache, deren Platz in der romanischen Sprachfamilie zu bestimmen ist, wird hinsichtlich des umformu-

lierten Fragenkatalogs überprüft, d.h. die Fragen werden mit A, N bzw. N_1, N_2 usw. beantwortet. Damit erhält jede Sprache einen Merkmalskatalog, der etwas über Veränderungsprozesse vom Latein zu den modernen Einzelsprachen aussagt.

Nach diesem Merkmalskatalog werden jetzt jeweils zwei Sprachen miteinander verglichen (insgesamt 66 Vergleiche), wobei folgende Vergleichsmöglichkeiten auftreten können:

AA (= beide Sprachen haben nicht geneuert)

AN oder NA (= eine Sprache hat eine Neuerung vollzogen, die andere nicht)

NN oder $N_1 N_1$ oder $N_2 N_2$ (= beide Sprachen haben gleichmäßig geneuert)

$N_1 N_2$ (= beide Sprachen haben unterschiedlich geneuert)

Der Vergleich zwischen zwei Sprachen wird quantifiziert und eine größere oder kleinere Verwandtschaft festgestellt. Verwandtschaft würde hier als Ratio der erfolgten Neuerungen zu verstehen und zu definieren sein. (zu allen hier erwähnten Fragen, Antworten usw. vgl. 6. Dokumentation)

3. Rechengang

3.1. Die Quantifizierung geschieht durch eine Indexbildung. Als Index sei hier der Index E (schon bei Kristophson 1978 verwendet) vorgeschlagen:

$$E = \frac{2\Sigma NA + \Sigma N_1 N_2 - \Sigma NN + 40}{120}$$

Der Index ergibt Werte zwischen 0 und 1, wobei 0 Identität, 1 totale Unverwandtschaft bedeutet. ΣNA und $\Sigma N_1 N_2$ stellen die Differenzen, divergente Entwicklungen dar, ΣNN die konvergenten. ΣNA ist mit 2 multipliziert, also gewichtet. Dies geschah in der Annahme, daß die stärkste trennende Größe der Fall darstellt, wo eine Sprache neuert, die andere aber gerade nicht. Der Wert ΣAA ist nicht berücksichtigt, da bewahrte Archaismen wenig über die Gliederung einer Sprachfamilie aussagen. Indirekt geht aber der Wert ΣAA in die Berechnung ein, da die anderen Werte durch sein Fehlen leicht variiert werden. Aus diesen Gründen ist auch nicht der Index RIW_{jk} (vgl. Goebl 1983) verwendet worden. Der Index RIW_{jk} würde, angepaßt an die hier formulierten Fragestellungen und Begriffe folgende Formel bedeuten:

$$RIW_{jk} = 100 \quad \frac{\Sigma NN + \Sigma AA}{n}$$

bzw. um eine Wertetabelle zwischen 0 und 1 zu erreichen, vereinfacht:

$$RIW_{jk} = \frac{\Sigma NN + \Sigma AA}{n}$$

In diesem Index erhalten gerade die gemeinsamen Archaismen ein starkes Gewicht (s.o.), außerdem ergibt sich wegen der relativen Kleinheit von n (= 40) eine zu geringe Varianz der erechneten Werte, was bei Goebl (1983) wegen der hohen Zahl der verwendeten Merkmale nicht ins Gewicht fällt.

3.2. Die Berechnung des Verwandtschaftsgrade nach Index E ergibt eine Matrix mit folgenden Werten s. Tabelle I (Abkürzungen der Sprachnamen: R = Rumänisch, V = Vegliotisch, I = Italienisch, S = Sardisch, Fr = Friaulisch, E = Engadinisch, Pr = Provenzalisch, FPr = Frankoprovenzalisch, F = Französisch, K = Katalanisch, Sp = Spanisch, P = Portugiesisch).

Nun genügt eine solche Matrix noch nicht, um ein umfassendes Bild der Verwandtschaftsgrade zu zeichnen. Man kann wohl ablesen, wie nahe oder wie weit eine Sprache zur anderen steht, ja man kann sogar die Ausgangsfrage, mit welcher anderen romanischen Sprache das Rumänische am meisten verwandt ist, beantworten. Man muß in Tabelle I nur den kleinsten Wert für das Rumänische finden, das wäre 0.350. Damit hätte sich das Vegliotische als die am nächsten verwandte Sprache des Rumänischen erwiesen. Für die anderen Sprachen gilt Analoges, aber der Beobachter erfährt nur etwas über Zweierbeziehungen, nicht aber etwas über die Gliederung der gesamten Familie.

Abhilfe leistet hier die Erstellung eines hierarchisierten Dendrogramms nach dem Johnson-Maximumverfahren (vgl. Johnson 1967): s. Abb. I. Als Grenzwert ist hier die durchschnittliche Distanz angenommen, so

daß von einer über- bzw. unterdurchschnittlichen Verwandtschaft gesprochen werden kann.

Als nächster Arbeitsgang geschieht eine Umordnung der Matrix der Verwandtschaftsgrade in eine Tabelle (s. Tabelle II), die die Verwandtschaftsgrade jeder einzelnen Sprache in Rangstufen enthält. Als einfachste Auswertungsmöglichkeit bietet sich die Summierung der Verwandtschaftsgrade an. Die errechneten Summen lassen sich wieder nach ihrem Rang ordnen (s. Tabelle III).

Aus dieser Tabelle ist die Stärke der Verwandtschaftsgrade einer Einzelsprache oder ihre Einbindung in die Sprachfamilie zu ersehen. Die Zahlen der Tabelle kann man auch als Grad der Isolierung interpretieren, wobei der durchschnittliche Isolierungsgrad wieder als Grenze dienen mag. Die nach Rangstufen geordneten Verwandtschaftsgrade (s. Tabelle II) lassen sich auch als Kurve abbilden. Eine Kurvenanpassung (vorgenommen von G. Altmann, s. Tabelle IV) ergibt für alle Sprachen eine Exponentialkurve der allgemeinen Form: $y = ae^{b\,x}$. Die Konstante 'b' des Exponenten besagt, daß die relative Distanz in Bezug auf den Rang konstant ist, d.h. auch neue sprachliche Veränderungen einer Sprache verschieben nicht die Beziehungen zu den übrigen Sprachen sondern lassen die Rangstufe ihrer Verwandtschaftsgrade bestehen. Auf eine graphische Abbildung der Kurven wurde hier verzichtet, da die Zusammenhänge innnerhalb der Familie aus den zwölf Kurven nicht sichtbar werden.

Jedoch bietet sich eine andere Möglichkeit, die Verwandtschaftsgrade abzubilden, nämlich die Werte der Matrix (Tabelle I) in ein Koordinatensystem umzusetzen (durchgeführt nach und von a Campo, vgl. A Campo 1988). Ein solches System gibt die Lage der einzelnen Sprache wieder, es bietet eine Art Modellgeographie (s. Abb. II).

4. Ergebnisse, Interpretationen

4.1. Damit ist die verfahrenstechnische Seite der Problemlösung beendet, und die erreichten Ergebnisse müssen linguistisch interpretiert werden. Das Dendrogramm (Abb. I) ist nicht mit dem traditionellen Stammbaum (s. Schleicher 1863) zu verwechseln, sondern nur ein mathematisches Modell, das Zahlen ordnet. Der traditionelle Stammbaum basiert auf Divergenzen, deren distinktive Qualität auf ihrer von Linguisten angenommenen Wichtigkeit beruht. Ein starkes Argument für die Wichtigkeit ist ein hohes Alter der trennenden Erscheinung. Ist eine Trennung einmal eingetreten, so ist sie in ihrer Distinktivität durch andere Neuerungen nie mehr ausgleichbar. Dieses Modell erkent also nur Divergenzen an und daraus folgernd nur Brüche oder saubere Schnitte. Die nicht so eindeutige sprachliche Wirklichkeit wird durch das flexiblere Wellenmodell (s. Schmidt 1878) besser wiedergegeben. Es lassen sich in diesem Modell Veränderungsprozesse fast geographisch abbilden, man kann unterschiedliche Erstreckungen und Überschneidungen von Veränderungen erkennen. Eine Dynamisierung des Wellenmodells ist das eingangs erwähnte Explosionsmodell (s. 1.), das Zerfall, aber auch dabei neu entstehende Beziehungen erfassen soll. Aus einer Mischkalkulation von trennenden und verbindenden Veränderungsprozessen ist der Index E hervorgegangen, der die Grundlage der Matrix der Verwandtschaftsgrade (Tabelle I) geworden ist. Das Dendrogramm (Abb. I) bildet wohl als Resultat, genau wie der traditionelle Stammbaum, die Gliederung einer Sprachfamilie ab, nur eben nach ganz anderen Kriterien, nämlich nach numerisch ausgedrückten Distanzen.

Distanzen innerhalb einer Sprachfamilie setzen eine einheitliche, zeitlich frühere Vorsprache voraus. Beides steht mit außersprachlichen Faktoren, hauptsächlich zeitlichen, räumlichen und sozialen in einem gewissen Zusammenhang. Da sprachliche Einheit enge Kontakte der Sprecher bedeutet, bedeutet sprachliche Diversifikation, d.h. Entstehung einer Sprachfamilie Verminderung der Kontakte, bzw. räumliche Einschränkung, was örtlich durchaus zu einer Verstärkung der Kontakte führen kann. Der Zeitfaktor ist hier nur insofern wichtig, daß zu unterschiedlichen Zeitpunkten die räumliche Kontaktdichte sich verschoben haben kann. Diese allgemeine Annahme von Kontakten, die zu unterschiedlichen Zeitpunkten unterschiedliche Räume umfassen, ist besonders gut auf den Fall der Entstehung der romanischen Sprachfamilie zu übertragen (s. 1.) und war der Anlaß, das sogenannte Explosionsmodell anzusetzen, das, wie gesagt, Wellen- und Stammbaummodell weiterentwickelt. Der Zeitfaktor ging in den Index durch die Teilung der Merkmale in Neuerungen und Archaismen ein. Eine Chronologie für die Trennung der romanischen Sprachfamilie ist aus dem Dendrogramm nicht zu schließen, da für den Zerfall der einheitlichen Vorsprache, des Latein, Gleichzeitigkeit angenommen wurde. Das Dendrogramm (Abb. I), mehr aber noch das Koordinatensystem (Abb. II) spiegeln eine ideale Geographie wider, widersprechen aber auch nicht der tatsächlichen. Beide Abbildungen stützen die an sich naheliegende Annahme eines Zusammenhangs zwischen geographischer Nachbarschaft und sprachlicher Verwandtschaft.

Zu diesem Bild passen die Rangstufen des Grades der Isolierung (Tabelle III), gemäß denen die Sprachen R, I, S, F als überdurchschnittlich isollert zu gelten haben Die Sprachen R, I, S sind in dieser Position zu

finden, weil sie mit keiner anderen Sprache (oder wie bei R nur mit dem zweifelhaften Partner V, vgl. 4.2) überdurchschnittlich verwandt sind. Dagegen macht die Position von F zwar einen Extrempunkt aus, aber F besitzt doch eine Reihe nahestehender Verwandter. Umgekehrt bedeutet der geringste Isolierungsrang von K, daß diese Sprache vielen Sprachen nahesteht. Daher auch das ständige Verschieben dieser Sprache zwischen der Iberia und der Gallia in der romanistischen Diskussion.

4.2. Als Ergebnis lassen sich für die Gliederung der Romania und zur Entscheidung einiger Einzelfragen folgende Aussagen machen:

Die Romania wäre in zwei klare Großgruppen, Gallia und Iberia, und in mehrere einzelne Äste, eine "Romania segregata" zu zerlegen. Was traditionell als westromanisch gilt, also Iberia und Gallia, wären eher zwei Gruppen, das traditionelle Ostromanische gibt es gar nicht. Italienisch tendiert eher zur Iberia, aber auch diese Bindung ist unterdurchschnittlich. Überhaupt sind die Bindungen innerhalb der "Romania segregata", aber auch zu den anderen Gliedern der Familie recht schwach, daher sind alle Sprachen dieser Gruppe überdurchschnittlich isoliert (vgl. Tabelle III und 4.1). Am meisten gerechtfertigt wäre also eine Fünferteilung für die Gesamtromania.

Die Position des Vegliotischen (Dalmatischen) erscheint geographisch durchaus sinnvoll, nur sind die Daten für diese Sprache recht unsicher, da es zur Zeit der Aufnahme nur noch einen Sprecher gab. Berücksichtigt man daher das Vegliotische nicht, so würde sich das Rumänische als nächster, aber sehr entfernter Verwandter des Sardischen erweisen.

Interessant bleibt die starke Isolierung des Italienischen. Aus der Isolierung des Italienischen darf man wahrscheinlich auf den konservativen Charakter dieser Sprache schließen. Gleiches würde auch für die anderen isolierten Sprachen, bis auf die erklärbare Ausnahme des Französischen gelten (s. 4.1).

Die beiden rätoromanischen Sprachen, Engadinisch und Friaulisch, sind sich am nächsten verwandt. Dies könnte ein zusätzliches Argument für die allgemein angenommene, aber auch bestrittene rätoromanischen Einheit abgeben. Diese Einheit steht der Gallia im engeren Sinne am nächsten, ja sie bildet einen Teil von ihr. Auch dies wäre als Argument für die Diskussion um die Position des Rätoromanischen zu nutzen.

Auch für das Katalanisch würde sich nach den vorgelegten Ergebnissen ein fester Platz in der Iberia ergeben (vgl. 4.1).

Keine Argumente liefert das vorgestellte Verfahren für die Sprache-Dialekt-Diskussion, die oft und gerne in der Romanistik geführt wird und wurde, letztlich eine Diskussion, für die es wenig linguistische Argumente

gibt. Auch sogenannte kulturelle Faktoren, die gerne dazu als letzte entscheidende Kriterien herangezogen werden, wirken nicht sehr überzeugend. Die Erhebungsdaten stammen aus einer Zeit, die jetzt ca. hundert Jahre zurückliegt und die noch die idealen Informanten lieferte mit den Eigenschaften: agrarisch, immobil, analphabetisch. Diese Eigenschaften waren damals keine Rarität, sondern durchaus noch repräsentativ. Letztlich entscheidend sind die politischen Ambitionen der großen Zentren Madrid, Paris, Rom gewesen, die politische Realitäten schufen und auch kulturelle Orientierungen für verbindlich erklären konnten.

5. Allgemeine Ergebnisse

Eine Sprachfamilie läßt sich als Netz von Beziehungen definieren, das im Laufe der Zeit, durch geographische Bedingungen und durch politische kulturelle Ereignisse von einem sehr engen Gewebe ausgehend immer lokkerer wird. Sprachlich sind die Beziehungen in erster Linie morphologische und lautliche Eigenschaften, die sich verändern und zeitlich verschobene neue Beziehungen widerspiegeln (vgl. Beispiel 0.1). Die Anzahl der Veränderungen muß weder allgemein noch innerhalb einer Sprachfamilie gleich groß sein, es ist stets mit konservativeren und mit progressiveren Sprachen zu rechnen. Konservative Sprachen innerhalb einer Sprachfamilie sichern aber nicht eine überzeugende Gliederung der Familie besser, sondern führen zu isolierten Strängen.

Phänomene der Wortschatzveränderung sind hier nicht aufgenommen worden, da sie im Gegensatz zu morphologischen und lautlichen Veränderungen zu singuläre Fälle darstellen. So zeigt das bunte Bild der Neologismen der romanischen Sprachen, die dazu zumeist aus lateinisch-griechischem Material geprägt sind, daß keine romanischen Neologismen vorliegen, sondern nur französische, italienische usw.

Da es konservative und progressive Sprachen gibt, dieses aber eher von der Kontaktsituation abhängt, läßt sich kein Zusammenhang zwischen Veränderungsrate und Zeitfaktor erkennen. Hier wurde einerseits der Wortschatz nicht berücksichtigt, andererseits gezeigt, daß Konservatismus auch nicht zu besonderer Nähe in der Familie führt, so daß die erreichten Ergebnisse keine Stütze für eine Glottochronologie, eher im Gegenteil Argumente zur Falsifikation liefern.

Offenbar ist eine Sprachfamilie auch zum immer weiterführenden Zerfall bestimmt, wobei die nächsten Verwandten immer die nächsten

bleiben, ihre Abstände voneinander im Laufe der Zeit aber wachsen. Zeitlich zurückversetzt würde dies bedeuten, daß sich die jeweils nächsten Verwandten näher standen als heute. Den Prozeß des relativ gleichmäßigen Zerfalls legt der Charakter der Exponentialkurve nahe, die die Verwandtschaftsgrade abbildet (vgl. Tabelle II, IV). Vergleichbar wäre die Zerfallsrate bei radioaktiven Elementen oder Isotopen.

Sprachliche Konvergenzprozesse innerhalb einer Sprachfamilie, wie totale Assimilation, partielle Assimilation, Sprachbunderscheinungen sind möglich und denkbar, nur würden solche Erscheinungen ganz anderen Kriterien unterliegen und wären ebenso unter nicht verwandten Sprachen möglich, so daß sie das Konzept der Sprachfamilie nicht berühren. Auch das Aussterben einer Sprache oder einer ganzen Familie falsifiziert die geäußerten Annahmen nicht.

Die vorgelegte Untersuchung ist um neue Daten erweiterbar, auch prinzipiell auf andere Sprachfamilien übertragbar. Allerdings ist zu bedenken, daß wenige Sprachfamilien über solche Datenmengen und über einen solchen Bekanntheitsgrad wie die romanische verfügen.

Die Konzeption der Sprachfamilie, der größte qualitative Sprung seit Erfindung der Buchstaben für die Linguistik, bietet ein Ordnungsprinzip sui generis, das erlaubt inner- und außersprachliche Zustände und Ereignisse in Beziehung zu setzen, ja bishin zu Vorstufen von Gesetzen zu kommen.

Überhaupt sollte gezeigt werden, daß eine Quantifizierung auch für Gliederungsfragen einer Sprachfamilie sinnvoll ist, daß sie einerseits Argumente und Indizien für alte Streitfragen liefert, daß sie andererseits auch andeutet, daß sprachliche Veränderungsprozesse gewissen allgemeinen, auch sonst vorkommenden Gesetzmäßigkeiten unterworfen sind.

A n h a n g

Tabelle I

Matrix der Verwandtschaftsgrade (Werte des Index E)

	R	V	I	Š	Fr	E
R						
V	0.350					
I	0.483	0.383				
Š	0.483	0.483	0.466			
Fr	0.475	0.366	0.450	0.535		
E	0.483	0.366	0.433	0.525	0.308	
Pr	0.500	0.400	0.475	0.583	0.341	0.375
FPr	0.566	0.425	0.475	0.666	0.333	0.341
F	0.625	0.483	0.541	0.723	0.400	0.400
K	0.490	0.383	0.450	0.533	0.341	0.358
Šp	0.433	0.425	0.433	0.490	0.416	0.433
P	0.441	0.500	0.441	0.541	0.490	0.425
	5.329	4.214	4.164	4.596	2.629	2.332

	Pr	FPr	F	K	Šp	P
R						
V						
I						
Š						
Fr						
E						
Pr						
FPr	0.300					
F	0.300	0.133				
K	0.341	0.350	0.408			
Šp	0.383	0.466	0.525	0.283		
P	0.433	0.416	0.475	0.316	0.216	
	1.757	1.365	1.408	0.599	0.216	0.000
28.609 (:66 = 0.4334)						

Tabelle II
Summe der Distanzen

R	V	I	S	Fr	E
0.350	0.350	0.383	0.466	0.308	0.308
0.433	0.366	0.433	0.483	0.333	0.341
0.441	0.366	0.433	0.483	0.341	0.358
0.475	0.383	0.441	0.490	0.341	0.366
0.483	0.383	0.450	0.525	0.366	0.375
0.483	0.400	0.450	0.533	0.400	0.400
0.483	0.425	0.466	0.535	0.416	0.425
0.490	0.425	0.475	0.541	0.450	0.433
0.500	0.483	0.475	0.583	0.475	0.433
0.566	0.483	0.483	0.666	0.490	0.483
0.625	0.500	0.541	0.723	0.535	0.525
5.329	4.564	5.030	6.028	4.455	4.447

Pr	FPr	F	K	Sp	P
0.300	0.133	0.133	0.283	0.216	0.216
0.300	0.300	0.300	0.316	0.283	0.316
0.341	0.333	0.400	0.341	0.383	0.416
0.341	0.341	0.400	0.341	0.416	0.425
0.375	0.350	0.408	0.350	0.425	0.433
0.383	0.416	0.475	0.358	0.433	0.441
0.400	0.425	0.483	0.383	0.433	0.441
0.433	0.466	0.525	0.408	0.433	0.475
0.475	0.475	0.541	0.450	0.466	0.490
0.500	0.566	0.625	0.490	0.490	0.500
0.583	0.666	0.723	0.533	0.525	0.541
4.431	4.471	5.013	4.253	4.503	4.694

Summe der Distanzen = 57.218 (: 12 = 4.768 = durchschnittliche Distanz bzw. Isolierung

Tabelle III
Rang der Isolierung

1.	S	6.028
2.	R	5.329
3.	I	5.030
4.	F	5.013
5.	P	4.694
6.	V	4.564
7.	Sp	4.503
8.	FPr	4.471
9.	Fr	4.455
10.	E	4.447
11.	Pr	4.431
12.	K	4.253

Tabelle IV
Kurvenanpassung von Tabelle II

(Zweite Spalte = beobachtete Werte aus Tabelle II
dritte Spalte = theoretische Werte nach $y = ae^{bx}$

Rumänisch

1	0,3500	0,3930
2	0,4330	0,4095
3	0,4410	0,4267
4	0,4750	0,4446
5	0,4830	0,4632
6	0,4830	0,4826
7	0,4830	0,5028
8	0,4900	0,5239
9	0,5000	0,5459
10	0,5660	0,5687
11	0,6250	0,5926

a = 0,3787 b = 0,0392
Det.koef = 0,8252

Vegliotisch

1	0,3500	0,3468
2	0,3660	0,3596
3	0,3660	0,3728
4	0,3830	0,3865
5	0,3830	0,4007
6	0,4000	0,4154
7	0,4250	0,4307
8	0,4250	0,4465
9	0,4830	0,4629
10	0,4830	0,4799
11	0,5000	0,4976

a = 0,3293 b = 0,0367
Det.koef = 0,9438

Sardisch

1	0,4660	0,4409
2	0,4830	0,4493
3	0,4830	0,4785
4	0,4900	0,4985
5	0,5250	0,5193
6	0,5330	0,5410
7	0,5350	0,5636
8	0,5410	0,5872
9	0,5830	0,6117
10	0,6660	0,6373
11	0,7230	0,6639

a = 0,4190 b = 0,0435
Det.koef = 0,8548

Italienisch

1	0,3830	0,4014
2	0,4330	0,4120
3	0,4330	0,4228
4	0,4410	0,4339
5	0,4500	0,4452
6	0,4500	0,4569
7	0,4660	0,4689
8	0,4750	0,4812
9	0,4750	0,4939
10	0,3830	0,5068
11	0,5410	0,5201

a = 0,3934 b = 0,0240
Det.koef = 0,8436

Engadinisch

1	0,3080	0,3187
2	0,3410	0,3340
3	0,3580	0,3501
4	0,3660	0,3669
5	0,3750	0,3846
6	0,4000	0,4031
7	0,4250	0,4225
8	0,4330	0,4428
9	0,4330	0,4641
10	0,4830	0,4864
11	0,5250	0,5098

a = 0,2998 b = 0,0476
Det.koef = 0,9597

Friaulisch

1	0,3080	0,2978
2	0,3330	0,3152
3	0,3410	0,3335
4	0,3410	0,3530
5	0,3660	0,3735
6	0,4000	0,3953
7	0,4160	0,4183
8	0,4500	0,4427
9	0,4750	0,4684
10	0,4900	0,4957
11	0,5350	0,5246

a = 0,2921 b = 0,0530
Det.koef = 0,9830

Frankoprovenzalisch

1	0,1330	0,2357
2	0,3000	0,2605
3	0,3330	0,2878
4	0,3410	0,3181
5	0,3500	0,3515
6	0,4160	0,3884
7	0,4250	0,4292
8	0,4660	0,4743
9	0,4750	0,5241
10	0,5660	0,5792
11	0,6660	0,6400

a = 0,2116 b = 0,1004
Det.koef = 0,9063

Provenzalisch

1	0,3000	0,2847
2	0,3000	0,3039
3	0,3410	0,3244
4	0,3410	0,3464
5	0,3750	0,3698
6	0,3830	0,3948
7	0,4000	0,4214
8	0,4330	0,4499
9	0,4750	0,4803
10	0,5000	0,5127
11	0,5830	0,5474

a = 0,2651 b = 0,0664
Det.koef = 0,9625

Portugiesisch

1	0,2160	0,3211
2	0,3160	0,3389
3	0,4160	0,3577
4	0,4250	0,3775
5	0,4330	0,3984
6	0,4410	0,4205
7	0,4410	0,4439
8	0,4750	0,4685
9	0,4900	0,4944
10	0,5000	0,5219
11	0,5410	0,5508

a = 0,3059 b = 0,0536
Det.koef = 0,7624

Französisch

1	0,1330	0,2667
2	0,3000	0,2941
3	0,4000	0,3243
4	0,4000	0,3576
5	0,4080	0,3944
6	0,4750	0,4349
7	0,4830	0,4796
8	0,5250	0,5289
9	0,5410	0,5832
10	0,6250	0,6431
11	0,7230	0,7092

a = 0,2443 b = 0,0967
Det.koef = 0,8819

Spanisch

1	0,2160	0,3026
2	0,2830	0,3214
3	0,3830	0,3401
4	0,4160	0,3600
5	0,4250	0,3801
6	0,4330	0,4032
7	0,4330	0,4268
8	0,4330	0,4517
9	0,4660	0,4781
10	0,4900	0,5060
11	0,5250	0,5355

a = 0,2880 b = 0,0564
Det.koef = 0,7743

Katalanisch

1	0,2830	0,2828
2	0,3160	0,2991
3	0,3410	0,3164
4	0,3410	0,3347
5	0,3500	0,3541
6	0,3580	0,3746
7	0,3580	0,3962
8	0,4080	0,4191
9	0,4500	0,4434
10	0,4900	0,4690
11	0,5330	0,4961

a = 0,2658 b = 0,0566
Det.koef = 0,8902

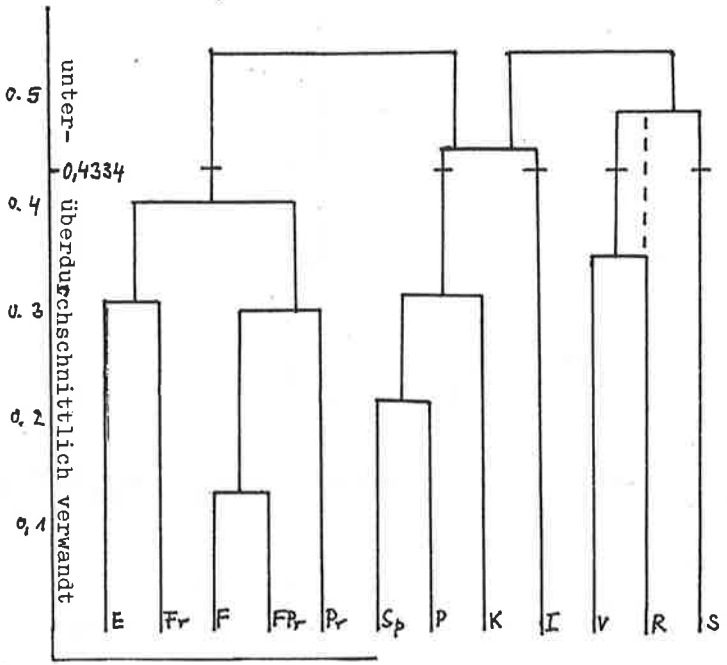


Abbildung I. Ordnung der Werte von Tabelle I in ein Dendrogramm (nach Johnson 1967)

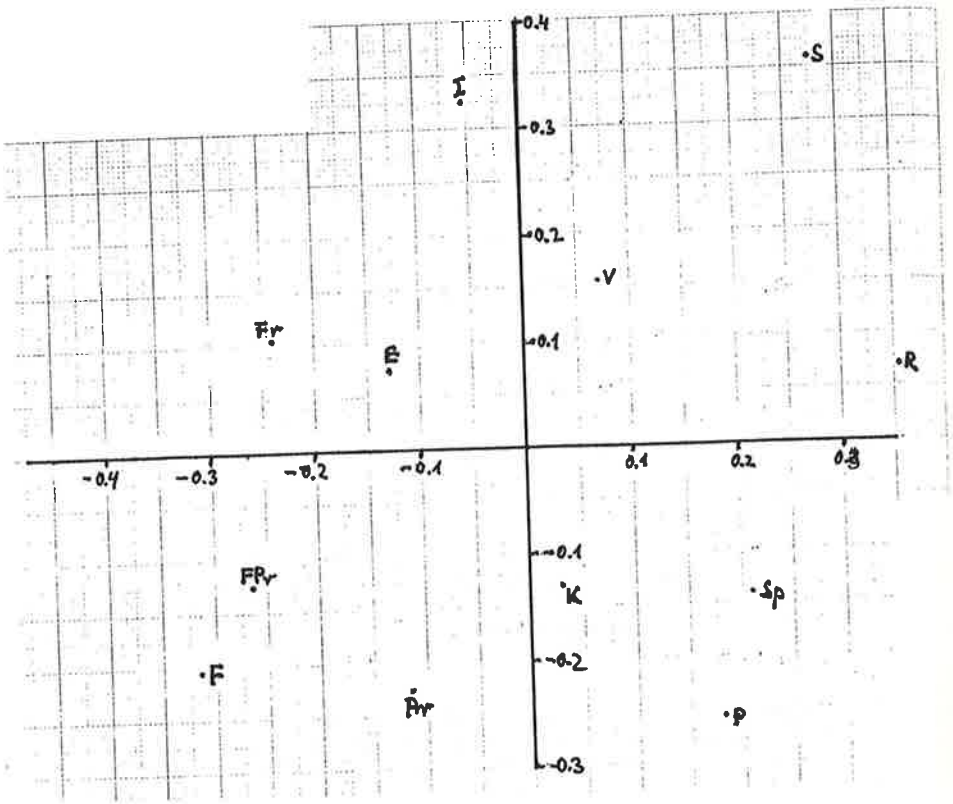


Abbildung II. Umsetzung der Werte von Tabelle I in ein Koordinatensystem (Verfahren und Durchführung von a Campo)

(R - Rumänisch, V - Vegliotisch (Dalmatisch), I - Italienisch, S - Sardisch, Fr - Friaulisch, E - Engadinisch, Pr - Provenzalisch, FPr - Frankoprovenzalisch, F - Französisch, K - Katalanisch, Sp - Spanisch, P - Potugiesisch)

Frageliste und beantwortete Fragen von Muljačić 1967 (40 spezifische Merkmale der romanischen Sprachen)

1. phonologischer/nichtphonologischer Akzent
2. zwei/mehr als zwei Vokalklassen
3. dreieckiger/viereckiger Vokalismus
4. dreistufiger/vier-(oder fünf-)stufiger Vokalismus
5. phonologisch gespannte (lange) Vokale bestehen/bestehen nicht
6. nasale Vokale bestehen/bestehen nicht als Phoneme
7. doppelte Konsonanten bestehen/bestehen nicht
8. lat. finales *-s* in der Aussprache erhalten/nicht erhalten
9. präpositiver/postpositiver Artikel
10. *ipse/ille* als Artikel
11. mehrere Kasusformen in gleichem Numerus unterschieden/nur eine Kasusform
12. Substantiv für Lebewesen als Objekt mit/ohne Präposition
13. Pluralbildung durch Endungen/auf andere Art
14. Dreigradigkeit/Zweigradigkeit des Demonstrativpronomen
15. Komparativ mit *magis*/mit *plus*
16. *-t* in der Endung der 3. Pers. Pl. erhalten/nicht erhalten
17. klass.-lat. Funktion des Inchoativ-Suffixes *-esc-* erhalten/nicht erhalten
18. Imperf. Ind. des Verbs "sein" von *esse*/von einem anderen Verb
19. mehrere Endungstypen/ein Typ für das Imperf. Ind.
20. Futur I des Typs *cantare* + *habeo* (u.ä.)/auf andere Art
21. Perfekt des Verbs "sein" mit demselben Hilfsverb (z.B. ital. *sono stato*)/mit einem anderen Hilfsverb (z.B. franz. *j'ai été*, port. *tenho sido*)
22. Perfekt reflexiver Verben mit dem Hilfsverb *habere* oder *tenere*/mit *esse*
23. Hilfsverb *habere/tenere*
24. haupttoniges *non* ergab zwei Formen/eine Form
25. lat. Vokale *i, e* ergaben/ergaben nicht ein Phonem
26. *stella* > *stela/stella* > *stella*
27. lat. Vokale *u, o* ergaben/ergaben nicht ein Phonem

28. Vokal in der Pänultima der proparoxytonen Erbwörter gewöhnlich erhalten/gewöhnlich nicht erhalten
29. das Kriterium der offenen Silbe war in längst vergangener Zeit wichtig (in Verbindung mit der Diphthongierung)/nicht wichtig
30. die intervokalischen stimmlosen Konsonanten *p, t, k, s* erhalten/weiterentwickelt (sonorisiert usw.)
31. die lat. Phoneme *k, g* vor den palatalen Vokalen *e, i* haben heute parallele Reflexe (z.B. ital. *cento, gente* mit /c/, /g/)/haben keine parallelen Reflexe (z.B. franz. *cent, gens* mit /s/, /z/)
32. die lat. Phoneme *k* (bzw. *g*) gaben gleiche Reflexe vor *e, i, j* (z.B. ital. *cento, citta, braccia*)/gaben nicht gleiche Reflexe
33. gleiche/verschiedenen Ergebnisse aus klassischem *i* und g^e, g^i
34. das labiale Element der Gruppe *qu* ist auf irgendeine Weise erhalten/ist längst verschwunden
35. *qu, gu* + *e, i* werden/werden nicht palatalisiert
36. die lat. Phoneme *k, g* + *a* werden nicht/werden palatalisiert
37. die Konsonantengruppen *kl-, gl-* am Anfang eines Wortes werden nicht/werden palatalisiert (ganz oder nur im zweiten Teil)
38. die Konsonantengruppen *pl, bl, fl* werden nicht/werden palatalisiert (ganz oder nur im zweiten Teil)
39. die Konsonantengruppen *-kt-, -ks-* entwickelten sich nicht/entwickelten sich in palataler Richtung
40. die Konsonantengruppe, die lat. *-gn-* geschrieben wurde, entwickelte sich nicht/entwickelte sich in palataler Richtung

Tabelle 1

Frage	R	V	I	S	Fr.	E	Pr.	FPr.	F	K	Sp.	P
1.	+	+	+	+	+	+	+	+	—	+	+	+
2.	—	+	+	+	+	—	—	—	—	+	+	—
3.	+	+	+	+	+	+	—	+	—	+	+	+
4.	+	+	—	+	+	—	+	—	—	—	+	—
5.	—	—	—	—	+	—	—	+	+	—	—	—
6.	—	—	—	—	—	—	—	+	+	—	—	+
7.	—	—	+	+	—	—	—	—	—	—	—	—
8.	—	—	—	+	+	+	±	±	±	+	+	+
9.	—	+	+	+	+	+	+	+	+	+	+	+
10.	—	—	—	+	—	—	—	—	—	±	—	—
11.	+	—	—	—	—	—	—	—	—	—	—	—
12.	+	—	—	+	—	+	—	—	—	—	+	+
13.	+	+	+	+	+	+	—	±	—	+	+	+
14.	—	—	±	+	—	—	—	—	—	+	+	+
15.	+	—	—	—	—	—	±	—	—	+	+	+
16.	—	—	—	—	—	—	—	±	±	—	—	—
17.	+	+	+	—	+	+	+	+	+	+	—	—
18.	+	+	+	+	+	+	+	—	—	+	+	+
19.	+	+	+	+	+	+	+	+	—	+	+	+
20.	—	—	+	±	+	±	+	+	+	+	+	+
21.	—	+	+	+	±	+	+	—	—	—	—	—
22.	+	+	—	—	—	+	—	—	—	+	+	+
23.	+	+	+	+	+	+	+	+	+	+	+	—
24.	—	—	—	+	—	—	—	—	—	—	—	—
25.	+	+	+	—	+	+	+	+	+	+	+	+
26.	—	—	—	—	+	+	+	+	+	+	—	—
27.	—	+	+	—	+	+	+	+	+	+	+	+
28.	+	—	+	+	—	—	±	—	—	—	±	+
29.	—	+	+	—	—	+	—	+	+	—	—	—
30.	+	+	+	+	—	—	—	—	—	—	—	—
31.	+	—	+	+	+	—	—	—	—	—	—	—
32.	—	—	+	—	+	+	—	—	—	—	—	—
33.	—	±	+	—	—	—	+	+	+	+	+	+
34.	—	—	+	+	—	+	—	—	—	—	—	—
35.	+	+	—	—	+	±	—	—	—	—	—	—
36.	+	+	+	+	—	—	±	—	—	+	+	+
37.	—	+	—	+	+	+	+	+	+	+	—	—
38.	+	+	—	+	+	+	+	+	+	+	—	—
39.	+	+	+	+	—	—	—	—	—	—	—	—
40.	+	+	—	+	—	—	—	—	—	—	—	—
	R	V	I	S	Fr.	E	Pr.	FPr.	F	K	Sp.	P

Tabelle 2

	R	V	I	S	Fr.	E	Pr.	FPr.	F	K	Sp.	P
R												
V	23											
I	37	26										
S	33	32	30									
Fr.	37	26	32	38								
E	38	25	31	40	19							
Pr.	38	27	33	41	21	24						
FPr.	47	32	36	52	22	23	19					
F	54	39	43	59	29	30	22	7				
K	37	26	32	38	22	23	19	20	29			
Sp.	31	30	30	34	30	31	25	34	41	14		
P	34	39	33	41	39	32	30	31	38	21	9	
	R	V	I	S	Fr.	E	Pr.	FPr.	F	K	Sp.	P

Rangliste

1. Sardisch	438	7. Frankoprovenzalisch	323
2. Rumänisch	409	8. Engadinisch	316
3. Französisch	391	9. Friaulisch	315
4. Italienisch	363	10. Spanisch	309
5. Portugiesisch	347	11. Provenzalisch	299
6. Vegliotisch	325	12. Katalanisch	281

Umformulierte Frageliste

Frage	=A	=N
1	nichtphonologischer	/ phonologischer Akzent
2	zwei Vokalklassen	/ mehr als zwei Vokalklassen
3	dreieckiger	/ viereckiger Vokalismus
4	dreistufiger	/ vier- (oder fünf-)stufiger Vokalismus
5	phonologisch gespannte (lange) Vokale bestehen nicht	/ bestehen
6	nasale Vokale bestehen nicht als Phoneme	/ bestehen als Phoneme
7	doppelte Konsonanten bestehen	/ bestehen nicht
8	lat. finales '-s' in der Aussprache erhalten	/ nicht erhalten
9	---	/ N_1 postpositiver Artikel / N_2 präpositiver Artikel
10	---	/ N_1 'ipse' als Artikel / N_2 'ille' als Artikel
11	mehrere Kasusformen in gleichem Numerus unterschieden	/ nur eine Kasusform
12	Substantiv für Lebewesen als Objekt ohne Präposition	/ mit Präposition
13	Pluralbildungen durch Endungen	/ auf andere Art
14	Dreigradigkeit	/ Zweigradigkeit des Demonstrativpronomens
15	---	/ N_1 Komparativ mit 'magis' N_2 Komparativ mit 'plus'
16	'-t'in der Endung der 3. Pers. Pl. erhalten	/ nicht erhalten
17	klass. lat. Funktion des Inchoativsuffixes '-esc' erhalten	/ nicht erhalten
18	Imperf.Ind. des Verbums "sein" von 'esse'	/ von einem anderen Verb

19 mehrere Endungstypen / ein Typ für das Imperf. Ind.

20 --- / N_1 Futur des Typs 'cantare + habeo'
N_2 auf andere Art

21 --- / N_1 Perfekt des Verbs "sein" mit demselben Hilfsverb (z.B. ital. 'sono stato')
N_2 mit einem anderen Hilfsverb

22 --- / N_1 Perfekt reflexiver Verben mit dem Hilfsverb 'habere' oder 'tenere'
N_2 mit 'esse'

23 --- / N_1 Hilfsverb 'habere'
N_2 Hilfsverb 'tenere'

24 haupttoniges 'non' ergab eine Form / zwei Formen

25 lat. Vokale 'i, e' ergaben nicht ein Phonem / ergaben ein Phonem

26 --- / N_1 'stella' > 'stela'
/ N_2 'stella' > 'stella'

27 lat. Vokale 'u, o' ergaben nicht ein Phonem / ergaben ein Phonem

28 Vokal in der Pänultima der proparoxytonen Erbwörter gewöhnlich erhalten / gewöhnlich nicht erhalten

29 das Kriterium der offenen Silbe war in längst vergangener Zeit nicht wichtig (in Verbindung mit der Diphthongierung) / war wichtig

30 die intervokalischen stimmlosen Konsonanten 'p, t, k, s' erhalten / weiterentwickelt / (sonorisiert, usw.)

31 --- / N_1 die lat. Phoneme 'k, g' vor den palatalen Vokalen 'e, i' haben heute parallele Reflexe (z.B. ital. 'cento, gente' mit /c/, /g/)

N_2 haben keine parallelen Reflexe (z.B. franz. 'cent, gens' mit /s/, /z/)

32 --- / N_1 die lat. Phoneme 'k' (bzw. 'g') gaben gleiche Reflexe vor 'e, i, j' (z.B. ital. 'cento, citta, braccia')
N_2 gaben nicht gleiche Reflexe

33 -- / N_1 gleiche Ergebnisse aus klassischem 'i' und 'g^e, g^i'
N_2 verschiedene Ergebnisse

34 das labiale Element der Gruppe 'qu' ist auf irgendeine Weise erhalten / ist längst verschwunden

35 'qu, gu + e, i' werden nicht palatalisiert / werden palatalisiert

36 die lat. Phoneme 'k, g + a' werden nicht palatalisiert / werden palatalisiert

37 die Konsonantengruppen 'kl-, gl-' am Anfang eines Wortes werden nicht palatalisiert / werden palatalisiert (ganz oder nur im zweiten Teil)

38 die Konsonantengruppen 'pl, bl, fl' werden nicht palatalisiert / werden palatalisiert (ganz oder nur im zweiten Teil)

39 --- / N_1 die Konsonantengruppen '-kt-, -ks-' entwickelten sich nicht in palataler Richtung
N_2 entwickelten sich in palataler Richtung

40 --- / N_1 die Konsonantengruppen, die lat. '-gn-' geschrieben wurde, entwickelte sich nicht in palataler Richtung
N_2 entwickelte sich in

Anmerkung: Der Wortlaut ist so weit wie möglich der gleiche wie bei Muljačić 1967. An sich müßte in Spalte N immer stehen: .. geworden zu.., .. entstanden aus.., .. weggefallen.. u.ä., da es sich um Prozesse und nicht um Zustände handelt. Für die Berechnungen ist allerdings der Wortlaut der Fragen nicht erheblich.

Beantwortung der umformulierten Frageliste

Abkürzung der Sprachen: R = Rumänisch, V = Vegliotisch (Dalmatisch), I = Italienisch, S = Sardisch, Fr = Friaulisch, E = Engadinisch, Pr = Provencalisch, FPr = Frankoprovencalisch, F = Französisch, K = Katalanisch, Sp = Spanisch, P = Portugiesisch.

Frage	Sprachen											
	R	V	I	S	Fr	E	Pr	FPr	F	K	Sp	P
1:	N	N	N	N	N	N	N	N	A	N	N	N
2:	N	A	A	A	A	N	N	N	N	A	A	N
3:	A	A	A	A	A	A	N	A	N	A	A	A
4:	A	A	N	A	A	N	A	N	N	N	A	N
5:	A	A	A	A	N	A	A	N	N	A	A	A
6:	A	N	A	A	A	A	A	N	N	A	A	N
7:	N	N	A	A	N	N	A	N	N	A	N	A
8:	N	N	N	A	A	A	NA	NA	NA	A	A	N
9:	N1	N2	N2	N2	N2	N2	N2	N2	N2	N2	N2	N2
10:	N2	N2	N2	N1	N2	N2	N2	N2	N2	N1N2	N2	N2
11:	A	N	N	N	N	N	N	N	N	N	N	N
12:	N	A	A	N	A	N	A	A	A	A	N	N
13:	A	A	A	A	A	A	N	NA	N	A	N	A
14:	N	N	NA	A	N	N	N	N	N	A	A	A
15:	N1	N2	N2	N2	N2	N2	N1N2	N2	N2	N1	N1	N1
16:	N	N	N	N	N	N	N	NA	NA	N	N	N
17:	N	N	N	A	N	N	N	N	N	N	A	A
18:	A	A	A	A	A	A	A	N	N	A	A	A
19:	A	A	A	A	A	A	A	A	N	A	A	A
20:	N2	N2	N1	N1N2	N1	N1N2	N1	N1	N1	N1	N1	N1
21:	N2	N1	N1	N1	N1N2	N1	N1	N2	N2	N2	N2	N2
22:	N1	N1	N2	N2	N2	N1	N2	N2	N2	N1	N1	N1
23:	N1	N1	N1	N1	N1	N1	N1	N1	N1	N1	N1	N2
24:	A	A	A	N	A	A	A	A	A	A	A	A
25:	N	N	N	A	N	N	N	N	N	N	N	N
26:	N2	N2	N2	N2	N1	N1	N1	N1	N1	N1	N2	N2
27:	A	N	N	A	N	N	N	N	N	N	N	N
28:	A	N	A	A	N	N	NA	N	N	N	NA	A
29:	A	N	N	A	A	N	A	N	N	A	A	A
30:	A	A	A	A	N	N	N	N	N	N	A	N
31:	N1	N2	N1	N1	N1	N2	N2	N2	N2	N2	N2	N2
32:	N2	N2	N1	N2	N1	N1	N2	N2	N2	N2	N2	N2
33:	N2	N1N2	N1	N2	N2	N2	N1	NJ	N1	N1	N1	N1
34:	N	N	A	A	N	A	N	N	N	N	N	N
35:	N	N	A	A	N	NA	A	A	A	A	A	A
36:	A	A	A	A	N	N	NA	N	N	A	A	A
37:	N	A	N	A	A	A	A	A	A	A	N	N
38:	A	A	N	A	A	A	A	A	A	A	N	N
39:	N1	N1	N1	N1	N2	N2	N2	N2	N2	N2	N2	N2
40:	N1	N1	N2	N1	N2	N2	N2	N2	N2	N2	N2	N2

Anmerkung: Die Beantwortung NA, N_1N_2 besagt, daß kein einheitlicher Reflex im Sinne der Frage vorliegt, sondern daß beide abgefragten Merkmale vorkommen. Bei den Vergleichen werden die Antworten wie folgt gewertet:

NA	: NA wie N : N	N_1N_2	: N_1 wie N_2 : N_1
NA	: N wie N_1 : N_2	NA	: A wie N : A

Literatur

a Campo, F., Gersić, S., Naumann, C.L., Altmann, G. (1989). Subjektive Ähnlichkeit deutscher Laute. *Glottometrika 10*, 46-70.

Goebl, H. (1983). "Stammbaum" und "Welle". Vergleichende Betrachtungen aus numerisch-taxonomischer Sicht. *Zeitschrift für Sprachwissenschaft* 2, 403-444.

Iliescu, M. (1969). Rassemblances et dissemblances entre les langues romanes du point de vue de la morpho-syntaxe verbale. *Revue de Linguistique Romane 33, 113-132.*

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika 32, 241-254.*

Kristophson, J. (1978). Überlegungen zur Anwendung mathematischer Methoden im Bereich der historischen Sprachwissenschaft. Slavistische Linguistik 1977, 86-99. *Slavistische Beiträge* 120. München.

Kristophson, J. (1984). Zur Meßbarkeit von Sprachverwandtschaft. *Folia Linguistica Historica IV, 305-314.*

Muljačić, Z. (1967). Die Klassifikation der romanischen Sprachen. *Romanistisches Jahrbuch 18, 23-37.*

Schleicher, A. (1863). Die Darwinsche Theorie und die Sprachwissenschaft. Weimar.

Schmidt, J. (1872). Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar.

Bibliographische Anmerkung: eine Kurzfassung dieses Aufsatzes ist im Tagungsband der 12. Jahrestagung "Klassifikation und Ordnung" der Gesellschaft für Klassifikation e.V. erschienen (1989).

Hřebíček, L. (ed.), Glottometrika 11, 1989

Semantische Motivation der Genuszuweisung

Ursula Rothe, Bochum

1. Die Aufteilung der Nomina nach den formalen Kriterien beinhaltet im Deutschen u.a. den Aspekt ***Genuszuweisung***. Obwohl bekannt ist, daß die einzelnen Genera korreliert sind mit bestimmten semantischen Eigenschaften der Nomina (Spitz 1965, Wienold 1967, Fleischer 1975, Wellmann 1975, Beito 1976, Bechert 1982) wie auch mit gewissen morphologisch-phonetischen Eigenschaften (Helbig/Buscha 1972, Altmann/Raettig 1973, Zubin/Köpcke 1981, 1984), wird doch im allgemeinen das Genus als eine arbiträre und nicht eindeutig determinierbare Variable des Nomens betrachtet (Brinkmann 1962, Admoni 1970, Greenberg 1978, Eisenberg 1986).

Die Determinierbarkeit sprachlicher Größen aber ist - außer für lernpsychologische Zwecke - in der Sprachwissenschaft nur von sekundärer Bedeutung, da in der natürlichen Sprache nichts nach deterministischen Regeln geschieht.

Deshalb ist es nicht betrüblich, wenn scheinbar kein vollkommen konsistentes Beschreibungsschema für die Genuszuweisung erstellbar ist und die Frage, ob sich bestimmte sprachliche Eigenschaften in ein Schema einordnen lassen oder nicht, nicht immer beantwortet werden kann. Ein Schema spiegelt im Grunde lediglich unseren Versuch wider, Ordnung in die vielfältigen und teilweise recht heterogenen sprachlichen Zusammenhänge zu bringen; es ist ein abstraktes Konstrukt, anhand dessen wir die für theoretische und rechnerische Zwecke benötigten Klassenbildungen (Kategorisierungen) vornehmen können.

Eine 1982 von Köpcke angestellte umfangreiche Untersuchung zum Genussystem der deutschen Einsilber bildet die Grundlage für die Überlegungen und Analyse des vorliegenden Beitrags.

In Köpcke wird zum Genussystem ein Regelapparat entworfen, der auf phonetische, morphologische und semantische Eigenschaften der Nomina Bezug nimmt, wobei der Autor sich sehr wohl der Unschärfe der Genuszuweisung bewußt ist und entsprechend einräumt, daß er Sprache betrachtet "als ein stochastischen Regeln gehorchender Prozeß..., in dem as Auftreten verschiedener Zeichen oder deren Kombination die Wahrscheinlichkeit

des Auftretens bestimmter anderer Zeichen bedingt. Ein bestimmtes einsilbiges Nomen der Struktur K_1VK_2 besitzt aufgrund seiner phonologisch-morphologischen Struktur einen gewissen Wahrscheinlichkeitsgrad, etwa mit dem Genus Maskulinum verbunden zu werden, gleichzeitig aber auch eine sehr geringe Wahrscheinlichkeit – quasi als Komplementärfunktion –, das Femininum oder Neutrum zu evozieren" (S. 44).

Die entworfenen Regeln sind also ein Raster, unsere Abstraktion der Struktur des Genussystems, in dessen Zellen die einzelnen Genera mehr oder weniger stark ausgeprägt sind. Ausnahmen zu den Regeln, die Köpcke aufführt, werden deshalb im folgenden ebenso als Zellen des Genussystems betrachtet wie die "Regeln" selbst.

Zunächst wollen wir hier nur den semantischen Aspekt des Genussystems betrachten.

2. Im Deutschen verteilen sich die drei Genera (m,f,n) auf sämtliche Nomina. Die Nomina sind unterschiedlichen Wortfeldern zuzuordnen. Deshalb ist zu erwarten, daß die einzelnen Genera ebenfalls ein diverses Spektrum von semantischen Funktionen widerspiegeln, vorausgesetzt, man billigt ihnen semantische Eigenschaften zu.

Köpcke's Hauptthese ist "In der Genuszuweisung zu den einsilbigen Nomen der deutschen Gegenwartssprache existieren Regelmäßigkeiten, die der sprachlichen Kompetenz des native speakers angehören und deshalb als Lern-, Speicherungs- und Generierungsprinzipien aufgefaßt werden können. Weil diese Regelmäßigkeiten sich dem heutigen Sprecher/Hörer über die Gegenwartssprache vermitteln, werden sie über ein synchron ausgerichtetes Untersuchungsverfahren auszufinden sein" (S. 2), denn, um anzuschließen, die Sprechergemeinschaft strebt stets nach einer möglichst effizienten Ausformung der Systemstruktur, um ihre Kodierungs- , Speicherungs- und Dekodierungsbemühungen gering zu halten.

Um eine solche effiziente Strukturierung zu ermöglichen und zu erhalten, ist eine "Intelligenz", ein Mechanismus nötig, der die Prozesse der Sprache auslöst. Die beobachteten Daten in den Klassen (vgl. Tab.1.) reflektieren die aktuelle Struktur des Genussystems, der dahinterstehende Prozeß macht die Ausbreitung der Strukturelemente in den einzelnen Klassen (zu denen auch die 'Ausnahmen' zählen) in probabilister Weise vorhersagbar.

Die Analyse von Köpcke stützt sich auf das Korpus von Hirsch-Wierzbicka (1971), welches für phonetische Untersuchungen aus dem Leipziger Rechtschreib-Duden (1967) erstellt wurde. Aus der insgesamt 1466 einsilbige Nomina umfassenden Liste wurden bei der Auszählung allerdings nur die Nomina berücksichtigt, die sich in die von Köpcke gefundenen semantischen Klassen eingruppieren ließen. Die Klassenzahl ist also von vornherein begrenzt worden, und die Zählung beinhaltet nur 301 Nomina aus der Liste. Deshalb hat die nachfolgende quantitative Auswertung vorerst lediglich Pilotcharakter. Eine Gruppierung aller Nomina des Wörterbuchs dürfte die Hypothese stärken.

In Tabelle 1 sind die semantischen Felder der Nomina und die Häufigkeiten pro Feld nach den drei Genera aufgeschlüsselt aufgeführt. Die 10. Klasse (chem. Elemente und Metalle) wurde von mir selbst unterteilt, die 16. Kategorie (Seemannssprache) wurde zwar in Köpcke noch nicht – wie bei den übrigen gehandhabt – als Regel formuliert, wird hier aber selbstverständlich mit aufgenommen. Die Zahlen in Klammern stellten Ausnahmen der von Köpcke beschriebenen semantischen Regeln dar. Aus Tabelle 1 läßt sich schon gut erkennen, daß bestimmte Klassen von bestimmten Genera bevorzugt werden, andere wiederum sich ausschließen.

Zum Test, ob es tatsächlich signifikante Assoziationen bzw. Dissoziationen zwischen einem Genus und bestimmten semantischen Klassen der Nomina gibt, wurde die Tabelle als Kontingenztafel aufgefaßt und wie in Schulz/Altmann (1988:27ff) nach einem von Altmann/Lehfeldt (1980:295ff) entwickelten Ansatz qualitativ ausgewertet. Nach dem Modell, das u.a. für die Beurteilung von Vokalinteraktionen aufgestellt wurde, läßt sich nachweisen, ob eine signifikante Tendenz bestimmter Vokale besteht, mit anderen Vokalen kombiniert zu werden.

Auf die Assoziation zwischen Genus und semantischen Klassen der Nomina bezogen, läßt sich mit Hilfe des Modells zeigen, ob eine signifikante Tendenz besteht, daß z.B. das Genus m Nomina des semantischen Feldes "alkoholische Getränke" annimmt.

In Tabelle 2 ist die quantitative Auswertung anhand eines z-Tests nach Rehák, Reháková 1980 (vgl. Altmann/Lehfeldt S. 301) widergegeben. Der Test prüft, ob einzelne Zellen signifikant unter- bzw. überbelegt sind und wird berechnet mit der Formel

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

Tabelle 1. Verteilung der einsilbigen Nomina nach semantischen Klassen und Genus

semantische Felder der Nomina	m	f	n
1. Naturereignisse und ihre Benennung:			
a) natürl. Zeiteinheiten	5	(1)	-
b) Bez. f. Himmelsrichtungen	5	-	-
c) Bez. f. Winde u. Windarten	4	(1)	-
d) Niederschläge	4	-	-
2. Bez. f. Mineralien u. Gesteine	23	(1)	(1)
3. Bez. für Menschen u. Berufe	65	-	-
4. Bez. für alkohol. Getränke	11	-	(2)
5. Grundzahlen	-	12	-
6. Abkürzungen	-	3	-
7.Bez.für Wortarten	-	-	1
8. Nominalierungen ohne Ableitungsmorpheme	-	-	37
9. Bez. f. physik., theor. Einh.	-	-	18
10. Bez. für a) chem. Elemente	-	-	6
b) Metalle	-	-	4
11. Bez. für Sprachen	-	-	2
12. Bez. für Tonarten	-	-	12
13. Auf Menschen referieren (entspr. natürl. Geschlecht)	7	5	(1)
14. Auf domestizierte u. jagdbare Tiere refer., entspr. nat. G.	10	4	-
15. Auf Wasserflächen referierend	15	5	(3)
16. Seemannssprache	6	18	5

Tabelle 2. z-Test der Interaktionen zwischen Genus und den semantischen Feldern der zugehörigen einsilbigen Nomina

semantische Felder der Nomina	m	f	n
1. Naturereignisse und ihre Benennung:			
a) natürl. Zeiteinheiten	1.54	-0.01	-1.66
b) Bez. f. Himmelsrichtungen	2.16	-1.01	-1.51
c) Bez. f. Winde u. Windarten	1.26	0.19	-1.51
d) Niederschläge	1.93	-0.91	-1.35
2. Bez. f. Mineralien u. Gesteine	4.16	-1.79	-3.05
3. Bez. für Menschen u. Berufe	8.73	-4.10	-6.11
4. Bez. für alkohol. Getränke	2.39	-1.66	-1.24
5. Grundzahlen	-3.69	7.86	-2.37
6. Abkürzungen	-1.82	3.87	-1.17
7.Bez.für Wortarten	-1.05	-0.45	-1.50
8. Nominalierungen ohne Ableitungsmorpheme	-6.79	-2.93	9.70
9. Bez. f. physik., theor. Einh.	-4.57	-1.97	6.53
10. Bez. für a) chem. Elemente	-2.59	-1.11	3.69
b) Metalle	-2.10	-0.91	3.01
11. Bez. für Sprachen	-1.48	-0.64	2.12
12. Bez. für Tonarten	-3.69	-1.59	5.28
13. Auf Menschen referierend (entspr. natürl. Geschlecht)	0.12	2.13	-1.86
14. Auf domestizierte u. jagdbare Tiere refer., entspr. nat. G.	1.48	1.20	-2.57
15. Auf Wasserflächen referierend	1.30	0.65	-1.94
16. Seemannssprache	-3.57	6.85	-1.68

Tabelle 3. Interaktionen zwischen Genus und den semantischen Feldern der zugehörigen einsilbigen Nomina

semantische Felder der Nomina	m	f	n
1. Naturereignisse und ihre Benennung:			
a) natürl. Zeiteinheiten	A	A	V
b) Bez. f. Himmelsrichtungen	P	V	V
c) Bez. f. Winde u. Windarten	A	A	V
d) Niederschläge	A	V	V
2. Bez. f. Mineralien u. Gesteine	P	A	M
3. Bez. für Menschen u. Berufe	P	I	I
4. Bez. für alkohol. Getränke	P	V	A
5. Grundzahlen	I	P	I
6. Abkürzungen	V	P	V
7.Bez.für Wortarten	V	V	A
8. Nominalierungen ohne Ableitungsmorpheme	I	I	P
9. Bez. f. physik., theor. Einh.	I	I	P
10. Bez. für a) chem. Elemente	I	V	P
b) Metalle	I	V	P
11. Bez. für Sprachen	V	V	P
12. Bez. für Tonarten	I	V	P
13. Auf Menschen referierend (entspr. natürl. Geschlecht)	A	P	A
14. Auf domestizierte u. jagdbare Tiere refer., entspr. nat. G.	A	A	I
15. Auf Wasserflächen referierend	A	A	A
16. Seemannssprache	M	P	A

Die qualitative Bewertung der Quantile der Normalverteilung in Tabelle 2 erfolgt aufgrund des Existenzkriteriums ($n_{ij} = 0$ oder $n_{ij} > 0$) und der Signifikanz der Abweichung der beobachteten Zellenhäufigkeit von der theoretischen wie folgt:

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

3. Wenn demnach also ein bestimmtes Genus die Tendenz hätte, eine ihm (möglicherweise) inhärente Bedeutung zu erweitern, handelte es sich um eine Form von semantischer Diversifikation, wie sie bereits an zahlreichen sprachlichen Daten nachgewiesen wurde (vgl. Beöthy, Altmann 1984a,b und 1989, Altmann, Best, Kind 1987, Rothe 1986 für sprachliche Einheiten, sowie Rothe 1989 für grammatische Kategorien). Das hieße, das Genus neigt dazu, spontan neue Bedeutungen anzunehmen, indem es Nomina anderer semantischer Klassen zusätzlich zu den bisherigen präferiert bzw. früher präferierte irgendwann ablehnt.

Zwei Gründe sprechen allerdings gegen diese Interpretation. Zunächst ist - wie bereits erwähnt - bis heute umstritten, ob dem Genus überhaupt eine Bedeutung (im weiteren Sinne) zugeordnet werden kann oder ob man es lediglich als (formale) Paradigmenkategorie des Substantivs (Eisenberg 1986: 40) betrachten sollte.

Zweitens handelt es sich bei der Aufteilung einzelner Genera in semantische Klassen (sofern man solche annimmt) eher um eine Aufteilung als um eine Zerteilung, wie es bei der Diversifikation der Fall ist:

Bei der Diversifikation zerfällt ein Element oder auch eine Kategorie in mehrere Klassen , wobei die Klassen aus syntaktisch gleichen wie auch unterschiedlichen Feldern übernommen werden können. Die Konjunktion *und* z.B. kann die Funktion einer anderen Konjunktion wie *aber* übernehmen, sie kann aber auch Adverbialausdrücke ersetzen oder alternativ für eine Gerundialkonstruktion stehen, etc. Der Kasus *Genitiv* kann den Dativ ersetzen, ist aber z.T. auch funktional äquivalent mit präpositionalen Ausdrücken wie etwa mit *von, aus*, etc. Die Übernahme und der Austausch neuer Funktioenen findet also innerhalb des (Konjunktional- oder Kasus-) Systems wie auch außerhalb statt, und zwar innerhalb des gesamten Rahmens "Element plus seine funktionalen Äquivalente". Es handelt sich deshalb um ein offenes System.

Beim Genus ist das System geschlossen. Es beschränkt sich auf die Gesamtheit aller Substantive. Jedem Substantiv ist (bis auf die Fälle von Genusschwankungen und Genuswechsel) genau ein Genus zugeordnet. Es gibt hier also keine funktionalen Äquivalente, deshalb ist die Zahl der Klassen, in die das Genus sich aufteilen kann, geschlossen.

Im ersten Fall ("echte" Diversifikation) zerfließt eine sprachliche Entität in theoretisch unendlich viele Klassen, praktisch bis zu einem Grenzwert, der sich im Wettstreit zwischen Sprecher und Hörer automatisch einstellt. Im zweiten Fall gibt es von vornherein eine begrenzte Zahl von Klassen (Zahl der semantischen Felder, in die sich alle Nomina einordnen lassen), aus der sich jedes Genus einen Anteil auswählt, welchen es dann für sich neu verteilt. Ein Genus zerfließt also nicht in mehrere Funktionen, sondern es ordnet die Substantive nach deren Funktionen nach bestimmten Eigenschaften um, wobei es manche Funktionen bevorzugt, andere wiederum ablehnt.

Es ist also kein Zerfallsprozeß, sondern ein Umordnungsprozeß. Prinzip ist dabei nicht, zu diversifizieren, sondern zu vereinheitlichen, d.h. zu unifizieren. Die Funktionen des Genus stellen zwar auch ein diverses Feld dar, das Feld gruppiert sich aber nach einem Vereinheitlichungs-Prinzip.

Während im Falle der Diversifikation der Sprecher bevorzugt eine Einheit mit mehreren Funktionen belastet, um den Inventarumfang zu begrenzen, wirkt er bei der Verteilung der Genusfunktionen eher unifizierend, nämlich dahingehend, daß er bestimmte Funktionen für ein Genus privatisiert, um auf diese Weise Zuordnungsregeln für das Genus einzusparen.

Es ist deshalb anzunehmen, daß dieser Prozeß – ebenso wie der Diversifikationsprozeß – in einer Verteilung resultiert, die nach dem Urnenmodell aufgebaut ist: die einzelnen Klassen ziehen umsomehr Funktionen an, je gefüllter sie bereits sind bzw. lehnen umsomehr Funktionen ab, je leerer sie sind.

4. Aus den obigen Überlegungen gilt analog wie bei der Diversifikation, daß die Erwartungswerte der Klassen eine Funktion der Ränge sind. Da es sich außerdem um eine Art "rückläufige" Diversifikation (Unifikationsbestreben in einem diversen Feld) handelt, dürfte die Verteilung mit der eines Diversifikationsmodells verwandt ist.

Als Modell wurde die "Dolinskij-Verteilung" (Dolinskij 1988) gewählt, mit der Formel:

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

($\alpha^* = f_1/N$, T = Normierungsgröße), die man linguistisch und psychologisch interpretieren kann und die im linguistischen Bereich u.a. auch für den Diversifikationsprozeß als Modell verwandt wurde (Altmann 1989). Eine Einbindung der Verteilung in das synergetische Modell ist ebenfalls möglich (vgl. Hammerl 1989).

Die Ergebnisse sind in Tabellen 4-6 widergegeben.

In den Tabellen (4-6) ist in der ersten Spalte die Schlüsselzahl der jeweiligen Klasse abzulesen, in der zweiten Spalte der Rang (nach Häufigkeit geordnet), in der dritten Spalte die beobachtete Häufigkeit, in der vierten Spalte die erwartete (theoretische) Häufigkeit. Der untere Block beinhaltet die Werte der beiden Parameter k und p, die Anzahl der Freiheitsgerade (f), den Chiquadrat-Wert und die Wahrscheinlichkeit (P, geforderte Untergrenze: Prob = 0.05) für das ermittelte Chi-Quadrat.

Wie man sieht, ergibt sich bei allen drei Genera trotz der oben angesprochenen Vorauslese der Datenkonfiguration ein recht gutes Resultat.

Interessanterweise sind gerade in den Gruppen, wo Köpcke in mehreren Fällen die Klassenbelegungen als Ausnahmen (siehe die geklammerten Klassen in der ersten Spalte) verstanden haben möchte, die Anpassungen besonders gut sind, nämlich bei den Nomina mit femininem Genus und bei den Neutra. Das bedeutet, daß die probabilistische Modellierung des Prozesses auch dort oder besser, eben dort greift, wo sich die Daten dank ihrer Variabilität eindeutig festlegbaren deterministischen Regeln verschließen.

5. Vorausblickend kann man schon jetzt davon ausgehen, daß eine umfangreichere Beschreibung des Genussystems unter semantischem Aspekt das Modell wiederum bestätigen dürfte. Dabei würde sich die Klassenzahl erheblich vergrößern, theoretisch unendlich, da für jedes Nomen ein semantischer Index angesetzt würde.

Die Untersuchung müßte dann auch Mehrsilber einschließen, selbst wenn – wie Köpcke andeutet, die von ihm aufgestellten semantischen Regeln in gleicher Weise für Ein- und Mehrsilber gelten. Denn bei Mehrsilbern sind vor allem morphologische Aspekte an der Bedeutung der Nomina beteiligt, die dann selbstverständlich mit berücksichtigt werden müßten.

Tabelle 4
Verteilung der einsilbigen maskulinen Nomina über die semantischen Klassen nach Köpcke (1982)

Klasse	x	n(x)	NP(x)
3	1	65	65.00
2	2	23	22.93
15	3	15	15.25
4	4	11	11.31
14	5	10	8.94
13	6	7	7.35
16	7	6	6.21
1a	8	5	5.36
1d	9	5	4.71
1c	10	4	4.18
1b	11	4	3.76

a = 0.9336
b = 0.0412
f = 7
chi² = 0.23
P = 0.99996

Tabelle 5
Verteilung der einsilbigen femininen Nomina über die semantischen Klassen nach Köpcke (1982)

Klasse	x	n(x)	NP(x)
16	1	18	18.00
5	2	12	11.46
13	3	5	6.54
15	4	5	4.30
14	5	4	3.05
6	6	3	2.28
(2)	7	1	1.77
(1a)	8	1	1.42
(1c)	9	1	1.16

a = 1.1417
b = 0.1319
f = 5
chi² = 1.52
P = 0.91127

Tabelle 6
Verteilung der einsilbigen Nomina im Neutrum über die semantischen Klassen nach Köpcke (1982)

Klasse	Rang	n(X)	NP(X)
8	1	37	37.00
9	2	18	17.92
12	3	12	10.98
10A	4	6	7.44
16	5	5	5.38
10B	6	4	4.06
14	7	4	3.17
15	8	3	2.53
11	9	2	2.07
(4)	10	2	1.72
7	11	1	1.45
(2)	12	1	1.23
(13)	13	1	1.06

a = 0.8393
b = 0.2061
f = 9
chi2 = 0.94
P = 0.99996

Literatur

Admoni, W. (1970:3), Der deutsche Sprachbau. München: Beck'sche Verlagsbuchhandlung.

Altmann, G. (1989), Two models for word association data. In: Köhler, R. (Ed.), *Studies in Language Synergetics* (erscheint).

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

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

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

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

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

Dolinskij, V.A. (1988), Raspredelenie reakcij v experimentach po verbal'nym associacijam. *Acta et Commentationes Universitatis Tartuensis 827, 89–101.*

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

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

Greenberg, J.H. (1979), How does a language acquire gender markers? In: Greenberg, J.H. (Ed.), *Universals of Human Language, Vol. 3: Word Structure.* Stanford 1979, 47–82.

Hammerl, R. (1989), Untersuchungen zur Verteilung der Wortarten im Text. *In diesem Band.*

Hirsch–Wierzbicka, L. (1971), Funktionelle Belastung und Phonemkombination am Beispiel einsilbiger Wörter der deutschen Gegenwartssprache. Hamburg, Buske.

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

Köpcke, K.–M. (1982), Untersuchungen zum Genussystem der deutschen Gegenwartssprache. Tübingen, Niemeyer.

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

Rothe, U. (1989), Diversification processes in grammar. An introduction. In: Rothe, U. (Ed.), *Diversification processes in language: grammar* (erscheint).

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

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

Hřebíček, L. (ed.), Glottometrika 11, 1989

Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen

K.-H. Best, Göttingen

1. Am Beispiel der Wortbildungsaffixe der deutschen Sprache der Gegenwart konnte gezeigt werden, daß die Ranghäufigkeitsverteilung der Bedeutungsklassen einzelner Affixe der negativen Binomialverteilung bzw., falls eine Ableitung aus verschiedenen Basiswortarten vorliegt, der gemischten negativen Binomialverteilung folgt (vgl. Altmann, Best, Kind 1987).

2. Inzwischen wurden im Erlanger Forschungsprojekt "Nürnberger Frühneuhochdeutsch um 1500" entsprechende Daten für den fnhd. Wortbildungstyp ***g(e)–ø/(t/d) + (e)*** (glid, glenck,...) erarbeitet. Es bietet sich an, auch diese Daten daraufhin zu überprüfen, ob sie dem oben genannten Modell folgen.

Es wurden sämtliche einschlägigen Belege aus dem handschriftlichen Nachlaß Albrecht Dürers erfaßt, nicht aber aus den Texten Dürers, die nur durch wesentlich spätere Druckfassungen bekannt sind (Habermann, Müller 1987: 123, 126). Alle Belege des obigen Wortbildungsmusters sind unter Angabe ihrer absoluten Häufigkeit aufgeführt und verschiedenen Funktionsständen zugewiesen worden (Habermann, Müller 1987: 128, 130). Außer dem Wort *gesicht* gehören alle anderen jeweils nur einem Funktionsstand an. Für *gesicht*$_1$ (Nomen patientis) wurde 1 Beleg gefunden, für *gesicht*$_2$ (Abstraktbildung) 42 Belege und für *gesicht*$_3$ (idiomatische Bildung) 3 Belege (Habermann, Müller, persönliche Mitteilung).

Die Belege verteilen sich folgendermaßen auf die Funktionsstände (vgl. Tabelle 1):

Tabelle 1

Die Verteilung der Belege des fnhd. Wortbildungsmusters g(e)–ø/(t/d)+(e) auf die Funktionsstände

(nach Habermann, Müller 1987:128,130)

1. ***Kollektiva***: 11

geeder (= Geäder), gefrüst (= Fröste), gemeüer, gemües, gepein, gepent (= Gebände), gepirg, geschtiren (= Gestirn), gestull (= Dachgestühl), getatten, getüll (= Wülste des Stirnbeins)

2. ***Nomina patientis***: 11 (10)

gedancken, gemell (= Gemälde), gepet, gepeü (= Gebäude), geprech (= Gepräge), geschöpf, gesetz, gesicht$_1$ (= Traumbild)[1)], getrangk, gewelb (= Gewölbe), gwind

3. ***Abstrakta***: 6 (5)

gedult, gericht, gescheft (= Anordnung), gesicht$_2$ (= das Sehen), gsang, getreng (= Gedränge)

4. ***Nomina agentis***: 1

geheng

5. ***Nomina instrumenti***: 1

geschos

6. ***Bildungen mit pleonastischem Affix***: 6

gemüt, gepildnus, geschrift, gespott, glid, gsims

7. ***Morphologisch motivierte, semantisch idiomatisierte Bildungen***: 12

gemach, gemecht (= männliches Geschlechtsteil), geplütz, geschlecht, geschütz, gesellen, gesicht$_3$ (= Angesicht), gewalt, gewand, gewicht, glenck, gwag (= Gewäge)

8. ***Simplicia***: 4

genick, gesind, gespenst, gnaden

1) Zu *gesicht*$_{1,2,3}$ vgl. Habermann, Müller (1987: 131 f.).

3. Bei der Anpassung eines Modells ergeben sich hier einige Schwierigkeiten:

(a) Während in Altmann, Best, Kind (1987) die Gruppierung aufgrund einer Transformation oder Paraphrasierung eines Derivats getestet wurde, spielt hier nur das abstrakte Endresultat, das weiter nicht analysiert wird, eine Rolle. Daher sind diese Daten etwas vergröbert.

(b) In einem Falle (*gesicht*) diversifiziert eine Entität in drei Funktionsstände gleichzeitig, wodurch eine leichte Verzerrung entsteht. Bei multidimensionaler Modellierung wäre diesem Umstand Rechnung getragen; hier werden wir in zwei Alternativen rechnen.

Die Häufigkeitverteilung der Funktionsstände ergibt folgendes Resultat:

(i) Wenn *gesicht* in allen Klassen belassen wird, dann ist

x	1	2	3	4	5	6	7	8
f_x	12	11	11	6	6	4	1	1

(ii) Wenn *gesicht* nur in der häufigsten Klasse belassen wird, dann ist

x	1	2	3	4	5	6	7	8
f_x	12	11	10	6	5	4	1	1

Da die Basiswortart nicht einheitlich ist, sondern sowohl Substantiv als auch Verb sein kann, handelt es sich nach unseren Annahmen um eine gemischte negative Binomialverteilung (konventionsgemäß auf den Anfang x = 1 verschoben). Die optimierte Anpassung zeigte in beiden Fällen in der Tat, daß nur die gemischte negative Binomialverteilung (mit 4 Parametern) monoton fallend ist, während die *beste* Anpassung der einfachen und der gestutzten negativen Binomialverteilung einen nicht monoton fallenden Verlauf aufweist, wobei allerdings die resultierenden Irrtumswahrscheinlichkeiten besser sind.

Die Resultate der Anpassung sind in Tabelle 2 aufgeführt.

Tabelle 2

Anpassung der gemischten negativen Binomialverteilung an die Daten

x	fx	NPx	fx	NPx
1	12	16.56	12	16.16
2	11	12.53	11	12.05
3	11	8.43	10	8.05
4	6	5.44	6	5.17
5	6	3.44	5	3.26
6	4	2.15	4	2.03
7	1	1.33	1	1.26
8	1	2.12	1	2.02
	$k = 1.2943$			1.2722
	$p_1 = 0.3956$			0.3929
	$p_2 = 0.4565$			0.4551
	$\alpha = 0.7166$			0.7027
	$X^2 = 6.38$			5.17
	$FG = 3$			3
	$P = 0.09$			0.16

Wie man sieht, ist die Anpassung nach der "Bereinigung" (= Eliminierung von *gesicht*$_2$ und *gesicht*$_3$) beträchtlich besser. Man kann also vorläufig annehmen, daß das ursprüngliche Modell adäquat ist.

Literatur

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

Habermann, M., ***Müller***, P.O. (1987), Zur Wortbildung bei Albrecht Dürer. Zeitschrift für deutsche Philologie 106. Sonderheft: Frühneuhochdeutsch, 117-137.

Hřebíček, L. (ed.), Glottometrika 11, 1989

Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina

Ursula Rothe, Bochum

1. Die vorliegende Untersuchung versteht sich als ein Beitrag zur Diversifikation von Wortbildungsmustern nach bestimmten semantischen Bildungskriterien. Der Beitrag stützt sich auf Ergebnisse, die unter dem Titel "Deutsche denominale Verben" von Vl. D. Kaliuščenko (1988) veröffentlicht wurden.

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

Die Verbalisierungen aus den Nomina unterliegen unterschiedlichen semantischen Motivierungsbeziehungen. Dies nicht zuletzt dadurch, daß durch die verkürzte verbalisierte Wendung viele verschiedenste, ansonsten umständlich zu formulierende syntaktische Beziehungen ökonomisch ausgedrückt werden (vgl. Erben 1975:69f). Kaliuščenko ermittelte insgesamt 29 semantische Klassen, die das Verhältnis zwischen motivierendem Nomen und Verb-Neubildung widerspiegeln.

Zunächst sollen hier nur die semantischen Motivierungsbeziehungen der Präfixe an der Wortbildung betrachtet werden, denn sie dürften - im Gegensatz zu den Suffixen - den "typischeren" Fall der Diversifikation aufzeigen. Präfixableitungen nämlich sind im heutigen Deutsch wesentlich produktiver als Suffixableitungen, sie werden demnach zahlreicher verwendet. Wie Köhler (1986) bereits herausstellte, besteht eine positive Korrelation zwischen Texthäufigkeit einer sprachlichen Entität und deren Diversifikationstendenzen. Bei Köhler wird zwar zunächst von Lexemen ausgegangen, die Tendenzen sind jedoch auf allen sprachlichen Ebenen vorhanden und deshalb auch bei Morphemen aufzufinden.

2. Die Diversifikation ist ein sprachinhärenter Prozeß, der bewirkt, daß sprachliche Einheiten Varianten bilden, mit neuen Funktionen ver-

sehen werden oder in mehrere Klassen eines Attributsraums eindringen. Solche Klassen können Bedeutungen, grammatische und syntaktische Eigenschaften, kategoriale Zuordnungen, etc. sein (vgl. Altmann 1987; Rothe 1989).

Eine Erklärung des Diversifikationsprozesses in der Sprache liefert das Zipf'sche Prinzip der geringsten Anstrengung (Zipf 1949), demgemäß eine Konkurrenz besteht zwischen den Bedürfnissen von Hörer und Sprecher: der Sprecher artikuliert sich am ökonomischsten, wenn er möglichst wenig Entitäten mit je vielen Funktionen benutzen kann, der Hörer rezipiert andererseits am ökonomischsten, wenn für jede Funktion eine eigene sprachliche Einheit zur Verfügung steht.

Aufgrund der Ökonomiebestrebungen kommt es z.B. dazu, daß die häufigsten Wörter (durchschnittlich) die kürzesten sind, daß diese wiederum die meisten Bedeutungen haben und häufiger in unterschiedlichen Sprachschichten und -stils vertreten sind bzw. in Texten unterschiedlicher Gattung (Polytextie) als längere mit wenig Bedeutungen (vgl. Köhler 1986).

Eine sinnvolle Kommunikation kommt aber nur zustande, wenn sowohl die Bedürfnisse des Sprechers als auch die des Hörers etwa gleich stark vertreten werden, sodaß die Sprache nicht aus dem Gleichgewicht gerät. Ein leichtes Kräfte-Ungleichgewicht ist allerdings immer vorhanden, es garantiert sprachliche Variabilität und dadurch den notwendigen Fortschritt der sprachlichen Entwicklung.

3. Die nachfolgend übernommene semantische Kategorisierung der Verben wurde von Kaliuščenko durchgeführt, indem "Deutungsformeln" entwickelt wurden, die – im Gegensatz zu einer reinen Bedeutungsbeschreibung des Verbs – hauptsächlich Bezug nehmen auf das motivierende Nomen und dadurch einen allgemeineren Charakter haben, als die lexikographische Beschreibung.

Beispielsweise würden die Verben *sägen, brausen, bürsten* lexikographisch unterschiedlich behandelt (vgl. Kaliuščenko 1988:21):

sägen: "etw. mit einer Säge zerschneiden"
bremsen: "die Geschwindigkeit von etw. vermindern"
bürsten: 1. "etw. mit einer Bürste von etw. entfernen"
2. "etw. mit einer Bürste bearbeiten" ,

wohingegen die Deutungsformel die Verben aufgrund gleicher Motivierungsbeziehung einheitlich behandelt:

S_1 wirkt auf S_2 mit Hilfe von S_m .

(S_1 und S_2 sind Argumente unterschiedlicher Satzgliedpositionen, S_m ist motivierendes Nomen der Verbalisierung). Die Nomina können unterschiedliche semantische Kasusrollen (Tiefenkasus, vgl. Fillmore 1968 u.a.) annehmen, die in der Metasprache von Kaliuscenkos Deutungsformel allerdings nicht einzeln benannt werden.

Insgesamt gelangt Kaliuščenko zu den in Tabelle 1 widergegebenen 29 semantischen Klassen (vgl. S. 114.

4. Untersuchungen zur semantischen Diversifikation von Affixen wurden bereits in verschiedenen Untersuchungen gemacht (zu ungarischen Präfixen vgl. Beöthy/Altmann 1984a, 1984b, 1989, zur deutschen Wortbildung im Althochdeutschen vgl. Best 1988).

Die letztgenannten Arbeiten untersuchten jeweils ausgewählte Entitäten in Hinblick auf ihre Diversifikation nach ihren unterschiedlichen Bedeutungen im lexikographischen Sinne.

Im vorliegenden Beitrag ist nicht die Frage, nach welchen Bedeutungen ein Präfix diversifiziert, sondern, nach welchen semantischen Motiven Affixe bei denominalen Verbalisierungen diversifizieren. Es ist also eine Frage, die nicht den aktuellen (semantischen) Status des Affixes, sondern direkt seinen Beitrag bei der Entstehung einer Neubildung betrifft.

Will man hierbei die Bedürfnisse der Sprachgemeinschaft, die ja durch ihren Druck den Diversifikationsprozeß in Gang hält, interpretieren, so lautet im vorliegenden Fall die Frage, in welchem Maß der Sprecher Druck auf das Sprachkonstrukt "denomiales Ableitungspräfix" ausübt, damit es möglichst viele Kontexte widergibt, die die Relation zwischen motivierendem Nomen und dem neu entstandenen Verb widergeben. Deshalb ist für *-ier* nicht nur der Typ "S_1 nimmt S_2 den Gegenstand S_m" (*ausbeulen)* mit 24 Vorkommen häufig, sondern der Typ "S_1 versieht S_2 mit S_m" (*auszinnen)* ist mit 10 Vorkommen ebenfalls recht oft vertreten, ebenso wie der Typ "S_1 wird zu S_m" (*ausfasern)* mit auch 10 Vorkommen.

Der Sprecher übt also (in Konkurrenz mit dem unifizierenden Gegendruck des Hörers) einen (diversifizierenden) Druck auf die Verteilung der Deutungsformeln des Präfixes aus. Das Resultat ist die aktuelle Verteilung der Präfixe nach den Deutungsformeln (hier des Nhd.).

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

Klasse	Deutungsformel	Beispiele
(1)	S_1 ist (wie) S_m	*gärtnern, wurmen, spiegeln*
(2)	S_1 wird zu S_m	*verdummteufeln, sich vernarren*
(3)	S_1 macht S_2 zu S_m	*adeln, mopsen, kapitalisieren*
(4)	S_m erscheint bei/ auf S_1	*verlausen*
(5)	S_1 verliert S_m	*nadeln*
(6)	S_1 versieht S_2 mit S_m	*grundieren, verursachen*
(7)	S_1 nimmt S_2 den Gegenstand S_m	*köpfen, entnerven*
(8)	S_1 schafft S_m	*bahnen*
(9)	S_1 handelt typisch an S_m	*wildern*
(10)	S_1 schafft ein S_m des Gegenstands S_2	*formen*
(11)	S_1 wirkt auf S_m = Teil von S_2	*blättern*
(12)	S_1 stellt S_2 aus S_m her	*knibbeln*
(13)	S_1 wirkt auf S_2 mit Hilfe von S_m	*bürsten*
(14)	S_1 handelt mit Hilfe von S_m	*ankern*
(15)	S_1 führt eine Handlg. S_m an S_2 aus	*alarmieren*
(16)	S_1 führt eine Handlg. S_m aus	*pokern*
(17)	S_1 befindet sich im Zustand S_m	*verunglücken*
(18)	S_1 löst bei S_2 ein Gefühl S_m aus	*verseuchen*
(19)	S_1 sagt S_m (zu einer Person S_2)	*beichten*
(20)	S_m = Beziehung (zw. S_1 und S_2)	*reimen*
(21)	S_1 befindet sich am Ort S_m	*zelten*
(22)	S_1 gerät an den Ort S_m	*sich einschiffen*
(23)	S_1 entfernt sich von S_m	*ausufern*
(24)	S_1 plaziert S_2 am Ort S_m	*zetteln*
(25)	S_1 entfernt S_2 aus S_m	*ausstopfen*
(26)	S_m = Art und Weise der Handlg. v. S_1 an S_2	*pachten*
(27)	S_m = Zeit ("Periode")	*übersommern*
(28)	S_m = Situation	*tagen*
(29)	(Vereinzelte denom. Verben)	*akademisieren, zetern, basteln*

5. In früheren Untersuchungen wurde der Diversifikationsprozeß stets mit dem Modell der negativen Binomialverteilung getestet (vgl. Beöthy/ Altmann passim, Rothe 1989), wobei sich sehr gute Anpassungen ergaben. Das Modell wurde aus einem Differentialansatz hergeleitet, der sich direkt mit dem synergetischen Modell der konkurrierenden Bedürfnisse von Sprecher und Hörer interpretieren ließ.

Aufgrund neuerer Erkenntnisse (vgl. Altmann 1989, Hammerl 1989) hat sich eine Variante des Modells ergeben, die sog. Zipf-Dolinskij-Verteilung, die z.T. sehr gute Anpassungen ergab:

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

wo T die Normierungsgröße und α, a und b Parameter sind. Das Modell läßt sich - ausgehend von Überlegungen aus dem Fechnerschen Gesetz der Abhängigkeit von Reaktionen auf physikalische Reize - ebenfalls in das synergetische Modell einbinden (vgl. Hammerl 1989) und scheint nach den bisherigen Erfahrungen insbesondere dann geeigneter zu sein, wenn die monotone Abnahme der Rang-Häufigkeiten nicht sehr steil verläuft. Dies ist gerade im unteren Teil der Verteilung typisch, wo nämlich viele unterschiedliche Klassen vertreten sind, die aber relativ gering belegt sind. Ebenfalls gute Anpassungen ergeben sich, wenn die Klassenzahl a priori festgelegt oder relativ gering ist, wie es etwa bei der Verteilung der Wortarten der Fall ist.

Da auch im vorliegenden Fall die Klassenzahl nicht sehr groß ist (es gibt insgesamt 29 semantische Klassen, aber das Präfix *be-* mit den meisten Gruppen bildet nur aus 16 von ihnen Verbalableitungen, siehe Tab.4), wurde hier ebenfalls das Modell der "Zipf-Dolinskij-Verteilung" gewählt.

Die Tabellen 2-7 beinhalten die entsprechenden Berechnungen für die Präfixe *ab-, aus-, ein-, ent-, ver-* für das Nhd. nach Kaliuščenko (1988:113-115).

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

1. die Schlüsselzahl der Deutungsformel (Bed.), gemäß Tab.1.
2. die Ränge der Klassen x, nach abnehmender Häufigkeit geordnet

3. die beobachteten Häufigkeiten n(x) der einzelnen Klassen
4. die berechneten Häufigkeiten Np(x)

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

1. und 2.: die Parameter a bzw. b der Verteilung
3. den Chi² - Wert mit der Zahl der Freiheitsgrade in Klammern
4. die Wahrscheinlichkeit P, mit der ein so kleines oder ein noch kleineres Chi² zu erwarten ist .

6. Wie man sieht, ergeben die Berechnungen für alle Präfixe eine sehr gute Anpassung. Allerdings fällt in 3 Fällen der Wert für den Parameter a so gering aus, daß man ihn vernachlässigen könnte. Es handelt sich hierbei um die Präfixe *aus-*, *ein-* und *ver-*.

Der Diversifikationsprozeß, der die Aufteilung der Präfixe in unterschiedliche Deutungsformeln bewirkt, ist selbstverständlich nicht gleichzusetzen mit dem historischen Prozeß. Semantischer Wandel mit seinen Bedeutungserweiterungen, und -verengungen gehört zwar zu den Erscheinungen, die auf sprachgeschichtlicher Ebene durch den Diversifikationsprozeß bewirkt werden; er ist aber deshalb nicht der Mechanismus selbst, sondern nur eines seiner vielfältigen Symptome.

Deshalb ist die häufigste Deutungsformel eines Präfixes nicht zwingend zugleich die älteste und ehemals einzige. Vielmehr ist die häufigste Deutungsformel diejenige, zu deren Gunsten im Wettstreit Sprecher/Hörer zu einem gegebenen Zeitpunkt am wahrscheinlichsten eine Zuordnung für das Präfix erfolgt.

Die Rang-Häufigkeitsverteilung drückt aus, in welchem Maße es der Sprechergemeinschaft gelungen ist, eine für die Kommunikation optimale Zuordnung der semantischen Klassen zu dem Präfix zu erreichen. Entscheidend dabei ist, daß wir annehmen, daß der Sprecher die Differenz zwischen der häufigsten und der zweithäufigsten Klasse, der zweithäufigsten und der dritthäufigsten Klasse, etc. "kennt".

Der Sprecher möchte den Kodierungsaufwand beim Speichern eines Präfixes und seines semantischen Feldes gering halten, d.h. er möchte möglichst viele Deutungsformeln mit *einem* Präfix abdecken. Der Hörer hingegen bremst dieses Bestreben, er möchte möglichst nur *eine* Deutungsformel für das Präfix. Je mehr Druck der Hörer ausübt, desto steiler wird der Kurvenabfall und desto gefüllter werden die häufiger belegten Klassen. Je mehr Druck der Sprecher ausübt, desto flacher wird die Ver-

Tabelle 2
Verteilung des Präfixes ***ab-***
nach den Deutungsformeln
der zugrundeliegenden
denomialen Verben

Bed.	x	n(x)	NP(x)
7	1	16	16.00
25	2	7	6.43
3	3	3	3.79
13	4	3	2.60
29	5	2	1.95
1	6	1	1.53
6	7	1	1.25
11	8	1	1.05
14	9	1	0.90
17	10	1	0.79
26	11	1	0.69

a = 1.3018
b = 0.0012
chi²(5) = 0.70
P = 0.98

Tabelle 3
Verteilung des Präfixes ***aus-***
nach den Deutungsformeln der
zugrundeliegenden
denomialen Verben

Bed.	x	n(x)	NP(x)
7	1	24	24.00
2	2	10	11.72
6	3	10	9.10
25	4	9	7.09
13	5	5	5.61
5	6	4	4.52
27	7	4	3.69
23	8	3	3.06
29	9	3	2.57
3	10	2	2.18
11	11	2	1.86
26	12	1	1.60

a = 0.0007
b = 0.3490
chi²(5) = 1.34
P = 0.99

Tabelle 4
Verteilung des Präfixes ***be-***
nach den Deutungsformeln
der zugrundeliegenden
denomialen Verben

Bed.	x	n(x)	NP(x)
6	1	86	86.00
4	2	13	14.61
1	3	10	9.14
3	4	8	6.32
18	5	6	4.66
2	6	3	3.58
13	7	2	2.84
15	8	2	2.31
29	9	2	1.91
7	10	1	1.61
11	11	1	1.37
17	12	1	1.19
19	13	1	1.03
24	14	1	0.91
26	15	1	0.80
27	16	1	0.71

a = 0.8399
b = 0.1768
chi²(10) = 2.00
P = 0.996

Tabelle 5
Verteilung des Präfixes ***ein-***
nach den Deutungsformeln
der zugrundeliegenden
denomialen Verben

Bed.	x	n(x)	NP(x)
24	1	18	18.00
3	2	4	4.46
27	3	4	3.54
6	4	3	2.82
2	5	2	2.29
22	6	2	1.88
29	7	2	1.57
12	8	1	1.32
13	9	1	1.13

a = 0.0018
b = 0.3156
chi²(5) = 0.37
P = 0.996

Tabelle 6
Verteilung des Präfixes *ent–* nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
7	1	71	71.00
3	2	12	10.88
5	3	3	4.67
23	4	3	2.83
25	5	3	2.02
13	6	1	1.60

a = 2.9623
b = -0.4895
$chi^2(2) = 1.42$
P = 0.49

Tabelle 7
Verteilung des Präfixes *ver–* nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
2	1	42	42.00
3	2	40	41.92
6	3	32	26.71
4	4	17	17.24
1	5	9	11.48
29	6	7	7.87
13	7	4	5.55
18	8	4	4.01
26	9	4	2.96
15	10	3	2.22
17	11	2	1.70
21	12	1	1.31
22	13	1	1.03

a = 0.0001
b = 0.3676
$chi^2(9) = 2.33$
P = 0.985

teilungskurve und desto gleichmäßiger verteilen sich die Klassenhäufigkeiten. M.a.W., ist die Diversifikationskraft des Sprechers sehr stark, so hat ein Präfix viele Deutungsformeln mit durchschnittlich wenigen Häufigkeiten (z.B. *be–*), ist der Unifikationsdruck des Hörers sehr stark, so hat das Präfix wenig Deutungsformeln mit durchschnittlich vielen Häufigkeiten (z.B. *ent–*).

7. Zum Vergleich wurden einige Verteilungen für das Mhd. (vgl. Kaliuščenko 1988: 119–120) getestet, um den universellen Charakter des Modells zu zeigen. Denn wenn das Gesetz, das dem Diversifikationsprozeß zugrundeliegt gilt, dann gilt es in jeder Sprachstufe. Die Häufigkeiten der einzelnen semantischen Gruppen mögen zwar für ein Präfix im Nhd. anders belegt sein als im Mhd., aber die Verteilung der Häufigkeiten selbst nach den Rängen folgt demselben Prozeß in beiden Sprachstufen.

Allerdings waren aus den Daten für das Mhd. nur noch 3 Präfixe (*be–, ent–, ver–*) für quantitative Auswertungen ausreichend häufig vertreten. Die Ergebnisse sind in Tabelle 8–10 festgehalten.

Die belegten Häufigkeiten in Kaliuščenko zeigen, daß im Mhd. andere Präfixe (für die aufgeführten Deutungsformeln) produktiv waren als im Nhd. (z.B. *aus* mit 77 von 2815 Belegen im Nhd., im Gegensatz zu nur 1 aus 1352 Belegen im Mhd., vgl. S. 114 bzw. 119), was zugleich darauf

hinweist, daß keine direkte Analogie zwischen historischer Entwicklung und dem Diversifikationsprozeß besteht.

Tabelle 8
Verteilung des Präfixes *be–* im Mhd. nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
6	1	36	36.00
3	2	4	4.78
1	3	3	3.12
13	4	3	2.41
11	5	2	2.03
15	6	2	1.79
24	7	2	1.64
26	8	2	1.52
27	9	2	1.44
2	10	1	1.38
10	11	1	1.33
18	12	1	1.29
29	13	1	1.26

a = 1.4696
b = -0.2327
$chi^2(9) = 1.05$
P = 0.999

Tabelle 9
Verteilung des Präfixes *ent–* im Mhd. nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
7	1	46	46.00
5	2	5	5.03
25	3	3	3.00
3	4	2	1.82
2	5	1	1.15

a = 0.1221
b = 0.6454
$chi^2(1) = 0.04$
P = 0.85

Tabelle 10
Verteilung des Präfixes *ver–* im Mhd. nach den Deutungsformeln der zugrundeliegenden denomialen Verben

Bed.	x	n(x)	NP(x)
6	1	14	14.00
3	2	10	8.89
13	3	4	5.08
4	4	3	3.53
29	5	3	2.71
2	6	2	2.21
7	7	2	1.87
26	8	2	1.64
27	9	2	1.46
1	10	1	1.32
5	11	1	1.21
11	12	1	1.12
15	13	1	1.05
19	14	1	0.98
24	15	1	0.93

a = 1.6662
b = -0.1607
$chi^2(10) = 0.92$
P = 0.99988

8. Bei allen Berechnungen sowohl für das Nhd. als auch für das Mhd. ergaben sich sehr gute Anpassungen. Der Parameter a allerdings, der im Nhd. in 3 von 6 Fällen einen so geringen Wert hatte, daß man ihn vernachlässigen konnte, scheint im Mhd. für die Präfix-Deutungsformel-Zuorndung (nach Kaliuščenko) eine wichtigere Rolle zu spielen.

Der Diversifikationsdruck, der ja vom Sprecher bewirkt wird, war möglicherweise im Mhd., als die Produktivität der Präfixe in der Entwicklung war, größer als im Nhd. Die Ausdehnung auf die vielen semantischen Gruppen konnte noch im größeren Umfang erfolgen. Das Feld der durch die Deutungsformeln widergegebenen semantischen Gruppen hätte sich demnach allmählich aufgefüllt, die Dynamik durch den Diversifikationsdruck hätte nachgelassen.

Das Diversifikationsgesetz – wirksam in allen Sprachstufen – wird durch das Modell der Verteilung ausgedrückt. Der Prozeß wird sichtbar in den einzelnen Parametern. Geht man davon aus, daß das Gesetz tatsächlich durch die gewählte Verteilung weitgehend treffend repräsentiert wird, dann lassen sich Schwächungen und Stärkungen im Prozeß durch die Parameter interpretieren.

Literatur

Altmann, G. (1989a), Diversification processes of the word. In: Köhler, R. (ed.), *Studies in Language Synergetics* (erscheint)

Altmann, G. (1989b), Two models for word association data. In: Köhler, R. (ed.), *Studies in Language Synergetics.* 1989 (erscheint)

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

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

Beõthy, E., Altmann, G. (1989), Semantic diversification of Hungarian verbal prefixes. I. "meg-". In: Rothe, U. (ed.), *Diversification Processes in Language: Grammar.* 1989 (erscheint).

Best, K.-H. (1989), Zur Diversifikation eines Wortbildungsmusters im Mhd. *In diesem Band.*

Dolinskij, V.A. (1988), Raspredelenie reakcij v experimentach po verbal'nym associacijam. *Acta et Commentationes Universitatis Tartuensis 827, 89-101.*

Erben, J. (1975), Einführung in die deutsche Wortbildungslehre. Berlin: Schmidt 1975.

Fillmore, Ch. (1968), The case for case. In: Bach, E., Harms, R.T. (Eds.), *Universals in Linguistic Theory.* New York, Holt, Rinehart and Winston.

Hammerl, R. (1989), Untersuchungen zur Verteilung der Wortarten im Text. *In diesem Band.*

Kaliuščenko, Vl.D. (1988), Deutsche denominale Verben. Tübingen, Narr.

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

Rothe, U. (1989) Diversification Processes in Language. An Introduction. In: Rothe, U. (Ed.), *Diversification Processes in Language: Grammar* (erscheint).

Zipf, G. K. (1972:2), Human Behavior and the Principle of Least Effort. New York, Hafner.

Hřebíček, L. (ed.), Glottometrika 11, 1989

A STUDY ON THE POSSIBILITY OF DISCRIMINATING BETWEEN HOMONYMS BY A SEMANTIC CORRELATION APPROACH CONSIDERING THEIR PRECEDING AND SUCCEEDING WORDS

Mutsuko KIMURA, Tokyo
Takano OGINO, Tokyo
Hiroshi KINUKAWA, Kawasaki
Yasuko KURABAYASHI, Tokyo

1. Introduction

Japanese texts generally consist of Kana alphabetic symbols (Japanese Phonetic Symbols) and Kanji characters (Chinese characters). In practice Japanese partially convert phonetic symbols to Kanji characters, creating mixed Kanji/Kana texts. For this reason the conversion of Kana to Kanji is one of the more important problems, when entering Japanese text on word processors, which only accept Kana and Roman letters as input.

Kana-Kanji conversion is difficult because in Japanese one word may have many different ideographic characters (Kanji). The Kanji for a word normally depends on the meaning of the word. The authors have already proposed a word recognition method for unsegmented Japanese phonetic symbol strings (Kinukawa, 1975). In the present paper we propose a method for discriminating between homonyms and evaluate it according to various criteria. We investigated homonym discrimination by semantically correlating the homonyms with the words which precede and succeed them. The homonyms studied here are words with different meanings expressed by the same phonetic string and being the same part of speech

2. Discrimination Algorithm

2.1. Input Data of the Discrimination Procedure

Before carrying out the discrimination procedure, we prepared the input data. with a morphological analyzer. The morphological analyzer divides its input, a character string, into separate words by inserting spaces

(usually absent in Japanese), and tags each word with syntactic category information. In addition, ideographic characters are assigned to words which are unambiguous in the dictionary. The morphological analyzer cannot assign ideographic characters to homonyms depending on their meaning, so homonyms remain in Kana form at this stage.

2.2. Procedure of the Discrimination Method

First of all, we define the probability of occurrence P (X_j) of the word X_j which has one or more homonyms by the transition probability between words or semantic word classes. The definition of the probability P(Xj) is as follows.

(X_1 , X_2) X_2 is a homonym of X_1 .
a_{-i} is the i-th independent word before X_1 or X_2 .
a_i is the i-th independent word after X_1 or X_2 .

$$P(X_1) = \sum_{i=1}^{-m} k_i P_1 \{X_1 | a_i\} + \sum_{i=1}^{m} k_i P_2 \{X_1 | a_i\} \qquad (2.1)$$

$$P(X_2) = \sum_{i=1}^{-m} k_i P_1 \{X_2 | a_i\} + \sum_{i=1}^{m} k_i P_2 \{X_2 | a_i\} \qquad (2.2)$$

k_i is the weight of the i-th position from homonym.
P_1 $\{X_j | a_i\}$ is the conditional probability of the word X_j when a_i appears in the preceding context of X_j.
P_2 $\{X_j | a_i\}$ is the conditional probability of the word X_j when a_i appears in the succeeding context of X_j .....

Here, we define discrimination of homonyms in the following ways:

(1) In case homonyms are two words:

where $| P(X_1) - P(X_2) | > s$ and s is a definite positive number, the word which has larger probability of occurrence is adopted.

(2) In case homonyms are more than three words:

If $P(X_L)$ is the largest probability of occurrence, homonyms $\{X_1 \ldots X_L \ldots X_M \ldots\}$, and if $P(X_L)$ is the second largest probability of occurrence of homonyms $\{X_1 \ldots X_L \ldots X_I \ldots\}$, where $P(X_L)/P^*X_I) > t$ and t is a definite positive number, then X_L is adopted.

(3) In case neither condition (1) nor (2) is satisfied, homonyms cannot be distinguished.

In this study we experimented with the case m = 1 and set the probability to one of 4 ranks, not a percentage. When homonyms appear in the preceding or succeeding context, we calculated the transition probability of m = 2. The calculation of the transition probability P_{ij} is shown below.

Example; 中国では、キョウカイのイシまで勘定する。

<China>	(kyokai)	(ishi)	<count>
[place name]	境界 a_1	石 b_1	
	<boundary>	<stone>	
	教会 a_2	医師 b_2	
	<church>	<medical doctor>	
	協会 a_3	意志 b_3	
	<society>	<will>	

< >: meaning ( ): Japanese pronunciation

[ ]: word class

$$P_{ij} = P_1\{a_i|[\text{place name}]\} \cdot P_2\{b_j|\text{勘定}\}(P_1\{b_j|a_i\} + P_2\{a_i|b_j\})$$
$$1 \leqq i \leqq 3, \quad 1 \leqq j \leqq 3$$

However, when more than two homonyms appear in a sentence with concatenation, we do not try to judge the discrimination among homonyms.

3. Basic Data

3.1. Homonym List

Sets of noun homonyms were selected from (1) the Japanese syllabary order list in *Vocabulary and Chinese Characters in Ninety Magazines of Today* (National Language Research Institute, Japan, 1962, hereafter: NRL). We supplemented these with homonyms from (2) a KWOC list which we made from transcripts of conversations.

Example:

seishi		
	製紙<paper manufacture>	(appears in data (1))
	製糸<silk reeling>	(appears in data (1))
	制止<restraint>	(appears in data (2))
shikaku		
	四角<square>	(appears in data (1))
	資格<qualification>	(appears in data (1))
	死角<dead angle>	(appears in data (2))

278 noun homonyms, representing a total of 634 words, were selected. The largest homonym set consisted of 6 words.

3.2. Semantic Code

In the *Word List by Semantic Principles* (NLR 1964), words are classified into 798 classes. For our word analysis, we further subdivided the lowest level of the 798 classes. The semantic code in the Word List (NLR 1964) is a 5 place number, but in this study we used a 7 place number as a semantic code. We replaced words in the homonym's immediate preceding or succeedinguse context with their semantic codes.

3.3. Co-occurrence Word Table

The authors made the co-occurrence word table of words of the Homonym List (Section 3.1), using the KWOC cards made from ninety magazines by NLR. The columns of the co-occurrence word table are: entry word, frequency of occurrence, position of occurrence (before/after) and semantic code of the co-occurrence word.

3.4. Word-class Relation Tables

(1) Word-class Relation Table 1:

The word-class Relation Table 1 (WRT 1) was made by sorting the Co-occurrence Word Table Data by computer in the ascending order of the entry word character string as the first key and semantic code of co-occurrence as the second key. The table includes, for every homonym and co-occurring semantic code, the frequency of words having that semantic code which appear in the preceding context, and the frequency of words having that semantic code which appear in the succeeding context.

Ex. Table 1:

homonym	semantic code of co-occurring word	frequency in preceding context	frequency in succeeding context
四角<square>	1.1802	5	3
	1.1803	0	4
	:		
	:		
	:		
資格	1.1802	0	0
<qualification>	1.1803	0	0

(2) Word-class Relation Table 2:

We also prepared a Word-class Relation Table 2 (WRT 2) from Table 1. We map the frequency of co-occurring words to four ranks. If the frequency is 0, the rank is 0; if the freqquency is 1 or 2, the rank is 2; if the frequency is over 2, the rank is usually 2, but if the frequency is remarkably large, the rank is 3.

4. Preliminary Survey

In this chapter we describe the reasons why we adopted m = 1 in the formulas (2.1) and (2.2) and subdivided the semantic code of the NLR.

4.1. Physical Distance and Syntactic Distance

As we described before, we defined physical distance between words as m in the formulas (2.1) and (2.2). It is preferable to use syntactic rather than physical distance to select co-occurrence words for the homonym when analyzing semantic relations. However, a parser was not available at the time of this study, so we chose to use physical distance. We tested the usefulness of this approach by examining the relationship between physical and syntactic distance. Sample words were selected in the following way:

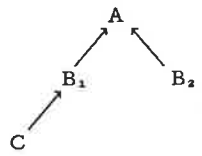
(1) We selected two words from words which are classified to the semantic class of the NLR code 1.1 - 1.5.

(2) For each word selected in step (1) we selected the first (m = -1) and the next (m = -2) preceding words, and the first (m = 1) and the next (m = 2) succeeding words in the context.

The result of the demonstration is shown in Table 4.1.

Table 4.1. Physical distance and syntactic distance

Syntactic relation	Distance				Total
	−2	−1	+1	+2	
Equivalent	1	2	4	0	7
Parent-child	22	76	84	14	196
Brother	3	15	10	2	30
Grandchild	29	0	0	35	64
Other	48	25	22	48	143
Total	103	118	120	99	440



(A, B_1) 親子 <Parent-child>
(A, B_2) 親子 <Parent-child>
(B_1, B_2) 兄弟 <Brother>

(A, C) 孫 <Grandchild>

Figure 4.1 Syntactic Relation of Table 4.1

Parent-child, brother and grandchild relationships in Table 4.1 are shown in Figure 4.1. In Table 4.1 we observe the following:

(a) In position m = ±1 80% of the words had syntactic relations.
(b) In position m = ±2 50% of the words had syntactic relations.
(c) In position m = ±1 67% of the words having some syntactic relations indicated in (a) had the equivalent realtion or parent-child relation.
(d) In position m = ±2 20% of the words having some syntactic relations indicated in (b) had the equivalent relation or parent-child relation.

There are, in fact, cases in which two words have little syntactic relation, yet they have a close semantic relation. These words are, however, infrequent and not easy to find. Therefore, in the co-occurrence word table (WRT1), we decided to replace only the immediately preceeding and succeeding words (m = +-1) of homonyms with their semantic code. Preceding and succeeding words without any semantic relations (as determined by a human editor) were ignored in table construction. There are three kinds of words:

(1) words which have strong relation to the preceding words; for example, 反映 (han-ei) <reflection> and 指示 (shiji) <indication>,

(2) words which have strong relation to the succeeding words; for example, 塀 (hei) <fence>, 箸 (hashi) <chopstick> and 気候 (kikou) <climate>, and

(3) words which show neither relation.

4.2. Difference of Meaning and Overlapping of Co-occurrence Word Sets

We examined in the context of m = ±1, whether the more different meanings of a word a and b are, the less overlapped sets of words co-occurred with a word a and b are. Our overlap ratio was defined as follows:

$$\text{Overlap Ratio} = \frac{\text{Number of common words overlapped in the same position}}{\Sigma(\text{Number of co-occurred words with each homonym word})}$$

We examined 61 pairs of homonyms to see how many times the semantic codes of the preceding word overlapped and how many times the semantic codes of the succeeding word overlapped. Overlap ratios of each pair of homonyms are shown in Table 4.2.

Table 4.2 shows the following:

(1) When the semantic codes are different, the overlap ratios between [2] and [3] of Table 4.2 are not significantly different.

(2) The average value of the overlap ratio of words of which the semantic codes are the same, i.e. in case of [1] of Table 4.2, is only 15% and is larger than the value in case of [2] and [3], but variance of the overlap ratio values of (1) is large.

These results indicate that semantic categorization of homonyms is not so useful in the discrimination of homonyms, because many words must be used appropriately in the given context even though they belong to the same semantic group. For example, both " 女 " (onna) and " 婦人 " (fujin) mean "woman", and the polite expression " 婦人 " is generally usually used in our society.

The following examples show that when co-occurrences are computed by a homonym's semantic group code rather than its Kanji, discrimination becomes impossible.

Example: The semantic codes of co-occurred words of "胃(1.574)" (i) and "意 (1.3045)" (i) are different. But sometimes these words have the same semantic codes, as in the following example:

(a) 胃(1.574) の 運動(1.1510) <activity of stomach>
　　意志(1.3045) の 発動(1.1510) <expression of one's will>

(b) 大酒のみの人(1.202) の 胃 <stomach of drinker>
　　占有者(1.202) の 意志 <possessor's will>

(c) 胃をこわす(2.1571) <have a stomach trouble>
　　意志をふみにじる(2.1571) <interfere one's will>

Table 4.2. Overlap Ratios of Co-occurred Words

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Numbers	Overlap Ratio
[1] Semantic Code Numbers of Homonym Words are same.	糸 (ito) <thread> 皮·革 (kawa) <skin>	51 122 + 89	24 %
	気候 (kikou) <climate> 陽気 (youki) <weather>	5 14 + 14	18 %
	砂糖 (satou) <sugar> 酢 (su) <vinegar>	32 72 + 21	34 %
	佐藤 (satou) : family name 原 (hara) : family name 関 (seki) : family name	42 62 + 24 + 9	46 %
	自己 (jiko) <ego> 自身 (jishin) <self>	6 18 + 73	7 %
	正体 (shotai) <true character> 資料 (shiryo) <material> 実体 (jittai) <entity>	7 20 + 65 + 18	7 %
	上段 (joudan) <upper row> 上部 (joubu) <upper part> 海上 (kaijou) <on the sea>	9 14 + 25 + 30	13 %
	所長 (shocho) <head> 署長 (shocho) <marshal>	7 21 + 13	21 %
	票 (hyo) <vote> 表紙 (hyoshi) <cover>	2 15 + 28	5 %
	繁栄 (han-ei) <prosperity> 不振 (fushin) <dullness>	2 30 + 19	4 %
	有償 (yusho) <onerous> 不利 (furi) <disadvantage>	0 22 + 20	0
	方策 (hosaku) <measure> 政策 (seisaku) <policy>	7 128 + 12	5 %
	[average]		15 %

Table 4.2. Continuation

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Numbers	Overlap Ratio
[2] One Decimal Place of Semantic Code Numbers is same and Two Decimal Places of Semantic Code are different.	河·川 (kawa) <river> 皮·革 (kawa) <skin>	64 141 + 89	28 %
	干渉 (kansho) <interference> 感傷 (kansho) <sentiment> 鑑賞 (kansho) <appreciation>	6 19 + 10 + 12	15 %
	勘定 (kanjo) <counting> 感情 (kanjo) <feeling>	7 25 + 86	6 %
	支持 (shiji) <support> 指示 (shiji) <indicate>	7 40 + 21	11 %
	実体 (jittai) <entity> 実態 (jittai) <actual condition>	4 18 + 21	10 %
	死亡 (shibou) <death> 脂肪 (shibou) <fat>	2 44 + 51	2 %
	生涯 (shogai) <life> 障害 (shogai) <obstacle>	9 53 + 38	10 %
	巣 (su) <nest> 酢 (su) <vinegar>	2 32 + 21	4 %
	政策 (seisaku) <policy> 製作 (seisaku) <manufacture>	19 128 + 89	9 %
	披露 (hirou) <announcement> 疲労 (hirou) <fatigue>	4 28 + 18	9 %
	不振 (fushin) <dullness> 不審 (fushin) <doubt>	0 30 + 25	0
	分 (bun) <minute> 文 (bun) <sentence>	8 258 + 30	3 %
	方策 (housaku) <measure> 豊作 (housaku) <rich harvest>	2 13 + 21	6 %
	保証 (hosho) <gurantee> 保障 (hosho) <security> 補償 (hosho) <compensation>	13 22 + 41 + 13	17 %
	優勝 (yusho) <victory> 有償 (yusho) <onerous>	0 103 + 20	0
	[average]		9 %

Table 4.2. Continuation

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Number	Overlap Ratio
[3] One Decimal Place of Semantic Code Numbers is differrent.	胃 (i) <stomach> 意 (i) <mind>	0 16 + 26	0
	石 (ishi) <stone> 医師 (ishi) <doctor> 意思 (ishi) <intention>	24 52 + 56 + 65	14 %
	意地 (iji) <will> 維持 (iji) <maintenance>	2 17 + 62	3 %
	糸 (ito) <yarn> 意図 (ito) <aim>	19 122 + 16	14 %
	駅 (eki) <station> 液 (eki) <liquid>	9 177 + 44	4 %
	海上 (kaijo) <on the sea> 会場 (kaijo) <meeting place	4 30 + 29	7 %
	会談 (kaidan) <conversation> 階段 (kaidan) <staircase>	0 33 + 32	0
	家庭 (katei) <home> 過程 (katei) <process>	17 61 + 40	17 %
	神 (kami) <god> 紙 (kami) <paper> 髪 (kami) <hair>	29 75 + 70 + 51	15 %
	機会 (kikai) <chance> 機械 (kikai) <machine>	13 86 + 124	6 %
	期間 (kikan) <term> 機関 (kikan) <engine>	16 81 + 138	7 %
	機構 (kikou) <mechanism> 気候 (kikou) <climate>	4 37 + 14	4 %
	佐藤 (satou) : family name 砂糖 (satou) <sugar>	4 62 + 72	3 %
	字 (ji) <character> 地 (ji) <place>	0 37 + 20	0
	自身 (jishin) <self> 自信 (jishin) <confidence> 地震 (jishin) <earthquake>	14 17 + 96 + 7	12 %
		(continued to next page)	

Table 4.2. Continuation

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Number	Overlap Ratio
[3]	島 (shima) <island> 縞 (shima) <stripes>	6 77 + 37	5 %
	姉妹 (shimai) <sister> 終い (shimai) <end>	2 24 + 21	4 %
	収容 (shuyo) <accomodate> 修養 (shuyo) <training>	2 42 + 16	3 %
	正体 (shotai) <one's true character> 招待 (shotai) <invitation>	4 20 + 32	8 %
	上段 (jodan) <upper row> 冗談 (jodan) <joke>	0 13 + 27	0
	焦点 (shoten) <focus> 商店 (shoten) <store>	2 24 + 20	4 %
	商人 (shonin) <merchant> 承認 (shonin) <recognition> 証人 (shonin) <witness>	4 22 + 26 + 9	7 %
	丈夫 (jobu) 上部 (jobu) <upper part>	5 54 + 24	6 %
	資料 (shiryo) <material> 飼料 (shiryo) <feed>	4 65 + 27	4 %
	進行 (shinko) <progress> 信仰 (shinko) <faith> 振興 (shinko) <promote>	7 37 + 20 + 22	9 %
	親切(shinsetsu)<kindness> 新設(shinsetsu)<newly established>	3 40 + 19	5 %
	身長 (shincho) <height> 慎重 (shincho) <careful>	0 14 + 24	0
	生 (sei) <life> 精 (sei) <spirit> 性 (sei) <sex> 所為 (sei) <owing to>	31 29+29+58+46	19 %
	正確 (seikaku) <accurate> 性格 (seikaku) <character>	13 30 + 126	8 %
		(continued to next page)	

Table 4.2. Continuation

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Number	Overlap Ratio
[3]	咳 (seki) <cough> 席 (seki) <seat> 関 (seki) <barrier> 籍 (seki) <membership> 堰 (seki) <dam>	10 11+60+9+8+3	11 %
	葉 (ha) <leaf> 歯 (ha) <tooth> 派 (ha) <group>	14 61 + 24 + 106	7 %
	箸 (hashi) <chopstick> 端 (hashi) <end> 橋 (hashi) <bridge>	28 25 + 47 + 40	25 %
	原 (hara) <plain> 腹 (hara) <abdomen> 原 (hara) : family name	30 18 + 121 +35	17 %
	針 (hari) <needle> 張 (hari) <tension>	7 123 + 28	5 %
	反映 (han-ei) <reflection> 繁栄 (han-ei) <prosperity>	2 22 + 13	5 %
	必死 (hisshi) <desperate> 必至 (hisshi) <inevitable>	0 20 + 13	0
	表 (hyo) <table> 票 (hyo) <vote> 評 (hyo) <criticism>	21 79 + 15 + 18	19 %
	拍子 (hyoshi) <rhythm> 表紙 (hyoshi) <cover>	0 22 + 10	0
	部隊 (butai) <party> 舞台 (butai) <stage>	26 50 + 12	42 %
	不利 (furi) <disadvantage> 振り (furi) <pretence>	0 22 + 10	0
	兵 (hei) <soldier> 塀 (hei) <fence>	17 95 + 23	14 %
	容器 (youki) <receptacle> 陽気 (youki) <season>	0 15 + 14	0
		(continued to next page)	

Table 4.2. Continuation

Semantic Code	Word (pronunciation) <meaning>	Cooccurred Word Number	Overlap Ratio
[3]	用事 (youji) <business> 幼児 (youji) <infant>	0 25 + 21	0
	寮 (ryo) <dormitory> 量 (ryo) <quantity>	0 28 + 132	0
	[average]		8 %

4.3. Semantic Categorization and Size of Co-occurrence Word Table

The more the semantic categorized groups are subdivided, the larger the size of the co-occurrence table is likely to become. The authors then studied whether this is true or not by comparing the number of semantic categorization groups replacing the co-occurrence words, with the number of words.

(1) Homonym:

We do not replace homonyms with their semantic codes because the semantic categorization reduces the discrimination capability of homonyms. Furthermore, there are cases when a homonym belongs to two or more semantic category groups, or two homonyms belong to the same semantic category group. In these cases, the number of items to search is greater than the number of words, and a complex algorithm is needed to handle many cases. For theses reasons, we decided not to replace the homonyms with semantic codes.

(2) Words in context:

The semantic categorization of words in context has to be considered in relation to the discrimination algorithm. If we represent the co-occurrence word table as a matrix, we must restrict the number of semantic category groups so that the matrix does not become too large. Even so, there will be wasted space where there are no co-occurrences.

Table 4.3. Semantic Category and the Degree of Co-occurrence of Useless Words

Semantic Code	Semantic Category <meaning> [category name]	Cooccurrence Times
2.1530	出入り <going in and out>	11
2.120	存在 <existance>	7
2.312	言語 <language>	6
	表現 <expression>	6
	報知 <report>	5
1.259	固有地名 <proper name of place >	5
2.342	行為 <action>	5
—	[numerical expression]	5
1.1770	内外 <inside and outside>	4
1.200	われ <I> , なれ <you> かれ <he> , だれ <who>	4
2.3060	思考 <thought> 認識 <recognition> 知解 <perception>	4
3.195	多い <many> 少ない <few>	4
1.100	この <this> , その <it> あの <that> どの <which>	3
1.202	人間 <human being>	3
1.204	男女 <man and woman>	3
2.111	関係 <relation>	3

Therefore we study the columns of the matrix which show the co-occurred words, and calculate the average values of the following items by using the data of section 4.1.

(a) Numbers of semantic categories which include co-occurring words

(b) Numbers of different words (Place names, proper names and numerical expressions are unified respectively to one category.).

The average value of (a) is 21.4, that of (b) is 21.2. From this result we found that (a) is not significantly different from (b). We then decided to classify the words in context with the semantic categories by using subcategorization of the NLR's code based on the above-mentioned investigation.

4.4. Words and Categories Useless for Discrimination

Words which are apt to co-occur with any words are useless in homonym discrimination. We examined semantic categories which include words likely to co-occur with every word in a homonym set and show them in Table 4.3.

Table 4.3 shows that the greatest number of overlaps occurred for co-occurrence words belonging to the category [姓] (a family name). 11 sets of homonyms had a co-occurrence word belonging to [姓] .

5. Experiments

5.1. Experimental Data

The following data were used in experimental study.

A-data: We selected 47 sentences which contain homonyms from conversational sentences.

B-data: We extracted 50 sentences from the investigation cards of the NLR.

5.2. Experimental Method and Results

The authors first investigated the A-data and B-data by using WRT-1 and then experimented with A-data and B-data by using WRT-2. We used the word dictionary (12,000 words) for word recognition. These experiments were performed according to the following steps.

STEP 1: Referring to the dictionary, write down the semantic codes of one preceding and one succeeding word.

STEP 2: In the first experiment: for each homonym, get the frequency from WRT-1 of co-occurrence words belonging to the same semantic code.

In the second experiment: for each homonym, get the rank from WRT-2 of co-occurrence words belonging to the same semantic code.

STEP 3: Select the first word if its score (the sum of the ranks) is two points or more than the score of the second word, third one and so on.

Example:

庭の〝イシ〟を数える.
↑
the homonym

STEP 1: The preceding semantic code the succeding semantic code

word	of preceding word	word	of succeeding word
〝庭〟	1.470010	〝数える〟	2.306204

STEP 2: Homonym appropriate word (1) (2)

イシ	石	1	2
	意志	0	0

(1): The rank of the co-occurrent preceding words belonging to the semantic code by using WRT-2

(2): The rank of the co-occurrent succeeding words belonging to the semantic code by using WRT-2

STEP 3: X1 = score of 〝石〟 = 1 + 2 = 3
X2 = score of 〝意志〟 = 0 + 0 = 0
X2 - X1 is greater that 2, and so 〝イシ〟 is 〝石〟.

The results of this experiment are shown in Table 5.1.

In the first experiment the results from the A-data differ from those from the B-data. But in the second experiment the results from the A-data and B-data hardly differ at all. "Indistinguishable" in Table 5.1 means that it is difficult to distinguish the homonyms because the co-occurrence words are semantically similar, resulting in no difference in the scores, or that words in context are not contained in the word dictionary.

Table 5.1. Experimental Results

Data		Success	Error	Undecided		
				First		Indistin-
				Right	Error	guishable
1	A-data	3	0	5	4	33
s	B-data	11	1	18	1	19
t	Total	14	1	23	5	52
2	A-data	20	0	6	0	21
n	B-data	21	1	9	0	19
d	Total	41	1	15	0	40

5.3. Discrimination by Human Subjects

Six persons discriminated the homonyms of the A-data by referring to one preceding word and one succeeding word. More than one answer was permitted. The results of this experiment are shown in Table 5.2.

Table 5.2. Subjects' Discrimination of 47 Examples

Person	A	B	C	D	E	F	Average Right Answer(%)
The first word is right.	44	44	43	40	47	44	92.9 %
Right answer is distinguishable.	25	27	15	24	26	26	50.7 %
Right answer is contained.	46	46	46	45	47	45	97.5 %

6. Conclusions

In homonym discrimination by six human subjects, in which the words have semantically related co-occurrence words and only the immediately preceding and succeeding words are used as context, we found that even if a person can distinguish the homonyms, he/she could select only half of them properly. This implies that we need more context to distinguish homonyms properly. However, in our automatic discrimination method, if the system can distinguish the homonyms, it ususally selects the right one.

The following example is one case in which the system could not select the proper word.

家の中にヒがついていて、子供が二人----

Homonyms of "ヒ" are "日" <sun>, "火" <fire> and "灯" <light>, and "火" were selected. In this case "灯" was the optimal word.

In case of extending the context we have to take account of the availability for noun modifying words (Example: コノ <this>, ソノ <the>, イロンナ <various>, etc.) as preceding context, and need to change the processing of methods according to the individuality of the homonym. For example:

メン(面<aspect>, 綿 <cotton>) is regarded as "面" when "綿" is not definitely selected.

ブン(分<part>, 文 <sentence>) is regarded as "分" when "文" is not definitely selected.

These processing methods are further issues which we have to consider.

Acknowledgements

The authors wish to express their gratitude for the cooperation of Mr. NISHIO Toraya and Miss TAKAGI Midori of the National Language Research Institute, Japan.

REFERENCES

DOI Akira, MAEHIGASHI Akira (1973: 9) Tango kan no kanren wo mochiita douon-igigo no hanbetsu (Discrimination of Homonyms Using Correlation between Two Words). *Keiryo Kokugogaku* No. 66, 32-45.

IMANISHI, Hiroko (1972: 12) Bun chu ni okeru bunrui goi no soukan (Interpendence among Each Semantic Group in Sentences). *Keiryo Kokugogaku* No 63, 7-18.

KINUKAWA, Hiroshi, TSUTSUI, Kenji, ODAGIRI, Ikui, KIMURA, Mutsuko (1975) Stenograph to Japanese Translation System. *Information Processing in Japan* 15, 158-162.

The National Language Research Institute of Japan (1962) Gendai zasshi 90 shu no yougo youji - dai 1 bunsatsu (Vocabulary and Chinese Characters in Ninety Magazines of Today - Volume 1). Tokyo, Syuei Syuppan.

----- (1964) Bunrui goi hyo (Word List by Semantic Principles). Tokyo, Syuei Syuppan .

Hřebíček,L.(ed.), Glottometrika 11,1989

Untersuchungen zur Verteilung der Wortarten im Text

R. Hammerl, Bochum

1. Sprachliche Erscheinungen stehen untereinander in verschiedenartigen Beziehungen und bilden vielfältige Abhängigkeitsstrukturen. Diese Abhängigkeiten haben stochastischen Charakter und konnten schon in mehreren Fällen durch die Anwendung mathematischer Modelle erklärt und beschrieben werden (Altmann 1980, Köhler 1986). Die Modellierung mehrerer dieser Abhängigkeiten führte auf verschiedene Wahrscheinlichkeitsverteilungen, wie z.B. bei der Beschreibung der semantischen Diversifikation (Altmann 1985), der Verteilung der Wortlänge (Grotjahn 1982) bzw. der Satzlänge (Altmann 1988), um nur einige zu erwähnen.

In diesem Artikel soll ein erster Versuch der Untersuchung der Verteilung der Wortarten im Text unternommen werden.

In jedem Text treten sprachliche Einheiten auf, die bestimmten Wortarten zugeordnet werden können. Da der lexikalische Bestand der Texte nicht zufällig gewählt wird, da der Textautor ja immer bestimmte Kommunikationsabsichten verfolgt, so kann auch die Zahl der Vertreter der jeweiligen Wortarten in Texten keine zufällige Größe sein. Die Zahl der Vertreter einer Wortart (z.B. der Substantive) hängt dann eben auch von der Zahl der Vertreter anderer Wortarten ab (z.B. der Verben, Adjektive usw.), so daß sich bestimmte Relationen von Häufigkeiten der Vertreter verschiedener Wortarten einstellen werden. Da man annehmen kann, daß die Gestaltung von Texten nach dem allgemeinen "Prinzip" der geringsten Anstrengung erfolgt, so kann auch für die Relationen von Häufigkeiten der verschiedenen Wortartenverteter im Text eine allgemeine Gesetzmäßigkeit angenommen werden (vgl. Gleichung 7), die jedoch nur die Verteilung der Texteinheiten auf die einzelnen "anonymen" Wortarten beschreibt; d.h. diese Gesetzmäßigkeit regelt nicht, welche konkrete der in der jeweiligen Sprache unterschiedenen Wortarten die häufigste, zweithäufigste usw.im jeweiligen Text ist, sondern nur, daß es eine Wortart mit der konkreten Häufigkeit F(1) gibt, die die größte Häufigkeit der Wortartenvertreter im Text ist, eine Wortart mit der Häufigkeit F(2), die

die zweithäufigste Wortart im Text betrifft usw. Es ist Ziel dieses Artikels, eine solche Wortartenverteilung abzuleiten und zu überprüfen.

Die Erscheinung der Wortart ist eine metasprachliche Erscheinung der Grammatik der jeweiligen Sprachen, die sowohl in verschiedenen Sprachen als auch in ein und derselben Sprache nach verschiedenen Kriterien ausgesondert werden kann, was natürlich dazu führen kann, daß Anzahl, Art und Umfang der jeweiligen Wortarten recht verschieden ist.

Die Wortart wird in der Regel als grundlegende lexikalisch-grammatische Kategorie verstanden, die unter Anwendung folgender Kriterien ausgesondert wird (jedem Kriterium kann dabei ein unterschiedliches Gewicht zugeschrieben werden):

a) referentielles Kriterium (Berücksichtigung einer allgemeinen Bedeutung, die allen "Objekten" der Wirklichkeit, die durch die jeweiligen Wortartvertreter gekennzeichnet werden, zukommt);

b) syntaktisches Kriterium (hebt bestimmte Gemeinsamkeiten in der syntaktischen Funktion der jeweiligen Wortartvertreter hervor);

c) morphologisches Kriterium (akzentuiert bestimmte Gemeinsamkeiten der Form und Formveränderung der jeweiligen Wortartvertreter) (vgl. Wörterbuch grammatischer Termini 1976,208).

Oft werden die Wortarten auch zunächst in 2 große Klassen geteilt, die Autosemantika und die Synsemantika. Diese Klassifikation geht auf eine Unterscheidung einer lexikalischen von einer grammatischen Bedeutung von Lexemen zurück, denn Autosemantika werden als Wörter mit einer eigenen lexikalischen Bedeutung verstanden, die diese auch unabhängig vom Kontext realisieren können, Synsemantika als grammatische Hilfs- und Formwörter, die vor allem andere sprachliche Elemente morphologisch und syntaktisch näher kennzeichnen (vgl. Wörterbuch grammatischer Termini 1976, 30 und 183). Auch die Unterscheidung in Auto- und Synsemantika ist nicht eindeutig anwendbar, da z.B. auch bestimmte Synsemantika (z.B Präpositionen) in vielen Wörterbüchern eine eigene lexikalische Bedeutung zugeschrieben wird, andererseits haben oft auch bestimmte Autosemantika in festen Wendungen keine eigenständige lexikalische Bedeutung.

Bekannt sind auch viele Zweifelsfälle, wo ein und demselben Lexem entweder mehrere Wortarten zugeschrieben werden können (z.B. Schwierigkeiten bei der eindeutigen Abgrenzung der Adverbien von den Partikeln im Polnischen), wo Lexeme in verschiedenen Bedeutungen verschiedenen Wortarten angehören (z.B. im Deutschen "danach" als Adverb in "sich danach umdrehen", als Konjunktion in der Bedeutung "dann" in

"erst lesen wir, dann schreiben wir"; "trotzdem" als Adverb und Konjunktion).

Da sich das Sprachsystem in ständiger Entwicklung befindet, muß es auch solche Fälle geben, wo bestimmte Wortartvertreter neue Funktionen übernehmen und dann auch in neue Wortklassen übergehen (z.B. die Präpositionen "dank, kraft" im Deutschen, die aus Substantiven entstanden sind).

Es gibt auch mehrere Fälle, wo ganze Wortgruppen, deren Bestandteile ihre eigene Bedeutung zugunsten einer gemeinsamen, übertragenen Bedeutung verloren haben, stellvertretend für bestimmte Wortartvertreter als Einzellexeme stehen und somit auch als solche angesehen werden können (z.B. im Deutschen "in bezug auf", "auf seiten", die als Präpositionen gekennzeichnet werden können).

Aus dieser kurzen und bei weitem nicht alle wichtigen Aspekte umfassenden Kennzeichnung der Problematik der Wortartenklassifikation können für unsere Untersuchungen zur Verteilung der Wortarten folgende Schlußfolgerungen gezogen werden:

a) Es muß unterschieden werden zwischen einer Verteilung der Wortarten im Text, im Textwörterbuch und im Lexikon, denn erst im Text ist in vielen Fällen eine eindeutige Anwendung des morphologischen und syntaktischen Kriteriums der Wortartenabgrenzung möglich. Da, wie Orlov (1982 a,b,c) zeigte, Texte vom Textautoren nach bestimmten Gesetzmäßigkeiten gestaltet werden, liegt es zunächst auf der Hand, bei der Untersuchung der Verteilung der Wortarten im Text zu beginnen. Das Textwörterbuch - und in noch stärkerem Maße das Lexikon -ist das Ergebnis bestimmter Abstraktionen vom Text nach Prinzipien, die konventionell festgelegt wurden. Kennt man die Verteilung der Wortarten im Text, so kann durch Einbeziehung weiterer Einflußfaktoren der Versuch unternommen werden, die Wortartenverteilung im Textwörterbuch und im Lexikon zu modellieren.

b) Untersucht man die Verteilung der Wortarten im Text, so muß bedacht werden, daß die Häufigkeit bestimmter Wortarten in Texten verschiedener Stile unterschiedlich ist (man spricht ja auch von einem "Nominalstil", "Verbalstil" usw.). Außerdem wird die Häufigkeit bestimmter Wortarten in konkreten Texten auch von der Thematik des Textes, dem Alter des Textautors und mehreren anderen Faktoren abhängen. Die von uns gesuchte Verteilung der Wortarten im Text soll für eine möglichst große Zahl von Texten Gültigkeit haben, unabhängig von den Einflußfaktoren wie Stil, Thema der Texte usw., deren Einfluß durch verschiedene Parameterwerte der Wortartenverteilung berücksichtigt werden soll. Wenn

wir eine allgemeine Wortartenverteilung im Text suchen, so muß diese Verteilung Resultat eines allgemeinen Modellansatzes sein, muß sowohl theoretisch als auch empirisch validiert werden können.

c) Wird eine allgemeine Wortartenverteilung im Text gesucht, so heißt das jedoch nicht, daß sie für jeden Text und für jede Wortartenklassifikation zutreffen wird. Wenn dem so wäre, so wäre auch die Aussagekraft des entsprechenden mathematischen Modells sehr gering. Wenn das mathematische Modell der Wortartenverteilung im Text ein allgemeines Textgesetz darstellen soll, so muß es auch für mindestens eine Wortartenklassifikation empirisch validiert werden können. Nicht mit dem Modell übereinstimmende empirische Daten aus konkreten Texten, die auf einer bestimmten Wortartenklassifikation beruhen, können somit auch nicht als Beweis dafür angesehen werden, daß das jeweilige Modell die allgemeine Wortartenverteilung im Text nicht adäquat beschreibt. Falls das untersuchte Modell der Wortartenverteilung im Text bei einer konkreten Wortartenklassifikation nur für einige Texte nicht bestätigt werden kann, so muß zunächst nach Einflußfaktoren gesucht werden (durch vergleichende Textanalysen), die die festgestellten Abweichungen der empirischen Wortartenverteilung in den konkreten Texten von der theoretischen Wortartenverteilung (dem Modell) verursachen können und im Modellansatz der abgeleiteten Verteilung noch nicht oder nicht in entsprechendem Maße berücksichtigt worden sind. Dies ist besonders dann möglich, wenn bisher noch nicht berücksichtigte Einflußfaktoren so gefunden werden können, daß beim Erreichen einer bestimmten Intensität dieser Faktoren die allgemeine Wortartenverteilung in der Regel nicht bestätigt werden kann. In diesen Fällen wäre demzufolge eine entsprechende Modifizierung des Modellansatzes möglich; es können aber auch gleichzeitig Bedingungen angegeben werden, für die die bisherige - nicht modifizierte - Wortartenverteilung angewendet werden kann. Diese Konsequenz findet bei der empirischen Überprüfung der von uns abgeleiteten Wortartenverteilung Anwendung (vgl. Punkt 3).

d) Auf der Suche nach der Wortartenklassifikation, für die unsere Wortartenverteilung in Texten zutreffend ist, gehen wir von der klassischen Wortartenklassifikation aus, für die auch reichhaltig empirische Daten vorhanden sind (vgl. Golovin 1974, 9/10; Becker 1988, 332ff.; vgl. auch Punkt 3).

2. Ableitung eines mathematischen Modells der Wortartenverteilung im Text

Bei der Modellierung der Wortartenverteilung im Text gehen wir von dem von Altmann, Köhler (1989) vorgeschlagenen Modellierungsansatz aus, der allgemein als

$$\frac{dp(x)}{p(x)} = \frac{f(x)}{g(x)}\,dx \qquad (1)$$

dargestellt werden kann, wo dp(x)/p(x) die relative Veränderungsrate der Wahrscheinlichkeiten sind, f(x) und g(x) Funktionen, die die sogenannten Zipfschen Kräfte darstellen (f(x) steht für die Sprecherbedürfnisse, g(x) für die Hörerkraft), und x die unabhängige Variable.

Bei der Beschreibung der Wortartenverteilung wäre x die Rangnummer der nach abnehmender Auftrittshäufigkeit im Text geordneten Wortarten, p(x) deren relative Auftrittshäufigkeit. Obwohl in diesem Falle x nur die Werte x = 1,2,...,n, also diskrete Werte annehmen kann, haben wir mit Gleichung (1) einen Ansatz für stetige Variable x gewählt, weil das für die von uns gewählte Funktion f(x) und somit für die Lösung der entsprechenden Gleichung wesentliche Rechenerleichterungen bringt.

Nun gilt es, die von x abhängigen Funktionen f(x) und g(x) zu bestimmen. In Abhängigkeit davon erhält man unterschiedliche Lösungen (Verteilungen). Man sollte jedoch bedenken, diese Funktionen möglichst einfach zu halten, möglichst wenige Parameter zu verwenden, da dies dann für die empirische Überprüfung der erhaltenen Verteilung von großer Bedeutung sein kann (wenn z.B. die Zahl der unterschiedenen Wortarten 7 beträgt, so wäre eine Verteilung mit 6 Parametern in empirischen Untersuchungen nicht mehr mit dem X^2 -Test überprüfbar.

Die Funktion f(x), die die Sprecherkraft repräsentiert, wurde von uns als

$$f(x) = a + d \ln x \qquad (2)$$

bestimmt, wo a und d Konstanten sind (d > 0). Die Konstante a steht hier als Störgröße stellvertetend für alle Störfaktoren, die einen Einfluß auf die Wortartenverteilung haben könnten und nicht gesondert berücksichtigt werden. Das Produkt d ln x dagegen bedeutet, daß der Sprecher bestrebt ist, die relative Abnahme der Auftrittswahrscheinlichkeit der Wortarten mit wachsendem x möglichst groß zu gestalten, d.h. die Zahl der unterschiedenen Wortarten möglichst zu reduzieren.

Ekman (1964) hat gezeigt, daß, wenn man Reaktionen (R) auf Zahlen (n), d.h. die psychischen Empfindungen von Zahlen, als

$$R = a' + d \log N \qquad (3)$$

ansetzt, wo a' und d Konstanten sind, und wenn man entsprechend dem Fechnerschen Gesetz

$$R = c + b \log S \qquad (4)$$

ansetzt (wo S die Intensität physikalischer Reize darstellt, die durch die Angabe der Zahlenwerte N von Versuchspersonen abgeschätzt wurden), so folgt daraus das Potenzgesetz von Stevens (1961):

$$N = e^{(c-a'/d)}\, S^{(b/d)}$$

bzw.

$$N = kS^n, \qquad (5)$$

wo $n = b/d$ und $k = e^{(c-a'/d)}$.

Dieses Gesetz wurde aufgrund von experimentellen Untersuchungsbefunden oft anstelle des Fechnerschen Gesetzes zur Beschreibung des Zusammenhanges zwischen physikalischen Reizintensitäten und deren Empfindungen durch den Menschen verwendet (vgl. auch Sydow, Petzold 1981, 155ff.).

Aus diesem Ansatz von Ekman leiten wir die Vermutung ab, daß in bestimmten Fällen die Intensität von menschlichen Empfindungen über die Größe von "Zahlen" nach Gleichung (3) abgeschätzt werden kann. Demzufolge nehmen wir an, daß der Sprecher die quantitative "Intensität" der Variablen x als a' + d ln x einschätzt und mit einer entsprechenden Kraft auf die Gestaltung der Verteilung der Wortarten wirkt.

Auf analoge Weise könnte man auch die Funktion g(x) modellieren, was natürlich zu einer relativ komplizierten Differentialgleichung führen würde. Wie unsere empirischen Untersuchungen bestätigen werden, kann die Hörerkraft durch die einfache Funktion

$$g(x) = cx \qquad (6)$$

beschrieben werden, wo c eine Konstante darstellt (c > 0). Eine Vergrößerung von x (viele Wortarten) führt demzufolge (laut Gleichung (1)) zu einer Verkleinerung der relativen Veränderungsrate dp(x)/p(x). Der Hörer ist somit bestrebt, eine möglichst große Zahl von unterschiedlichen

Wortarten im Text zu markieren, er wirkt somit der Sprecherkraft entgegen.

Gleichung (1) nimmt dann folgende Form an:

$$\frac{dp(x)}{p(x)} = \frac{a + d \ln x}{cx} dx. \qquad (7)$$

Durch Integration erhält man:

$$\ln p(x) = [(a/c) + (d/2c)\ln x]\ln x + E$$

$$p(x) = sx^{(k + m \ln x)}, \qquad (8)$$

wo $s = e^{E}$, $k = (a/c)$ und $m = (d/2c)$.

Da bei uns x nur bis zu einem endlichen Wert n läuft (x = 1,2,...,n) und diskret ist, muß Gleichung (3) noch normiert werden, wodurch man folgende Wahrscheinlichkeitsfunktion erhält:

$$p(x) = \begin{cases} p(x=1) & \text{für } x = 1 \\ T[1-p(x=1)]x^{k + m \ln x} & \text{für } x = 2,3,\ldots,n \end{cases} \qquad (9)$$

wo k und m Parameter und T die Normierungsgröße sind. Gleichung (5) stimmt mit der von Dolinskij (1988,91) angeführten Verteilung überein, die bei der Beschreibung der Daten eines freien Assoziationsexperimentes - Abhängigkeit zwischen der Häufigkeit F(x) der verbalen Reaktion auf ein Stimuluswort, die in der Rangordnung der Reaktionen nach deren Häufigkeit den Rang x einnimmt und dem Rang x selbst - angewandt wurde. Aus dieser Arbeit von Dolinski wurde leider nicht ersichtlich, wie die entsprechende Formel zur Berechnung von F(x) abgeleitet wurde, so daß wir uns hier nicht darauf berufen können. Es fällt jedoch auf, daß Gleichung (5) eine Modifizierung des bekannten Zipfschen Gesetzes

$$p(x) = hx^{-t} \qquad (7)$$

darstellt(wo h und t Konstanten sind), wenn man anstelle der Konstanten t die Funktion $u(x) = k + m \ln x$ einsetzt.

3. Empirische Überprüfung der abgeleiteten Wortartenverteilung

Die von uns abgeleitete Verteilung wurde an 110 Texten überprüft, an 60 russischen Texten (Golovin 1974, 9/10) und an 50 deutschen Texten (Becker 1988, 332/333).

Die von Golovin angeführten Daten betreffen eigentlich keine Gesamttexte, sondern Textpassagen mit etwa 500 Wortstellen aus größeren Texten. Da, wie wir unten zeigen werden, die Textlänge einen entscheidenden Einfluß auf die Beschreibung der Wortartenverteilung mit unserem Modell hat, sind die Daten von Golovin für unsere empirischen Untersuchungen geeignet, obwohl sie keine Gesamttexte betreffen. Die 60 russischen Textfragmente wurden aus Prosatexten des 19. Jahrhunderts entnommen:

20 Textfragmente aus den Werken von A.I.Gerzen (Textkorpus I),

20 Textfragmente aus den Werken von I.S.Turgenev (Textkorpus II)

20 Textfragmente aus den Werken von I.A.Gontscharov (Textkorpus III).

Alle Textfragmente wurden nach dem Zufallsprinzip ausgewählt. Golovin führt Daten zu 7 Wortarten an, ohne genauere Angaben zur Wortartenklassifikation selbst zu geben. Unterschieden wurden folgende Wortarten: Substantiv, Verb, Adjektiv, Adverb, Pronomen, Präposition, Konjunktion. Es handelt sich hier um 7 traditionelle Wortarten, die auch Becker (1988) in seinen Untersuchungen von deutschen Presseartikeln unterschieden hat. Aus den Daten von Becker haben wir ausgewählt:

20 Texte aus dem FAZ-Korpus (Artikel aus der Frankfurter Allgemeinen Zeitung) (Textkorpus IV),

10 Texte aus dem NZZ-Korpus (Artikel aus der Neuen Züricher Zeitung) (Textkorpus V)

20 Texte aus dem PRESSE-Korpus (Presse) (Textkorpus VI).

Becker unterscheidet insgesamt 9 Wortarten: Substantiv, Verb, Adjektiv, Adverb, Pronomen, Präposition, Konjunktion, Partikel, Artikel. Numeralien werden zu den Adjektiven gezählt.

In unseren Untersuchungen wird die Wortart der Artikel nicht berücksichtigt, Artikel werden als Determinativa und "obligatorische Begleiter des Nomens" (Engel 1988, 523) angesehen, Partikel werden zu den Adverbien gezählt. (Die Zahl der Partikel ist in der Regel relativ klein, deshalb scheint die Zusammenfassung mit der ebenfalls recht kleinen Klasse der Adverbien angebracht.) Auf diese Weise wurden wiederum 7 Wortarten unterschieden, wie auch in den Textkorpora I-III.

Es wurde natürlich auch untersucht, welchen Einfluß die Zusammenfassung der Adverbien und Partikel zu einer Wortart und die Nichtberücksichtigung der Artikel als gesonderte Wortart auf die Güte der Beschreibung der empirischen Wortartenverteilungen mit unserem Modell hatte. Dazu wurden die 5 ersten Texte des Textkorpus IV ausgewählt und die Wortartenverteilung in 3 Versionen untersucht:

a) für alle 9 von Becker unterschiedenen Wortarten,
b) bei Zusammenfassung der Adverbien und Partikel zu einer Klasse,
c) bei zusätzlicher Nichtberücksichtigung der Artikel.

Die Ergebnisse dieser Untersuchungen zeigt Tabelle 1.

Tabelle 1.
Beschreibung verschiedener Versionen der Wortarten-klassifizierung mit Gleichung (6) für Texte aus Textkorpus IV

Version	Textnummer	X^2	FG	$P(X^2)$
a	1	13.31	5	0.0206
	2	12.37	5	0.0301
	3	12.99	5	0.0235
	4	22.35	5	0.0005
	5	32.76	5	0.0000
b	1	7.90	4	0.0955
	2	7.86	4	0.0969
	3	8.87	4	0.0645
	4	5.83	4	0.2123
	5	18.00	4	0.0012
c	1	2.90	3	0.4073
	2	7.05	3	0.0703
	3	3.82	3	0.2815
	4	2.26	3	0.5205
	5	6.86	3	0.0763

Wie man aus Tabelle 1 ersehen kann, beschreibt unser Modell (Gleichung (6)) am besten die Wortartenklassifikation von Version c, was sich auch bei deren Anwendung auf alle 110 untersuchten Texte bestätigte. Die Ergebnisse dieser Untersuchungen führt Tabelle 2 auf (Anhang). In den Untersuchungen aller dieser Texte waren wir bemüht, neben der Einhaltung der Bedingung $m > 0$ auch eine solche Anpassung zu finden, wo auch $k > 0$ galt. Dies erleichtert eventuelle spätere Untersuchungen zum Vergleich der Parameterwerte, war aber nicht in allen Fällen möglich.

Wie aus Tabelle 2 ersichtlich ist, beschreibt Gleichung (6) die empirische Wortartenverteilung in 107 Texten von insgesamt 110 Texten gut. In den 11 mit einem "+" gekennzeichneten Texten wurden zusätzlich die Wortarten der Adverbien und Adjektive zusammengefaßt, da festgestellt wurde, daß gerade die Adverbien und Adjektive in diesen Texten sehr abweichende Werte im Vergleich zu anderen Wortarten in Texten derselben Länge haben und/oder im Vergleich zu diesen Wortarten in der Rangordnung andere Werte belegen (dies betrifft aber in diesen Fällen nicht die Summe der Adverbien und Adjektive).

Es müßte nun untersucht werden, ob in diesen Texten besondere stilistische Mittel angewandt wurden, die sich in einer solchen empirischen Wortartenverteilung äußern (denn 6 der 11 hier besprochenen Fälle betreffen Texte eines Autors (Gontscharov)) oder ob diese Daten aus Schwierigkeiten bei der Abgrenzung der Wortarten Adjektiv und Adverb resultieren. Möglicherweise kommen beide Ursachen in Betracht.

Unser Modell der Wortartenverteilung konnte jedoch für 3 Texte nicht bestätigt werden. Es fällt aber sofort auf, daß es sich hier im Vergleich zu anderen Texten um wesentlich längere Texte handelt (mit einem N von 2000 bis über 3000 Wörtern).

Um zu überprüfen, ob sich diese Tendenz auch an anderen "längeren" Texten bestätigt, wurden noch 3 weitere solcher Texte ausgewählt (siehe Tabelle 2: Zusatztexte). Auch für diese Texte konnte unser Modell - wie erwartet - nicht bestätigt werden.

Hieraus kann geschlußfolgert werden, daß unser Modell der Wortartenverteilung im Text nur auf kurze Texte (Textpassagen) anwendbar ist, die Wortartenverteilung in solchen Texten aber überzeugend beschreibt (was wir leicht an einer noch größeren Zahl solcher Texte hätten zeigen können).

Tabelle 2
Beschreibung der Wortarten-verteilung im Text mit Gleichung (6)

Text-korpus	Text-num-mer	Parameter k	m	$s=f_1$	X^2	FG	$P(X^2)$
I	1	0.0001	0.2864	150	4.79	3	0.19
	2	0.0001	0.3175	158	2.40	3	0.49
	3	0.0202	0.2823	164	2.30	3	0.51
	4	0.0000	0.2587	168	3.31	3	0.35
	5	0.0016	0.3591	169	5.18	3	0.16
	6	0.0005	0.3234	170	4.94	3	0.18
	7	0.2256	0.1462	171	0.95	3	0.81
	8	0.2036	0.0647	181	1.85	3	0.61
	9	0.0010	0.2570	183	3.62	3	0.31
	10	0.0010	0.2534	183	1.99	3	0.58
	11	0.0010	0.3481	186	1.83	3	0.61
	+ 12	0.0001	0.1884	187	3.62	2	0.16
	13	0.0222	0.2777	190	2.57	3	0.46
	+ 14	0.1596	0.1165	199	4.80	2	0.09
	15	0.0000	0.1888	201	5.40	3	0.14
	16	0.0001	0.4654	201	6.74	3	0.08
	17	0.0001	0.3708	201	7.53	3	0.06
	18	0.0001	0.2701	205	5.53	3	0.14
	19	0.0005	0.3459	213	6.97	3	0.07
	+ 20	0.0001	0.2221	218	0.69	2	0.71
II	1	0.1989	0.1908	133	5.82	3	0.12
	2	0.3285	0.1300	146	1.32	3	0.72
	3	0.5908	0.1457	155	2.95	3	0.40
	4	0.0002	0.3547	159	4.89	3	0.18
	5	0.7048	0.1259	162	3.92	3	0.27
	6	0.0001	0.3391	167	6.32	3	0.10
	7	0.1499	0.2932	168	0.47	3	0.92
	8	0.3013	0.0781	168	4.73	3	0.19
	9	0.1200	0.2199	168	1.15	3	0.77
	10	0.0502	0.2237	169	4.88	3	0.18
	11	0.0003	0.2458	172	2.40	3	0.49
	12	0.0001	0.3516	173	3.61	3	0.31
	13	0.0001	0.2978	177	6.05	3	0.11
	14	0.0001	0.3785	182	3.60	3	0.31
	15	0.0001	0.2292	186	2.76	3	0.43
	16	0.4451	0.0615	190	0.41	3	0.94
	17	-0.5694	0.0348	192	0.82	3	0.85
	18	0.4527	0.0002	195	0.14	3	0.99
	19	0.0005	0.1932	198	2.82	3	0.42
	20	0.5420	0.0511	208	1.50	3	0.68

Tabelle 2. Fortsetzung

III	1	0.0949	0.2533	149	3.24	3	0.36
	2	0.2214	0.1439	156	1.99	3	0.57
	+ 3	0.0009	0.0779	161	0.22	2	0.89
	+ 4	0.4119	0.0102	162	1.37	2	0.51
	5	0.5262	0.0866	164	4.89	3	0.18
	6	0.4113	0.0086	170	0.01	3	0.99
	7	0.0014	0.3303	176	5.05	3	0.17
	8	0.0057	0.2123	176	1.23	3	0.75
	9	0.0000	0.2423	180	5.14	3	0.16
	10	0.0001	0.2980	182	7.68	3	0.05
	+ 11	0.0549	0.0902	184	0.33	2	0.85
	12	0.5940	0.0022	185	2.23	3	0.33
	+ 13	0.5175	0.0001	185	5.29	2	0.07
	+ 14	0.0001	0.1134	194	4.03	2	0.13
	15	0.0054	0.2162	196	2.34	3	0.51
	16	0.0005	0.2448	196	5.32	3	0.15
	17	0.0001	0.2144	197	2.76	3	0.43
	18	0.0000	0.1613	220	1.66	3	0.65
	+ 19	0.2164	0.0037	223	0.33	2	0.85
	20	0.0017	0.2544	224	5.87	3	0.12
		k	m	f_1	X^2	N	FG	P
IV	1	0.0006	0.4233	104	0.90	289	3	0.83
	2	0.8356	0.2038	37	4.84	82	3	0.18
	3	0.0207	0.3525	253	7.67	664	3	0.05
	4	0.0004	0.5515	110	5.31	241	3	0.15
	+ 5	-1.9608	1.4534	248	3.42	689	2	0.18
	6	0.0004	0.3537	143	2.44	444	3	0.49
	7	0.7413	0.0948	163	7.74	422	3	0.13
	8	-0.2345	0.4376	1058	27.41	2838	3	0.00
	9	0.7189	0.0926	59	1.74	168	3	0.63
	10	0.7174	0.1474	129	5.46	358	3	0.14
	11	0.0001	0.4053	65	0.47	188	3	0.92
	12	0.0017	0.6251	44	0.68	103	3	0.08
	13	0.0003	0.3902	93	2.49	234	3	0.48
	14	0.0002	0.5259	42	6.21	108	3	0.10
	15	-0.9079	0.6844	139	2.62	400	3	0.08
	16	0.0002	0.4473	136	6.82	333	3	0.07
	17	0.0042	0.5001	156	6.37	446	3	0.10
	18	0.0001	0.4510	124	7.79	290	3	0.05
	19	-0.7168	0.6635	987	40.73	2628	3	0.00
	20	0.0001	0.3152	183	4.90	512	3	0.18

Tabelle 2. Fortsetzung

V	1	0.0043	0.5108	142	4.01	393	3	0.26
	2	0.0002	0.2820	157	7.05	454	3	0.07
	3	-1.6159	1.0787	260	3.82	631	3	0.28
	4	0.0016	0.4225	400	4.99	937	3	0.17
	5	-0.4288	0.5919	238	6.86	605	3	0.08
	6	-0.5359	0.5612	121	6.62	358	3	0.09
	7	-2.3478	1.3477	1318	7.04	3378	3	0.07
	8	-1.5516	1.1677	81	4.48	209	3	0.21
	9	0.0002	0.2903	206	7.19	607	3	0.07
	10	0.0014	0.3047	157	1.85	441	3	0.60
VI	1	-1.6358	1.1308	110	5.59	260	3	0.13
	2	0.0000	0.4230	145	7.69	366	3	0.05
	3	0.0481	0.6353	54	2.99	105	3	0.23
	4	0.2026	0.4328	71	4.28	156	3	0.23
	+ 5	0.0006	0.6930	146	4.84	398	2	0.09
	6	0.0015	0.3968	148	3.98	359	3	0.26
	7	0.0002	0.5333	46	1.48	110	3	0.69
	8	0.0006	0.5022	111	1.82	291	3	0.61
	9	0.0003	0.4184	91	2.99	265	3	0.39
	10	0.0002	0.4837	79	1.49	171	3	0.69
	11	0.0000	0.4332	1001	37.40	2469	3	0.00
	12	0.0029	0.5406	28	2.84	78	3	0.42
	13	0.0004	0.4210	127	2.76	333	3	0.43
	14	0.0001	0.4849	71	1.72	179	3	0.63
	15	0.0023	0.3438	139	4.62	340	3	0.20
	16	0.0004	0.3396	55	1.39	143	3	0.71
	17	0.0052	0.6220	36	0.34	72	3	0.84
	18	0.0002	0.4389	117	3.55	336	3	0.31
	19	-1.9477	1.2375	131	4.13	342	3	0.25
	20	0.0006	0.5622	55	2.95	140	3	0.40
Zusatztexte:								
	1	-0.8224	0.7878	1422	40.54	3669	3	0.00
	2	0.0005	0.3216	897	8.44	2449	3	0.04
	3	-0.0611	0.3950	3453	57.88	8945	3	0.00

4. Zusammenfassung

In diesem Artikel konnte ein Modell der Wortartenverteilung in kurzen Texten oder Textfragmenten abgeleitet und an mehr als 100 Texten aus 2 Sprachen bestätigt werden. Dabei wurde den empirischen Untersuchungen die klassische Wortartenklassifikation zugrunde gelegt, die für 2 Wortarten modifiziert wurde.

Daß die Textlänge einen Einfluß auf die statistische Textstruktur besitzt, wurde schon von anderen Autoren bestätigt (Orlov 1982 a,b,c). Demzufolge gelten entsprechende statistische Gesetzmäßigkeiten auch nicht für Texte jeder beliebigen Länge, denn lange Texte sind thematisch, stilistisch und somit auch hinsichtlich sprachlicher Phänomene wesentlich weniger homogen als kürzere Texte oder Textfragmente. Hierin wird auch die Ursache gesehen, warum unsere Wortartenverteilung nur an kürzeren Texten bestätigt werden konnte. Dies kann eben auch so interpretiert werden, daß die bewußte Textgestaltung durch den Textautor hinsichtlich dieser (und eventuell auch anderer) lexikalisch-grammatischer Erscheinungen nur in kleinen Textfragmenten möglich ist, daß sich größere Texte eben aus bewußt gestalteten "kleineren" Textfragmenten zusammensetzen, die sich aber hinsichtlich dieser Erscheinungen relativ stark unterscheiden können.

Orlov (1982 a,b,c) hat gezeigt, daß ein ähnliches Phänomen auf die lexikalische Gestaltung von Texten zutrifft, daß das Zipfsche Gesetz zur Beschreibung der Verteilung der Häufigkeit von Lexemen nur auf Texte einer bestimmten Länge zutrifft, auf sogenannte "abgeschlossene" Texte (vgl. hierzu auch Arapov 1976) mit einer optimalen Textlänge.

Unsere Untersuchungen haben auch gezeigt, daß sich hinsichtlich der Verteilung der klassischen Wortarten im Text bestimmte Gesetzmäßigkeiten feststellen lassen, die möglicherweise auch rückführend auf eine Begründung dieser Wortartenklassifikation angewandt werden können. Dies scheint ein günstiger Ansatzpunkt für neue Untersuchungen in der Diskussion um die Wortartenklassifizierung zu sein.

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), Verteilungen der Satzlänge. *Glottometrika 9, 147-167.*

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

Arapov, M.W. (1976), Struktura ilosciowa tekstu skonczonego. In: *Semantyka tekstu i jezyka.* Wroclaw, 145-163.

Becker, H. (1988), Die Wirtschaft in der neuen Presse. Sprachliche Untersuchungen zur Wirtschaftsberichterstattung in der "Frankfurter Allgemeinen Zeitung", der "Neuen Züricher Zeitung" der "Presse" und im "Neuen Deutschland". Diss., Bochum.

Dolinskij, V. A. (1988), Raspredelenie reakcij v experimentach po verbal'nym associacijam. *Acta et Commentationes Universitatis Tartuensis 827, 89-101.*

Ekman, G. (1964), Is the power law a special case of Fechners law? *Perceptual and Motor Skills 19, 730.*

Engel, U. (1988), Deutsche Grammatik. Heidelberg.

Golovin, B.N. (1974), Opyt primenenia korrelacionnogo analiza w izučennii jazyka. In: *Voprosy statističeskoj stilistiki.* Kiev, 5-16.

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

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

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

Orlov, Ju.K. (1982b), Dynamik der Häufigkeitsstrukturen. In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š., *Sprache, Text, Kunst.* Bochum, Brockmeyer, 82-117.

Orlov, Ju.K. (1982c), Ein Modell der Häufigkeitsstrukturen des Vokabulars. In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š., *Sprache, Text, Kunst.* Bochum, Brockmeyer, 118-192.

Stevens, S.S. (1961), To honor Fechner and repear his law. *Science 133, 80-86.*

Sydow, H., Petzold, P. (1981), Mathematische Psychologie. Berlin, VEW Dt. Verlag der Wissenschaften.

Wörterbuch grammatischer Termini (1976). Greifswald.

Hřebíček, L. (ed.), Glottometrika 11, 1989

Interpreting textual dimensions through factor analysis: Grammatical structures as indicators of textual dimensions

Pauli Saukkonen, Helsinki

What are the factors which give rise to different types of text? What are, in general terms, the main features that characterize textual types and serve to distinguish between them? What makes an article into an article, a news item into a news item or a narrative into a narrative?

Descriptions and comparisons of texts and textual types have normally been based either on macrostructures or on lexical, grammatical and textual microstructures. Both of them are insufficient: macrostructures include too small information, and microstructures lack intregration.

I shall concentrate here on microstructures, in order to intregrate them by looking for broader common properties, the textual dimensions which they describe and represent. The starting-point is the fact that it is neither particularly revealing nor in all probability in accordance with the language user's linguistic competence to list only a vast heterogeneous set of isolated linguistic variables and the values associated with them.

Efforts have been made to look for macroscopic properties by statistical methods (e.g. Biber 1985, 1988; cf. Kraus, Polak 1967; Schwibbe, Räder 1984), but the task has proved a very difficult one. Thus there still seems to be a lack of any overall organization or theory of detailed text typology. Faced with an overwhelming number of linguistic features which could potentially be relevant, the investigator usually finds imself unable to see the wood for the trees and fails to perceive the common dimensions which the various linguistic features may form.

The material and its clustering analysis

I have been involved in studying text types in Finnish for the last twenty years (through an extensive project carried out at the Department of Finnish and Lappish at the University of Oulu, Finland), and have attempted to determine the stylistic-rhetorical structure and typology of different types of text by means of statistical analysis. I would like to put forward here certain generalized observations on the distinctive properties, or dimensions, of the subtypes of the informative text type (such as news reports, articles, oral narratives etc.).

Language samples were used to calculate statistically the grammatical and lexical differences between the text types, after which factor analysis was employed to obtain a pattern for the location of the types in spaces defined by the factors. The fundamental question really concerns the interpretation of the factors in text type matrices. What are the major normative properties and mutual differences which the individual is required to master intuitively?

The corpus consists of certain classes of text, each represented by a set of random samples. A general impression of the material may be gained from the following list of 20 text classes grouped as indicated by a cluster analysis based on 50 grammatical and lexical categories (Fig. 1). (The material is described in detail in Saukkonen 1982).

An important thing to study would be the question of how relevant and how sufficient the 50 variables used in the clustering analysis are, but this is not possible here. The variables are listed in Table 1. They serve to show, according to Figure 1, that certain text classes belong closely together, forming text types of varying degrees, e.g. newspaper reports and radio talks, which are in turn most closely associated with magazine articles for young people and magazine articles for adults and then with current affairs radio programmes (radio reports), newspaper articles and on to news items. Radio serials for young people, radio serials for adults and radio serials for schools then fall into a separate group of their own, as also do radio discussions, radio interviews and oral narratives from unscripted person-to-person conversations. The remaining classes lie somewhat apart, each more clearly representing a text type of its own.

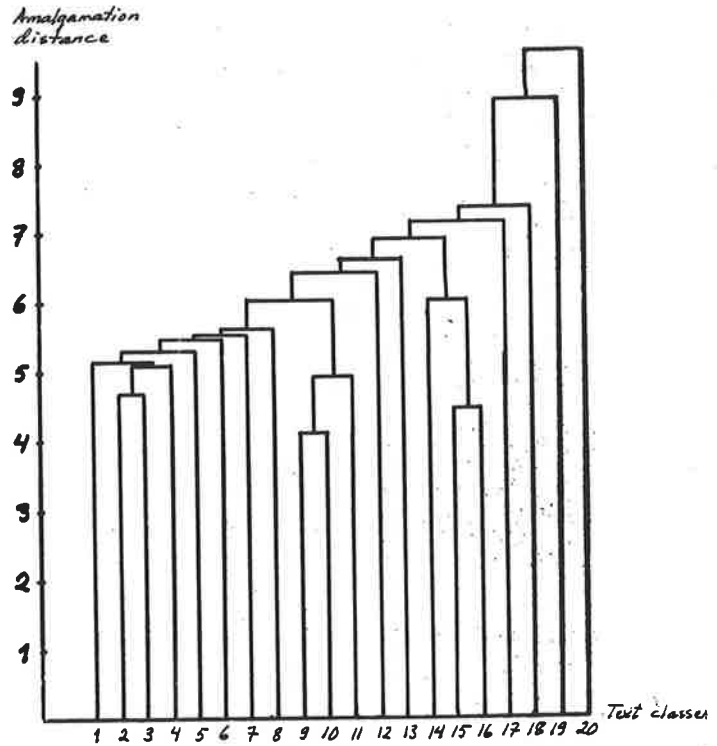


Figure 1. Tree diagram of the clusters according to 50 linguistic variables (cf. Table 1)

Text classes: 1 = magazine articles; 2 = newspaper reports; 3 = radio talks; 4 = magazine articles for young people; 5 = current affairs radio programmes (radio reports); 6 = newspaper articles; 7 = magazine news items; 8 = newspaper news items; 9 = radio serials for young people (narrative frames); 10 = radio serials for adults (narrative frames); 11 = radio serials for schools (narrative frames); 12 = newspaper sports commentaries; 13 = radio news reports; 14 = oral narratives; 15 = radio discussions; 16 = radio interviews (interviewees' speeches); 17 = reports of committees; 18 = radio sports commentaries; 19 = laws; 20 = advertisements

Of the 20 text classes listed above, I have investigated more closely those incorporating articles, news items and speeches, each being represented by two text classes acting as controls one for the other: i.e. newspaper articles and magazine articles, news items in newspapers and news items in magazines, and radio discussions and oral narratives in unscripted dialogues. These six text classes were used to analyse the role of 194 linguistic variables, with respect to which the clustering analysis grouped the classes in the manner depicted in Fig. 2.

Figure 2
Tree diagram of the clusters according to 194 linguistic variables

This more extensive cluster analysis shows the pairs of text classes to form clearly distinct text types, in spite of the obvious differences between the members of each pair. Also, the speeches differ more markedly from the articles and news reports than the latter do from each other. The result is similar in direction to that obtained in the previous diagram.

Factor analysis

Factor analysis formed eight factors with a loading of at least 0.6 from among the 50 linguistic variables chosen to evaluate the corpus of 20 textual classes mentioned above. These served to explain 91% of the total variance. Factors 1 and 2, accounting for 54% of the variance, cause the text classes to be located in the factor space in the following manner (Fig. 3):

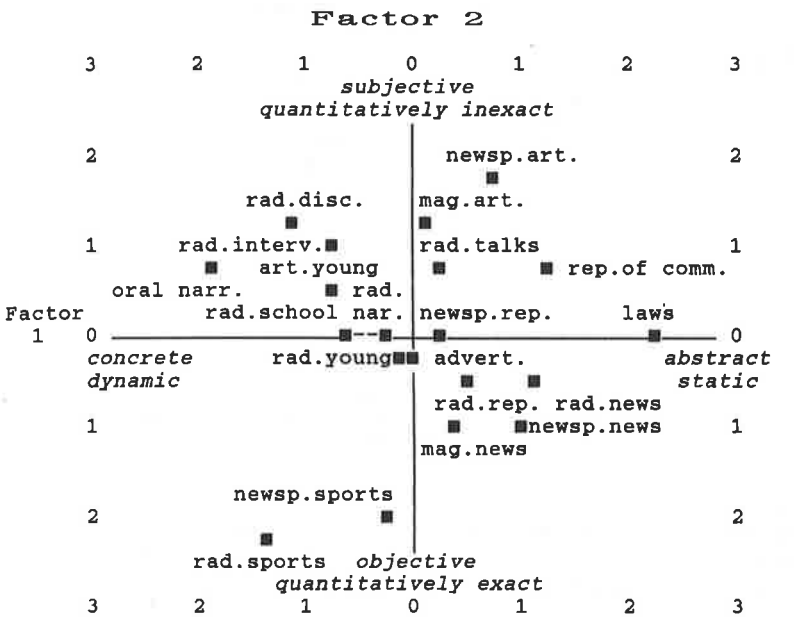


Figure 3. Factor scores according to 50 linguistic variables. Factors 1 and 2

The more restricted material comprising six textual classes of articles, news reports and speeches analysed in terms of 194 variables yielded five factors, which together explain all the observed variance, i.e. 100%. The first two of these together explain 74% of the variance and form coordinates on which the text classes are located as follows (Fig. 4):

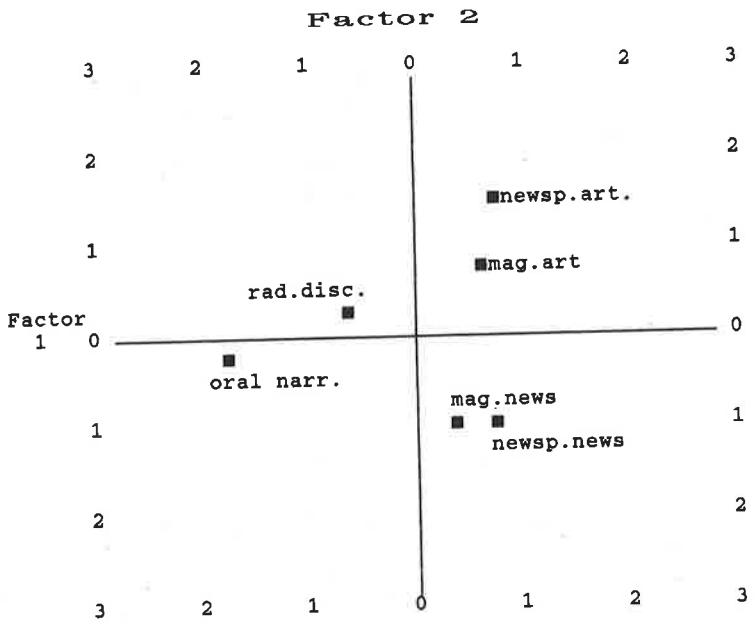


Figure 4. Factor scores according to 194 linguistic variables. Factors 1 and 2

Note that the articles, news reports and speeches occupy almost exactly the same positions in both diagrams in spite of the different sets of variables used and the different composition of the material. This suggests that factors 1 and 2 must be roughly similar in nature in the two cases. Therefore I will try to interprete only the set of 50 variables.

The composition of the first four factors, explaining 77% of the variance, in terms of the 50 variables used to study the full range of 20 text classes is shown in Table 1.

Table 1
Sorted rotated factor loadings. Loadings less than 0.250 are replaced by zero.

	Variable	Factor 1	Factor 2	Factor 3	Factor 4
1.	Compound words	0.938	0.000	0.000	0.000
2.	Length of Words	0.923	0.000	0.000	0.000
3.	Genitives	0.914	-0.267	0.000	0.270
4.	Contracted Sentences	0.906	0.000	0.000	0.000
5.	Modifiers	0.891	-0.351	0.000	0.000
6.	Predicate Verbs	-0.880	0.000	0.000	0.000
7.	Adverbs	-0.866	0.346	0.000	0.000
8.	Clause Length	0.816	0.000	-0.376	0.000
9.	Modifying Genitive	0.793	-0.379	0.000	0.318
10.	Denominal Verbs	0.787	0.000	0.429	0.000
11.	Translatives	0.782	0.312	0.000	0.000
12.	Pronoun Modifiers	-0.710	0.432	0.000	0.000
13.	Temporal Structures	0.703	0.000	0.000	-0.256
14.	Pronouns	-0.692	0.546	0.000	0.000
15.	Interjections	-0.691	0.251	0.000	0.000
16.	Postpositions/Prepos.	0.668	0.000	-0.280	0.430
17.	Nominatives	-0.652	-0.574	0.000	0.000
18.	Objects	0.621	0.593	0.304	0.000
19.	Predicatives	-0.618	0.458	0.457	0.000
20.	Essives	0.607	0.000	0.000	0.000
21.	Abessives	0.601	0.000	0.000	-0.348
22.	Conjectural Sentences	-0.571	0.000	-0.441	0.000
23.	Infinitive Constr.	0.000	0.912	0.000	0.000
24.	Noun Modifiers	0.000	-0.893	0.000	0.000
25.	Predicative Nominat.	0.000	-0.866	0.000	0.000
26.	Partitives	0.000	0.857	0.000	0.000
27.	Partitive Objects	-0.332	0.856	0.000	0.000
28.	Numerals	0.335	-0.820	0.000	0.000
29.	Adjective Modifiers	-0.470	0.741	0.000	0.000
30.	Subordinating Conj.	0.000	0.740	0.000	0.000
31.	Negative Clauses	0.000	0.648	-0.323	0.000
32.	Object Clauses	-0.315	0.632	0.000	0.581
33.	Partitive Subjects	-0.379	0.617	-0.402	0.000
34.	Adjectives	0.467	0.530	0.400	0.000
35.	Imperatives	0.000	0.000	0.867	-0.281
36.	Explicit Clauses	0.000	0.000	-0.845	0.353
37.	Temporal Clauses	0.000	0.000	0.765	0.000
38.	Asyndetic Lists	-0.433	0.000	0.699	-0.251
39.	Coordinating Conj.	0.000	0.000	0.266	-0.880
40.	'That' Clauses	0.000	0.000	0.000	0.866
41.	Quotations	0.340	0.000	0.000	0.850
42.	Referative Constr.	0.413	0.000	0.000	0.788
43.	Deverbal Verb Deriv.	0.271	-0.260	0.000	0.000
44.	Denominal Noun Deriv.	0.319	0.000	0.000	0.000
45.	Adj. as Adj. Modif.	-0.297	0.000	0.000	0.000
46.	Questions	0.000	0.000	0.000	0.000
47.	Assertions	0.000	0.000	-0.566	0.267
48.	Instructive Cases	0.401	0.000	0.000	-0.301
49.	Conditional predicates	0.000	0.610	0.000	0.265
50.	Contracted Modal Cl.	0.436	0.000	0.000	-0.251

Factor 1

Factor 1 comprises the first 22 variables as listed here and involves negative loadings on predicate verbs, adverbs, pronoun modifiers, pronouns, interjections, nominatives, predicatives and conjectural clauses and positive loadings on the remainder. What is that dimension which gives rise to this same quantitative distribution in 20 distinct types of text?

The following variables with a positive loading have at least one feature in common, namely that they all represent tightly knit, complex syntactic constructions: compound words (variable no. 1), genitives in general (no. 3) and modifying genitives (no. 9) in particular (genitives are used in Finnish principally to modify head nouns and in connection with postpositions and prepositions), modifiers (attributes) in general (no. 5), contracted sentences (with agent expressed) (no. 4) and in particular those representing temporal structures (no. 13), and denominal verbs (no. 10). As may be expected, these features show a positive correlation with word and clause length (no. 2 and 8). All these elements express more extensive, synthetic nominalized or verbalized entities which incorporate semantic relations denotable analytically in terms of clauses or phrases, e.g.:

The compound word **kattolamppu**, 'ceiling light', implies a relation between its parts, namely 'the light belongs on the ceiling'.

The attribute construction **valkoinen talo**, 'the white house', implies the predicative relation 'the house is white'.

The verbal root **lyhe-ntä-ä**, 'shorten', corresponds to the phrase **tehdä lyhye(mmä)ksi**, 'make short(er)', the focus being upon either the action, 'make something', or the result, 'short, shorter'.

The contracted sentence **kirjoitta-essa-ni**, 'while writing', corresponds to the clause **kun kirjoitan**, 'while I am writing', in which the focus may be on either the situation, 'while', or the action, 'write'.

Thus certain semantic roles overtly present in analytical clauses or phrases are found here in embedded positions, incorporated as sub-features of synthetic nominalizations or verbalizations. The analytical alternatives express the content implied by the semantic space in a more actual, more concrete, more explicit and more rhematic form, whereas the synthetic expressions are more abstract and less actual, so that the corresponding reality, and particulary the relations or roles operating between its parts, are not revealed so explicitly or with such precision or emphasis. Reduction of the predicate part of a clause in particular

implies a loss of actuality and a movement towards nominality and a static situation. Thus these nine variables may be said to represent **abstract, synthetic properties in a text**, a distanced, superstructural perspective on the world.

The translative case (no. 11) expresses purpose or result, the passive or static member in a process (**Hän aikoo opettaja-ksi**, 'He intends to become a teacher'), and postpositions and prepositions (no. 16) are also used in these texts principally to denote purpose or an abstract object of some action (e.g. **varten**, 'for', **puoleen**, 'on behalf of', **vastaan**, 'against', **mukana**, 'along with', **lähelle**, 'close to (motion)', **kohti**, 'towards'. These aspects of purpose, result and abstract objects of actions call for the imaginative, synthetic perception of more extensive portions of reality, taking place over a longer stretch of time, rather than an actual, concrete observation made "here and now". In this sense these two variables may also be assigned to the above group of structures adding to the abstract nature of the text.

The loading on objects (no. 18) and the essive (no. 20) and abessive cases (no. 21) is weaker and their importance less marked, nor is it at all clear how they are related to the above group. The relevance of objects may nevertheless be explained by noting that a text that contains a relatively high proportion of these may be said to be inclined to view the world through the perspective of passive or static participants in processes, i.e. the patients, objects, results and goals of these processes. This is to a certain extent connected with the question of translatives, postpositions and prepositions mentioned above, and with the perception of reality over a longer temporal span.

The essive (no. 20) mostly denotes location in time, e.g. **tänä syksynä**, 'this autumn', **tiistaina**, 'on Tuesday', or in a state of being, e.g. **sairaana**, 'being ill', **johtajana**, 'being the leader'. Like the translative, it is a local case that has taken on an abstract function and does not carry as concrete a meaning as the local cases proper. The latter type, expressing a state of being, is particularly abstract and static.

Little weight should be attached to the abessive (no. 21) in this interpretation as a whole, since it is of very low frequency, although it does represent a tightly snythetic construction, usually of the type **Hän piti päänsä muiden mielipiteistä välittämättä**, 'He kept his head, without heeding other people's opinions', corresponding to a clause: **Hän piti päänsä eikä (niin että ei) välittänyt muiden mielipiteistä**, 'He kept his head and (so that he) did not heed other people's opinions'.

Thus all the variables receiving a positive loading in factor 1 can in one way or another be traced back to a common property of a semantically relatively abstract, complex, synthetic and static perception of reality. But before we can finally regard this interpretation as correct, we must also consider whether the variables that gained a negative loading then represent the opposite property, as should logically be the case.

The highest negative loading is on predicate verbs (no. 6), a frequent occurrence of which is usually taken to imply that the text depicts reality in a structurally simpler manner, more analytically, more dynamically on average, temporally bound as far as possible, in the form of an actual action or state viewed from a given moment in time, in other words as a deictic expression, related to concrete experience and expressing an explicit relation. It is under these circumstances that the subject matter is viewed from a closer perspective. In this sense the variable does represent the opposite of the abstract constructions discussed above.

The total frequency of adverbs (no. 7) is largely determined by the occurrence of place and time adverbs, which account for the majority of this word class. These are relatively concrete, simple forms, and are usually deictic in character. The most common adverbs are of the kind **täällä**, 'here', **siellä**, 'there', **nyt**, 'now', and **sitten**, 'then, next'.

Abundant use of pronouns (no. 14) is indicative of redundancy and a more analytical, more explicit mode of expression, making the text more suggestive, introductory and more concrete. When occurring in a modifying position (no. 12), pronouns possess defining, individualizing, emphasizing and concretizing functions: **tässä tapauksessa**, 'in this case', **se asia**, 'that matter', **ketkä henkilöt**, 'which people'.

The interjections (no. 15) in the present material are practically discourse particles which serve to organize the stream of speech in the spoken textual classes: **no, jaa, joo, juu, niin, tuota, kyllä**, 'well, so, yes', etc. These break up the text into more analytical sequences and provide the listener with a simple, direct feedback. Although these again have the effect of rendering the text more concrete, they exist to some extent on a different plane from the other variables mentioned here and are peripheral to the identification of the factor concerned since they contain signals that are concerned only with the process of text production.

The nominative case (no. 17) and predicative function (no. 19) are linked together, since the nominative is the case form used for subjects and predicatives. The nominative may be said to be the simplest and

most explicit way of naming aspects of the real world. In the present texts subjects and predicatives contrast with objects and receive the opposite factor loading. A concentration on subject functions implies increased emphasis on the initial participant in actions and states, the raising of a crucial argument, a viewing of events from the perspective of active, dynamic effectors, variables or processors, or elements that merely exist in the situation. If a process is not perceived as involving an object it is looked on as being simpler and of shorter duration, while a process of longer duration need not be concieved of as having a goal when viewed in terms of a momentary observation. In this sense a hypothesis may be put forward that lack of an object implies a more actual, more straightforward and more concrete mode of perception. Likewise the predicative, as an expression of a static state, naturally does not imply an abstraction from a longer time span. These speculations do not necessarily have to be assigned very much weight, as the loadings of these variables are not as powerful as the loadings of the earlier variables.

The final variable, conjectural sentences (no. 22), does not seem to be connected in any way with the main character of the factor, which can now be designated as the dimension of **abstract, synthetic, static vs. concrete, analytical, dynamic**. The texts studied arrange themselves on this dimension in the manner shown in Fig. 3, with the speeches and spoken accounts in general at the more concrete end and laws, committee reports and news items at the more abstract end.

Factor 2

Factor 2 comprises the next 12 variables in Table 1, infinitive constructions, subordinating conjunctions, etc. The infinitive constructions (no. 23) are predominantly modal structures, expressing subjective perspective, i.e. the degree of certainty of a statement being true or coming true on scales of 'possible' – 'necessary' (**voi tehdä**, 'may do', **täytyy olla**, 'must be') or evaluation on the scale 'positive' – 'neutral' – 'negative' (**on hyödyllistä tehdä**, 'it is advantageous to do').

The more precise statistical frequency distribution of subordinating conjunctions (no. 30) shows the principal weight to be attached to conditional 'if' clauses, causal 'because' clauses, concessive 'although' clauses, consecutive 'so that' clauses, final 'in order that' clauses and com-

parative 'than/as' clauses. These also express either verification or evaluation, the majority being concerned with verification of the certainty of the world of discourse in terms of 'cause and effect' relations. The concessive clauses have a verificational meaning in the sense of contrastive negation, 'in spite of the fact that'; while the comparative clauses have an evaluative force.

Negative clauses (no. 31) also represent verificational modality, while adjectives (no. 34) and adjective modifiers (no. 29) have largely an evaluative function in the above scheme. The loading on the conditional (no. 49) shows it to belong firmly to factor 2 (although it is also associated slightly more strongly with factor 7), and this modus is obviously expressive of a verificational attitude.

All the above variables with a positive loading are intercorrelated and denote a single common property of texts, subjective perspective. Other variables correlated with them include the partitive case (no. 26), and in particular partitive objects (no. 27) and partitive subjects (no. 33). These are indicators of indefinite quantity and imprecision. In the sentence **Pihalla leikkii lapsia**, 'There are children playing in the yard', for instance, indefinite, inexact quantity is attached to both the playing and the children by virtue of the partitive subject, while correspondingly in **Hän syö aamiaista**, 'He is eating breakfast', indefinite quantity is attached to both the eating and the breakfast by virtue of the partitive object.

The negative equivalents of the above are noun modifiers (no. 24), predicative nominatives (no. 25) and numerals (no. 28). In contrast to predicative partitives, predicative nominatives imply quantitative accuracy and definiteness, while the use of numerals naturally also enhances quantitative precision, forming an objective counterpart to the variables indicative of subjectivity. The noun modifiers form a more objective alternative correlating negatively with the more subjective adjectival modifiers among the various modifier structures.

The remaining category, object clauses (no. 32), lacks any clear explanation, but the other variables may indeed be said to form a clear dimension of **subjectivity and quantitative indefiniteness and imprecision vs. objectivity and quantitative definiteness and accuracy**. As seen in Fig. 3, the most subjective and quantitatively imprecise among the texts in terms of factor 2 are the articles and speeches, while the most objective and quantitatively accurate are the sports reports and news items.

Factor 3

Factor 3 again represents in broad terms an **abstract–concrete** dimension, but of a different type. Where concreteness became apparent in factor 1 in the property of explicitness, it is **implicit**, in the form of a presupposition, in factor 3. A certain implicit concreteness and straightforwardness, an itemization of reality, arises form the use of imperatives (no. 35), temporal clauses (no. 37) and asyndetic lists (no. 38), as are typical of advertisements in particular. Imperatives, although at the same time incorporating subjective perspective, are associated with a concrete situation in which an appeal is made to the listener or reader, while temporal subordinate clauses are used to tie an event down to a specified time. Asyndetic, implicit juxtaposition of items consists of no more than a list, in which such a straightforward, unambiguous, specific relation obtains between the itemized members that it does not even have to be made explicit by means of a conjunction. These show a negative correlation with explicit clauses containing predicates (no. 36), while their counterparts, incomplete clauses with no predicate relation, are again so unambiguously simple that they can easily remain without any predicate provided they receive adequate contextual support (typically in advertising text).

Factor 4

Factor 4 is largely a relational dimension. Coordinating conjunctions (no. 39) express **static relations of coordination** and the other variables **subordination**, embedding relations to be regarded as more **dynamic** in character. Clauses introduced by **että**, 'that' (no. 40), quotations (no. 41), either direct or indirect, from someone else's words, and contracted referative construction (no. 42) expressing indirect speech (**Hän sanoi tulevansa**, 'He said he was coming') usually serve as either subjects or objects in clauses.

The textual classes are arranged in a space defined by factors 3 and 4 in Fig. 5. Our interpretation of factor 3 shows advertisements and radio sports commentaries to be the most concrete in terms of situational reference, while texts of the news report type and newspaper articles are the least so. With respect to factor 4, the most coordinative or static view of the world is presented by legal texts and newspaper articles and

the most subordinative and dynamic view by radio news broadcasts and current affairs programmes (reports).

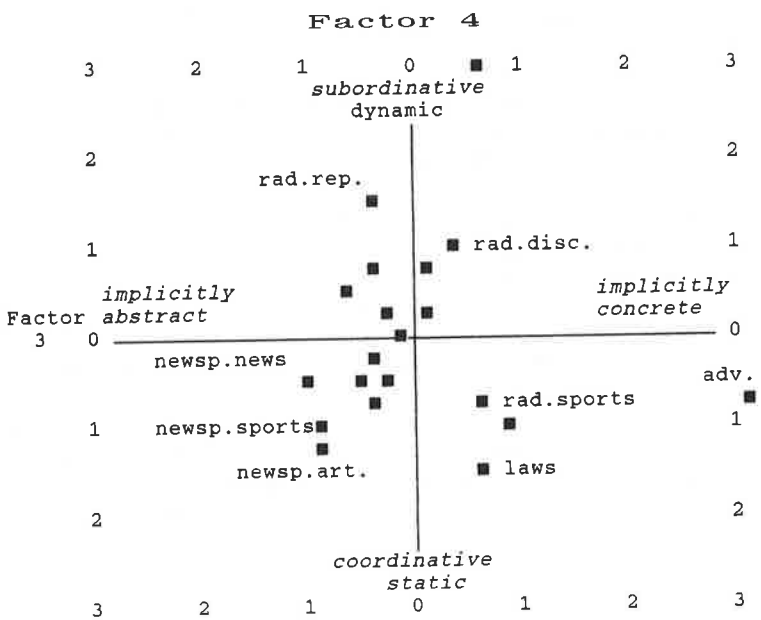


Figure 5. Factor scores according to 50 linguistic variables (only the most extreme text classes are named). Factors 3 and 4

I leave the remaining, more peripheral factors out of discussion. They can be explained as more restricted subdimensions of the above general dimensions. I exclude also the factor analyses performed on the six specially selected text classes in terms of 194 variables. Briefly stated, they led to the same principal dimensions as detailed above.

Although the interpretations of the individual variables may not always be entirely correct, it is possible to claim as a general conclusion that the dimension types and the whole system to which they belong would seem to have emerged in a certain incontrovertible manner. This system is broadly of the following form:

Objective perspective
Accuracy of the world
Qualitative accuracy
Concrete - abstract arguments
Dynamic - static relations
Quantitative accuracy
Inexact - exact

Subjective perspective
Certainty of the world
Possible - necessary
Value of the world
Positive - neutral - negative

(Cf. Saukkonen 1986, 1987, 1988)

A text may be to a greater or lesser extent subjective or objective in character, with subjectivity subdimensions of verification/directiveness and evaluation. The objective dimension expresses the degree of accuracy with which the real world is viewed, in what depth of detail, in what breadth, to what extent the text aims at a concise overview, or with what degree of concrete illustration or abstract conceptualization the theme is treated. Here distinctions may be made between qualitative and quantitative accuracy and between accuracy of arguments and accuracy of relations. It is my belief that semantic dimensions of this kind can be used to construct a rhetorical grammar based on the semantics of language usage.

References

Biber, Douglas (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics 23, 337-360.*

Kraus, Jiří & Polák, Josef (1967). Text factors and characteristics. *Prague Studies in Mathematical Linguistics 2, 155-171.*

Saukkonen, Pauli (1982). Oulun korpus. Research Reports of the Department of Finnish and Lappish. Oulu, University of Oulu.

Saukkonen, Pauli (1986). Flexionskategorien der finnischen Sprache als Indikatoren der Sprecherperspektive. *Finnisch-ugrische Mitteilungen 10, 353-358.*

Saukkonen, Pauli (1987). Discourse types - cultural norms. Forthcoming in the Proceeding of the 8th World Congress of Applied Linguistics in Syndey 16 - 21 August 1987.

Saukkonen, Pauli (1988). Die Perspektive als Quelle des Stils. *Göttingische Gelehrte Anzeigen 240, 156-172.*

Hřebíček, L. (ed.), Glottometrika 11, 1989

Stochastic Model of Sentence Structure in Japanese Literature

Akio Tanaka, Tokyo

0. Introduction

The main aim of this article is to consider what would be the result if the sentence structure of Japanese were assembled in a completely machine-like manner.

Put simply, one could say that the structure of a sentence is formed by the first phrase at the beginning fo the sentence being decided, the second phrase being selected as a result of this, and then the kind of phrase able to occur as the third phrase being decided due to the fact of the second clause having been decided, and so on, until the Full Stop at the end of the sentence is reached.

For example, we postulate that, if a sentence starts with '- ga' as a first phrase, what sort of phrase will occur as the second phrase will be strongly influenced by the fact that '- ga' occurred as the first phrase. Consequently, we hypthesise that if a phrase with the form '- o' occurs as the second phrase, this will influence what form of phrase will occur the third phrase. We postulate that a sentence is formed in this way from the first phrase at the beginning of the sentence with each phrase influencing the phrase after it, until the end of the sentence.

This means that when one cinsiders the formation of a sentence in this way, one can obtain the probability of a particular structure being fromed, using the product of the Probabilities of Transition from phrase to phrase.

On the other hand, when one is considering a structure from the end of the sentence, one may follow the sorts of phrases which are likely to come before a Sentence Final Phrase as a result of it having been decided as a particular type and what phrases are likely to come before them as a result of their having been decided, until one reaches the beginning of the sentence. Thus, as in the above situation, when a specific type of Sentence Final Phrase is selected, one can obtain the

kinds of structure easily formed using the Probabilities of Transition from one phrase to another.

Consequently, we see sentences as being fromed as a result of a process of probabilities of phrases and we have in fact tested a technique which produces typical Japanese and English sentences based on the Probability of Transition of word units or syllable units, which is used in the field of Computational Linguistics. Using the terminology of Computational Linguistics, structures obtained in this way are called 'Approximate Sentence Structures of Phrase Units'.

As one can obviously view sentence structures produced with such a technique as structures which occur very easily, the consideration of phenomena in sentence structure thus revealed is an important task in the clarification of basic facts in syntax.

Furthermore, grasping the ease of continuation from one phrase to another, in other words, the Probability of Transition, will give valid clues in the consideration of research into word order and agreement phanomena, and the establishment, etc. of Formation Rules in Phrase Structure Grammars.

1. Methods Used In The Survey

We obtained the data for this research from Naoya Shiga's 'Ki no Saki nite' and Yûzô Yamamoto's 'Sutoo Huzin'. We chose these because we thought that, as the former is written in Plain Style aimed at the general reader and the latter in 'Desu/Masu' Style aimed at young people, differences in this respect may be reflected in the results of the survey.

We should first mention our treatment of passages of dialogue. Obviously, the forms of sentence structure in dialogue passages and narrative passages are different. Moreover, broken sentence structure, so called incomplete sentence structure often occurs in dialogue passages. Bearing such things in mind, it is from the start dangerous to deal with dialogues, so for this survey we removed all dialogues from the object fo the survey and dealt with places where dialogues are quoted in the course of the narrative as follows

>Kare wa 'boku mo iku' to syutyoo sita.
->kare wa / (Dialogue) to / syutyoo sita/

>'Kirei na hana ne' nado to, kanozyo wa itta.
->(Dialogue) nato to / kanozyo wa itta /

In other words, we recognized '(Dialogue) + to' and '(Dialogue) + nado to' as sorts of phrases.

As a result of carrying out the above treatment, the number of sentences which are the object of thes survey is 209 in 'ki no saki nite' and 774 in 'Sutoo Huzin'.

Concerning the content of what we have so far refered to until now as 'phrases', this refers to the following sort of treatment after dividing the sentence into phrases, following the concept of 'bunsetu' (phrase) in Hashimoto Grammar.

BS ——>Yuki no——>yama ga——>takusan——>mieta——>.

T ——>(N + no)——>(N + ga)——>(Adv)——>(V + ta)->.

Thus in this survey, we have not treated the actual phrases themselves, such as 'yuki no', and 'takusan', etc.; we have converted these into 'Noun + no' and 'Adverb', etc., and then dealt with them.

T in the above example is the indication of the beginning of the sentence and '.' is the 'Full Stop' which indicates the end of the sentence.

2. Phrase Entropy

The concept of entropy is usually employed concerning the probability of letters and symbols, etc. following each other. However, here it is used in the sense of how much uncertainty there is about the next phrase when a phrase has been decided. Consequently, when a phrase always follows a particular phrase, in other words, when there is a Probability of Transition of 1, its entropy is smallest. For example, if an Attributive Phrase is always followed by a Noun Phrase, this would be a case of the Probability of Transition being 1 and the entropy being 0.

Considering sentence structure using such 'Phrase Entropy' is, in fact, put simply, viewing sentences as a system of sequences of the Probabilities of Transition from one phrase to another. Hence, the first problem is grasping how phrases appear in actual sentences and, moreover, how they are linked up and used.

Thus, if we first observe how phrases occur in the two works we have taken up, the token number of phrases in 'Ki no Saki nite' was 1635 and the type number 168. In 'Sutoo Huzin' the token number was 8405 and the type number 557.

If we take the phrases with a high frequency of occurrence from among these, Table 1 is obtained. In 'Ki no Saki nite', the most frequent phrase was of the form 'Verb + ta' and after that came 'Adverb Phrase', 'Noun + wa', etc. In 'Sutoo Huzin', 'Noun + no' came first, followed by 'Noun + ni', 'Noun + wa', etc., with noun-type phrases consistently occupying the highest frequencies. This fact may reflect a difference in the two styles.

Table 1
Occurrence Rate of Phrase Types

1.1

Ki no saki nite		
Phrase Type	Freq.	%
V + ta	132	8.07
Adv	128	7.83
N + wa	122	7.46
N + no	120	7.34
V + te	116	7.09
N + ni	109	6.67
V	102	6.24
N + o	92	5.63
N + ga	84	5.14
A	53	3.24
Attrib.	41	2.51
N	38	2.32
AV	33	2.02
N + e	29	1.77
C	27	1.65
N + mo	25	1.53
...	...	...
Total	1635	

1.2

Sutoo Huzin		
Phrase Type	Freq.	%
N + no	898	10.68
N + ni	552	6.57
N + wa	549	6.53
N + o	529	6.29
V + te	520	6.19
Adv	476	5.66
N + ga	410	4.88
V	391	4.65
Attrib.	305	3.63
N	247	2.94
V + masita	230	2.74
A	219	2.61
V + ta	202	2.40
AV	162	1.93
N + to	151	1.80
N + mo	128	1.52
N + de	121	1.44
V + masen	81	0.96
V + masu	79	0.94
V + to	73	0.87
N + niwa	71	0.84
N + kara	71	0.84
N + desu	50	0.59
V + de	45	0.54
V + rete	43	0.51
N + de	42	0.50
N + desita	32	0.38
N + de	31	0.37
V + yoona	31	0.37
...	...	...
Total	8405	

Table 2
Distribution by Word Classes of Phrase Types

Phrase Type	Ki no Saki nite		Sutoo Huzin	
	Freq.	%	Freq.	%
Noun-Type	767	46.91	4308	56.26
Verb-Type	537	32.84	2567	30.54
Adj-Type	73	4.46	312	3.71
A-V-Type	37	2.26	170	2.02
Adv-Type	153	9.36	552	6.57
Attrib-Type	41	2.51	320	3.81
Conjunction	27	1.65	134	1.59
Exclamation	0	0	29	0.35
Quotation	0	0	13	0.15
...	...	...	...	...
Total	1635		8405	

Table 3
Probability of Transition between Phrase Types

Table 3.1. 'Ki no Saki nite'

Order	Preceeding--Succeeding	Freq.	Prob. of Transition	
			---->	<----
1	(V + te) -- (V + ta)	38	32.67	28.79
2	(N + no) -- (N + ni)	26	21.67	23.85
3	(N + wa) -- (Adv)	25	20.49	19.53
4	(V + te) -- (V)	22	18.97	21.57
5	(N + ni) -- (V + te)	20	18.35	17.24
6	(N + o) -- (V + ta)	16	17.39	12.12
7	(N + o) -- (V + te)	15	16.30	12.93
8	(V) -- (N + ga)	13	12.75	15.48
8	(N + no) -- (N + no)	13	10.83	10.83
8	(N + ga) -- (V + te)	13	7.07	11.21
11	(N + ga) -- (Av)	12	6.57	9.38
11	(N + no) -- (N + ga)	12	10.00	14.29
11	(Adv) -- (N + ni)	12	9.38	11.01
11	(Adv) -- (V + te)	12	9.38	10.34
15	(V) -- (N + o)	11	10.78	11.96
15	(N + no) -- (N + o)	11	9.17	11.96
15	(N + ni) -- (V + ta)	11	10.09	8.33
15	(Adv) -- (N + no)	11	8.59	9.17
19	(N + o) -- (V)	10	10.87	9.80
19	(Attrib) -- (N + o)	10	24.39	10.87
..	...	...	...	...

--> shows the probability from the preceding phrase
<-- shows the probability from the succeeding phrase

However, if one shows the difference in occurrence of phrases in the whole of both works by arranging them by word class, one arrives at Table 2 and there is hardly any difference between the two works in the rates of the wholes.

The analysis and investigation of such points is not the subject of this article, so we will confine ourselves to a general view of the data. Next we will observe the combination of such phrases.

Table 3.2. Satoo Huzin

Order	Preceeding--Succeeding	Freq.	Prob. of Transition	
			---->	<----
1	(N + no) -- (N + ni)	143	15.92	25.91
2	(N + o) -- (V + te)	134	25.33	25.77
3	(N + ni) -- (V + te)	113	20.47	21.73
4	(N + no) -- (N + o)	110	12.25	20.79
5	(N + wa) -- (Adv)	104	18.94	21.85
6	(N + no) -- (N + ga)	93	10.36	22.68
7	(N + wa) -- (N + no)	89	16.21	9.91
8	(N + o) -- (V)	87	16.45	22.25
9	(N + no) -- (N + wa)	82	9.13	14.94
10	(N + no) -- (N + no)	81	9.02	9.02
10	(Adv) -- (N + no)	81	17.02	9.02
11	(Attrib) -- (N + no)	71	23.28	7.91
12	(V + te) -- (V+masita	70	13.46	30.43
13	(V + te) -- (V)	67	12.88	17.14
14	(V) -- (N + ga)	59	15.09	14.39
15	(V) -- (N + no)	56	14.32	6.24
16	(V) -- (N + wa)	53	13.55	9.65
17	(N + ni) -- (V)	50	9.06	12.79
..	...	...	...	...

--> shows the probability from the preceding phrase
<-- shows the probability from the succeeding phrase

The proportions in Table 3 were calculated as follows:

Fr(V+te) = 116, Fr(V+ta) = 132; Fr[(V+te)--(V+ta)] = 38

Prop[(V+te) --> (V+ta)] = (38/116)100 = 32.76
Prop[(V+te) <-- (V+ta)] = (38/132)100 = 28.79.

Table 3 shows the forms of combination of phrases. Thus, if we look at the token number of 168 types of phrases which occur in 'Ki no Saki nite', the form of combination which occurred most frequently was 'Verb + te' and 'Verb + ta'. This is followed by '(Noun + no) - (Noun + ni)' and '(Noun + wa) - (Adverb Phrase)'. In 'Sutoo Huzin', the form of com-

bination of 'Noun + no' and 'Noun + ni' occurred most frequently; followed by '(Noun + o) - (Verb + t)' and '(Noun + ni) - (Verb + te)', etc. If we survey Table 3, predictably, we find that the kinds of combinations one would expect from common sense occupy the top of the table and these occur very frequently. Put another way, there are 168 types of phrase in 'Ki no Saki nite', so the square of 168 combinations could be made, and with 'Sutoo Huzin', the square of 557 combinations could exist. However, in fact, a very limited number of combinations are just repeated. For example, 'Noun + no' in 'Sutoo Huzin', as is shown in Table 1, is the phrase with the highest rate of occurrence. However, of its frequency of 898, over half are used in the form of being followed by one of only four types; 'Noun + ni (143 times)', 'Noun + o (110 times)', 'Noun + ga (93 times)' and 'Noun + wa (82 times)'. Moreover, if we examine 'Verb + ta' in 'Ki no Saki nite', of its frequency of occurrence of 132, over half the phrases preceding it are of the types 'Verb + ta (38 times)', 'Noun + o (16 times)', 'Noun + ni (11 times)' and 'Adverb Phrase (6 times)'. Consequently, it is found that although logically various combinations may be predicted, it may be said that the types of combinations of phrases which occur in actual sentences are very limited. If we observe the Probabilities of Transition in Table 3, we can find four types. For example, the combination of 'Verb + te' and 'Verb + ta' in 'Ki no Saki nite' where the Probability of Transition from the Preceding Phrase and the Probability of Transition from the Succeeding Phrase are both high; a type like '(Adverb Phrase) - (Noun + no)' where they are both low, a type like the combination '(Attributive Phrase) - (Noun + no)' in 'Sutoo Huzin' where the Probability of Transition from the Preceding Phrase is high but from the Succeeding Phrase is low and, on the other hand, a type like '(V + te) - (V + masita)' where the Probability of Transition from the Preceding Phrase is low but from the Succeeding Phrase is high. Of course, the fact that both Probabilities of Transition are high shows that the combination of them is strong and the fact that they are both low indicates that the combination is weak. Moreover, the fact that, for example, only the Probability of Transition from the Preceding Phrase or only the Probability of Transition from the Succeeding Phrase is high reveals that the combination is strong from one side, but not necessarily so from the other.

What is important here is the fact that the condition of strength or weakness of a combination, as we see in Table 3, is not necessarily equal to its frequency. This hints at the dangers in research in syntax, especially in research into word order and in the consideration of agreement phenomena, of relying on only the frequency of occurrence when investigating the combinations and relations of words and phrases.

Above, we used two passages as material and examined the patterns of the combinations of phrases which occur in actual sentences and some of the aspects of this combination. It is very important to grasp and accumulate such facts in order to make syntax more substantial. We think that an approach using Probability of Transition is an effective method for doing this.

3. Sentence Initial Occurrence Rate

If we show the types of Sentence Initial Phrases in the two works we took up in order to discover what type of phrases sentences are likely to start with, Table 4 is obtained.

In both works, the form 'Noun + wa' occupies the top position, and, in 'Ki no Saki nite', of the total of 209 sentences, 57 start with this type of phrase. Moreover, in 'Sutoo Huzin' 17.05% of the total number of sentences are begun with 'Noun + wa'. We will call this rate the 'Sentence Initial Occurrence Rate'. The top of the list of Sentence Initial Occurrence Rates is made up of very common Sentence Initial Phrases.

Table 4
High Frequency Sentence Initial Phrases

Table 4.1. Ki no Saki nite

Order	SI Phrase	Freq.	%
1	N + wa	57	27.27
2	N + no	26	12.44
3	Av	21	10.04
3	Conjunction	21	10.04
5	Attrib	14	6.69
6	N + ga	10	4.78
7	N	6	2.87
8	A	5	2.39
8	N + de	5	2.39
8	N + ni	5	2.39
8	N + o	5	2.39
12	A-V	4	1.91
	...	...	...
	Total	209	

Total Number of Beginnings of Sentences (= Total Number of Sentences): 209

Table 4.2. 'Satoo Huzin'

Order	SI-Phrase	Freq.	%
1	N + wa	132	17.05
2	Conjunction	123	15.89
3	N + no	101	13.04
4	Attrib	78	10.07
5	Av	73	9.43
6	N + ga	41	5.29
7	N	36	4.65
8	N + o	25	3.22
9	Exclam.	20	2.58
10	N + ni	19	2.45
11	N + to	11	1.42
12	N + mo	10	1.29
	...	...	...
	Total	774	

Total Number of Beginnings of Sentences (= Total Number of Sentences): 774

Furthermore, there is little difference between these two works, or rather, although there is some amount of discrepancy, they are in absolute agreement at least for the top seven rates.

Moreover, such particles as 'wa', 'no', 'ni', 'ga' and 'o', which M. Nagao set up as particles likely to occur in Sentence Initial Phrases, as a result of examining the newspaper vocabulary survey data of the National Language Research Institute, were all present as Sentence Initial Phrases within the top ten positions in the table for the two works we surveyed.[1]

When this is taken into consideration, it may be thought that the types of Sentence Initial Phrase are fairly similar even in the case of different passages.

Next if we show what kind of phrase is likely to occur as the phrase after the Sentence Initial Phrase when the Sentence Initial Phrase has been decided, using 'Noun + wa' which is the most frequent Sentence Initial Phrase, Table 5 is obtained.

Thus, in 'Ki no Saki nite', of 122 occurrences of 'Noun + wa', 25 are followed by an 'Adverb Phrase'. In 'Sutoo Huzin', of 549 occurrences of 'Noun + wa', 18.94% are followed by an 'Adverb Phrase'. This rate is, of course, the Probability of Transition from 'Noun + wa' to 'Adverb

1 Koobun no Imi no Kaiseki no Kokoromi, M. Nagao. Keiryoo Kokugogaku, 64 Syuu (M.Nagao, Preliminary analysis for sentence structures and word semantics. Mathematical Linguistics 64, 1973.

Table 5
'Noun + wa' Succeeding Phrases

Table 5.1. 'Ki no Saki nite'

Order	Succeeding Phrase	Freq.	%
1	Adv	25	20.49
2	N + no	12	9.48
3	A	9	7.38
4	N + ni	7	5.74
5	A-V	5	4.10
6	N + de	5	4.10
6	N + o	5	4.10
8	Attrib	4	3.28
8	N	4	3.28
10	N + e	3	2.46
...	...	...	...
	Total	122	

Total Number of 'Noun + wa': 122

Table 5.2. Sutoo Huzin

Order	Succeeding Phrase	Freq.	%
1	Adv	104	18.94
2	N + no	89	16.21
3	N + ni	41	7.46
4	Attrib	33	6.01
5	N + o	30	5.46
6	A	24	4.37
6	N	24	4.37
8	A-V	16	2.91
9	N + to	12	2.18
10	N + mo	11	2.00
10	N + ga	11	2.00
12	V + masen	10	1.82
...	...	...	...
	Total	549	

Total Number of 'Noun + wa' Phrases: 549

Phrase'. When there is an Adverb Phrase, the most likely type of phrase to follow is 'Noun + ni' in 'Ki no Saki nite' and 'Noun + no' in 'Sutoo Huzin'. If we follow in this way the types of phrase which have the highest Probability of Transition from the Sentence Initial Phrase, we obtain the following kind of sentence structure in 'Ki no Saki nite' (the figures are the Probability of Transition):

$$T \xrightarrow{27.27} (N+wa) \xrightarrow{20.49} (Adv) \xrightarrow{9.38} (N+ni) \xrightarrow{18.35} (V+te) \xrightarrow{32.76} (V+ta) \xrightarrow{81.82}$$

It is as follows in 'Sutoo Huzin':

$$T \xrightarrow{17.05} (N+wa) \xrightarrow{18.94} (Adv) \xrightarrow{17.02} (N+no) \xrightarrow{15.92} (N+ni) \xrightarrow{20.47} (V+te) \longrightarrow$$
$$\xrightarrow{13.46} (V+masita) \xrightarrow{99.57}$$

This means that, in the case of 'Ki no Saki nite', when one reaches 'Verb + ta' the most likely thing to follow is the Full Stop and the operation ends. Likewise, in the case of 'Sutoo Huzin', when one has a 'Verb + masita' phrase, one arrives next at a Full Stop with a Probability of Transition of 99.57%.

Thus, the sentence structures we have made are typical sentence structures in the respective works, when we take the Probability of Transition of the phrases two at a time from the beginning of the sentence. When these two sentence structures are compared, they can be viewed as being almost the same structure, the only difference being whether there is a Plain Form or a Polite Form at the end of the sentence. If whether '(Noun + no)' occurs or not is viewed with '(Noun + no) -- (Noun)' being counted as equivalent to a Noun Phrase, this is not a basic difference in sentence structure.

Figure 6
Part of the Tree Diagram Starting with 'Noun + wa'

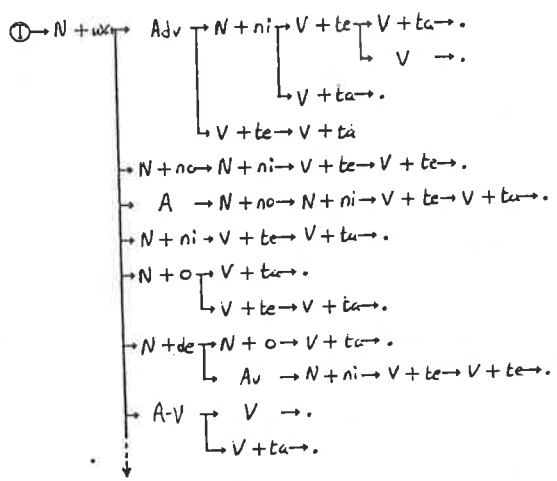


Figure 6.1. 'Ki no Saki nite'

Of course, we cannot conclude at this stage whether this is representative as a Japanese sentence pattern. However, we feel that it may be said that it is one type of sentence structure which is very likely to be formed, at least when a sentence structure based on the Probability of Transition between two phrases at a time is created.

Above we traced only the areas where there is the highest Probability of Transition in Sentence Initial Phrases of the 'Noun + wa'-type. However, as was shown in Table 5, even in the case of 'Noun + wa', various phrases may follow depending on the Probability of Transition. Hence, even if we make 'Noun + wa' the Initial Condition, a very large tree diagram may be drawn. Part of it is shown in Figure 6.

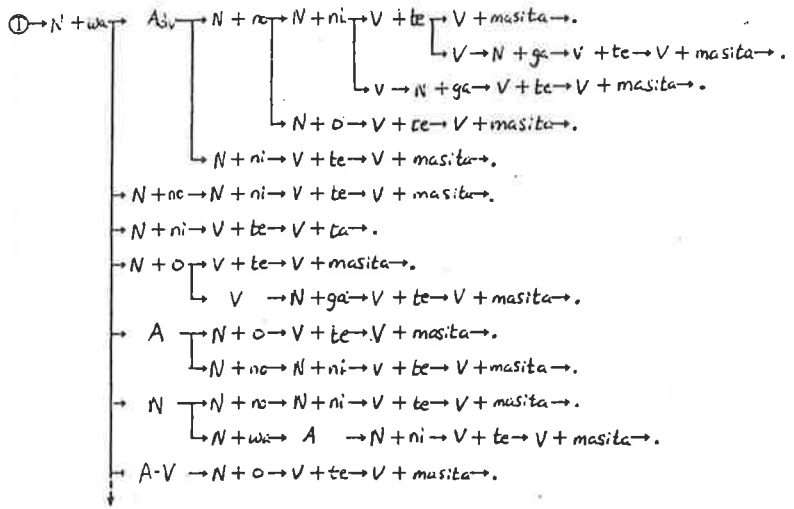


Figure 6.2. 'Sutoo Huzin'

The structure which first springs to mind involving 'Noun + wa' is the 'Zoo wa hana ga nagai' sort. However, this does not occur within the scope of Table 6. Perhaps this is due to the nature of the data. We feel that this sort of sentence structure is not likely to be formed when one follows through using this method. This will be taken up again in Section 6.

Moreover, what has been called the 'Rate of Occurrence' is in fact the Probability of Transition of the various Sentence Initial Phrases following T, the indicator of the beginning of the sentence.

4. Sentence Final Conclusion Rate

Table 7
The Most Frequent Sentence Final Phrases

Table 7.1. Ki no Saki nite

Order	S F Phrase	Freq.	%
1	V + ta	108	51.67
2	V	29	13.87
3	A + ta	7	3.34
3	V + nakatta	7	3.34
5	V + nai	6	2.87
6	N + da	5	2.39
6	N + datta	5	2.39
8	V + reta	4	1.91
8	A	4	1.91
10	A-V + ta	3	1.43
...	...	...	...
	Total	209	

Table 7.2. 'Sutoo Huzin'

Order	S F Phrase	Freq.	%
1	V + masita	229	29.58
2	V + masen	80	10.33
3	V + masu	79	10.20
4	N + desu	50	6.45
5	N + desita	32	4.13
6	V + masedesita	25	3.22
7	V + to no desu	24	3.10
8	N	18	2.32
9	N + da	8	1.03
9	N + desyoo	8	1.03
9	V	8	1.03
12	V + no desu	7	0.90
...	...	...	...
	Total	774	

If the Sentence Final Phrases of the sentences in the works, are surveyed as in the case of the Sentence Initial Phrases in the last section, Table 7 is obtained. When we examine this table, the Sentence Final Phrase with the highest frequency in both works is a Verb Phrase with past tense 'ta'. In 'Ki no Saki nite', of the total of 209 sentences, 108 are closed with 'Verb + masita'. If we call this rate the 'Conclusion Rate', Sentence Final Phrases with Past Forms and with Negative Forms

The Total Conclusion Rate of Past Form Sentence Final Phrases reaches 64.11% in 'Ki no Saki nite' and 40% in 'Sutoo Huzin', with those shown in Table 7 alone. Furthermore, the Total Conclusion Rate of Negative Form Sentence Final Phrases is 6.22% in 'Ki no Saki nite' and 13.57% in 'Sutoo Huzin'.

Logically, various types of Sentence Final Phrases could be predicted, but in the actual sentences, it seems that there is quite a heavy bias in favour of Past Forms, with Negative Forms in second place.

If the Sentence Final Phrases in Table 7 of the sentences in both works are compared, as expected, the difference between sentences in Polite Style and sentences in Plain Style is very clearly revealed. A particularly remarkable example of this is the fact that, whereas plain verb phrases occupy second position in 'Ki no Saki nite', this type of Sentence Final Phrase falls below the 1% level in 'Sutoo Huzin'. This is, of course, because 'Verb + masu' has covered them in 'Sutoo Huzin'. If the Sentence Final Phrases of both works at the top of the Conclusion Rate Table are compared, the sets of 'Noun + da' in 'Ki no Saki nite' and 'Noun + desu' in 'Sutoo Huzin', and 'Verb + nai/Verb + masen', 'Verb + nakatta/Verb + masen desita' and 'Noun + datta/Noun + desita', etc. can be made. The characteristics of the Sentence Final Phrases of both works can be said to be revealed by this.

Table 8
Preceding Phrase with the Highest Conclusion Rate

Table 8.1. 'Verb + ta' Preceeding Phrases in Ki no Saki nite

Order	Succeeding Phrase	Freq.	Prob. of Transition
1	V + te	38	28.79
2	N + o	16	12.12
3	N + ni	11	8.33
4	A	6	4.55
5	N + ga	5	3.79
5	N + de	5	3.79
5	N + no	5	3.79
8	V + de	4	3.03
8	A-V	4	3.03
...	...	...	...
	Total	132	

Total Number of 'Verb + ta': 132

Table 8.2. 'Verb + masita' Preceeding Phrases in Sutoo Huzin

Order	Succeeding Phrase	Freq.	Prob. of Transition
1	V + te	70	30.43
2	N + o	37	16.09
3	N + ni	23	10.00
4	N + ga	17	7.39
5	A	10	4.35
6	V + yoo ni	7	3.03
6	V + rete	7	3.04
8	N + de	6	2.61
...	...	...	...
	Total	230	

Total Number of 'Verb + masita': 230

Here we will try creating sentence structures with Sentence Final Phrases as a starting point, using the same method as that tried out in the last section with the Occurrence Rate of Sentence Initial Phrases. In the case of 'Ki no Saki nite', the phrase which is most likely to come before the Full Stop, in other words, the type of phrase with the highest Conclusion Rate is as shown in Table 7, 'Verb + ta'. Once 'Verb + ta' has occurred, the clauses which are likely before it are as in Table 8. From this table, we understand that once 'Verb + ta' has occurred, 'Verb + te' is most likely to come before it. If we repeat this operation until we arrive at T. the indicator of the beginning of the sentence, the following sentence structure is formed (the figures are the Probability of Transition).

$$\xrightarrow{51.67}(V+ta)\xrightarrow{27.54}(V+te)\xrightarrow{17.24}(N+ni)\xrightarrow{23.85}(N+na)\xrightarrow{21.31}T$$

If this is tried with 'Sutoo Huzin', we obtain the following:

$$\xrightarrow{29.58}(V+masita)\xrightarrow{30.43}(V+te)\xrightarrow{25.77}(N+o)\xrightarrow{20.79}(N+no)\xrightarrow{11.25}T$$

These are the typical sentence structures in both works which are formed when we follow the phrases with the highest Probability of Transition, two at a time, starting from the Full Stop. When the two sentence structures are compared, apart from a difference over whether the Object Case is 'ni Case' or 'o Case', they are exactly the same in construction in that they both have 'Noun + no' sentence initially

Figure 9
Part of the Tree Diagram Beginning with the Phrase with the Highest Conclusion Rate.

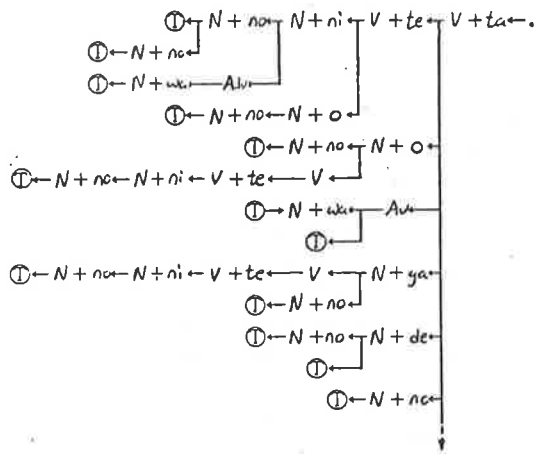


Figure 9.1. 'Ki no Saki nite'

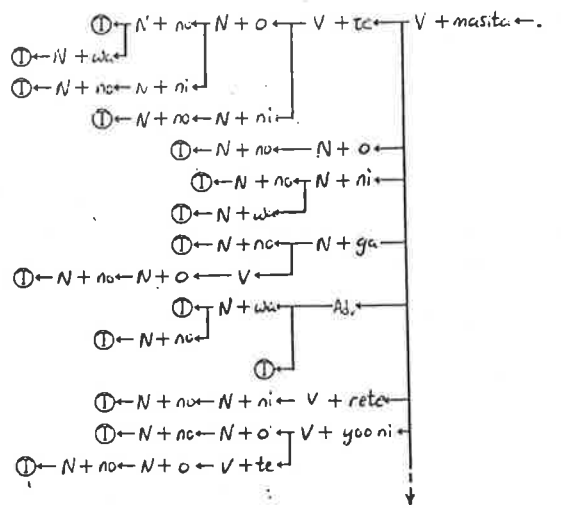


Figure 9.2. 'Sutoo Huzin'

followed by an Object Case, then 'Verb + te' and are concluded with a phrase with the verb in the Past Form.

There is room for discussion over whether these results are to be seen as particular to the works used here or whether they can be viewed as more general. Nevertheless, they cannot be seen as being coincidentally the same. It is noteworthy that structurally extremely similar sentence structures were obtained from different works using a completely mechanical operation.

Figure 9 is obtained when we try drawing a part of the tree diagram of the Sentence Final Phrase with the highest Conclusion Rate using the same method as we applied on the Sentence Initial Phrase 'Noun + wa' in Figure 6 of the last section. The interesting thing about this figure is that, although in Figure 9.2 of 'Sutoo Huzin' the word order 'ni Case -- o Case -- Verb'is apparent, the order 'o Case -- ni Case -- Verb' does not appear, at least within the range of Table 8. If we follow the Probability of Transition with these methods, the order 'o Case -- ni Case' seems not to occur as much as 'ni Case -- o Case'. This point will be taken up again in Section 6.

Furthermore, what has been called the 'Conclusion Rate' corresponds, of course, with the Probability of Transition from the Full Stop to the Sentence Final Phrase.

5. A Model of the Probability of Sentence Structures

The sentence structures which we tried to draw up in Sections 3 and 4 both followed only the succession of phrases with the highest Probability of Transition. However, for each type of phrase, there are many types of phrases which can stand adjacent, although with varying Probabilities of Transition. If these were all included and a tree diagram were made, an enormous tree diagram could be drawn with T the indicator of the beginning of the sentence, or the Full Stop as the Initial Condition. All of this cannot be shown here, but when a part of the tree diagram from the beginning of the sentence for both the works taken is shown, the result is Figure 10.

When the tree diagram models obtained from 'Ki no Saki nite' (Figure 10.1) and 'Sutoo Huzin' (Figure 10.2) are compared, extremely similar sentence structures are also found here for Sentence Initial Phrases which are the same. Consequently, it may be inferred that once the Sentence Initial Phrase has been decided there is not much

Figure 10
Part of the Tree Diagram from the beginning of the Sentence

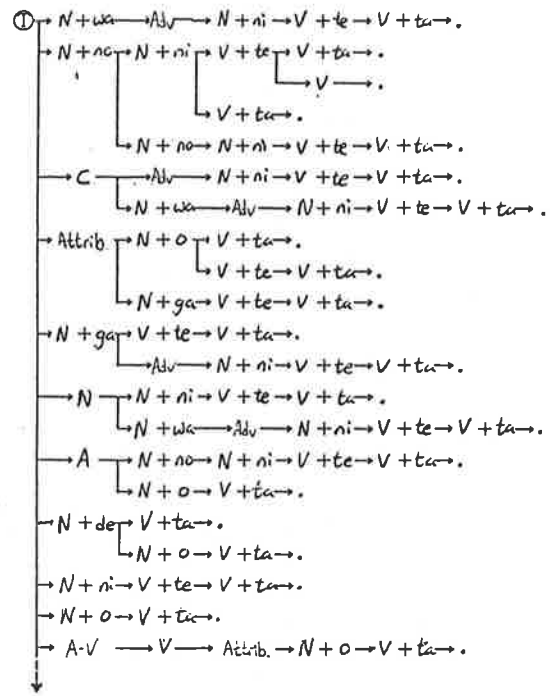


Figure 10.1. 'Ki no Saki nite'

difference according to the passage in question in the type of sentence structure which this introduces.

Next, if a part of the tree diagram from the Full Stop is shown, Figure 11 is obtained.

Of course, when the end of the sentence is the starting point, the same Sentence Final Phrase is hardly ever obtained because of the strong influence of the differences between the Polite Style and the Plain Style. Hence, quite a difference in the types of sentence structures may be seen between the two works. It may be said that the difference

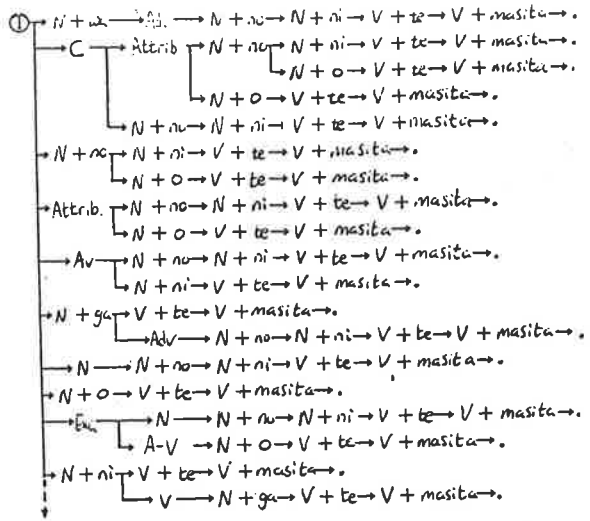


Figure 10.2. 'Sutoo Huzin'

Figure 11
Part of the Tree Diagram from the End of the Sentence

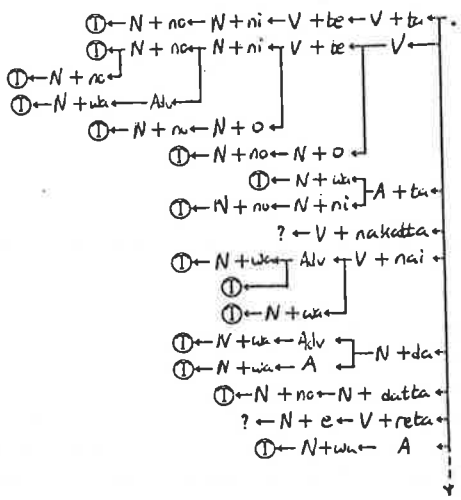


Figure 11.1. 'Ki no Saki nite'
? indicates that there is a competition of phrases with the same Probability of Transition

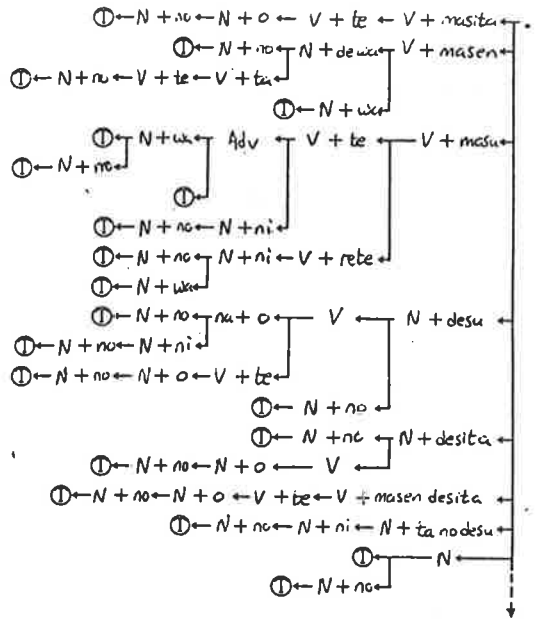


Figure 11.2. 'Sutoo Huzin'

in style of the two works is reflected in this point. Whatever the case may be, in the model from the end of the sentence, to what extent these models may be viewed as general cannot be discussed as common points cannot be grasped because of the differences between the Polite Style and the Plain Style. However, when the fact that extremely common-sense sen-tence structure types have been obtained is considered, at least it may be thought that these sentence structure types are quite likely to be formed when the Sentence Final Phrase is decided.

As we stated at the beginning, the sentence structure types which we have attempted to draw up should be called 'Approximate Sentence Structures of Phrase Units' and are 'Approximate Sentence Structures of Secondary Approximation', if the concept used in Computational Linguistics is borrowed. Consequently, if actual vocabulary which often comes up in the texts is inserted where there are the names of word classes in these sentence structures, typically 'Ki no Saki nite' sentences or sentences likely to occur in 'Sutoo Huzin' should be produced. If actual vocabulary is applied to the sentence structures mentioned in Tables 10.1 and 11.1 using the results of the vocabulary survey on 'Ki no Saki

nite'2, the following sorts of sentences are obtained (the figures are the order of frequency depending on the word class in the vocabulary survey):

T——>(N+wa)——>(Av)——>(N+ni)——>(V+te)——>(V+ta)——>
Zibun wa soo imori ni site ita.
<1st> <1> <3> <1> <2>

T——>(C)——>(Adv)——>(N+ni)——>(V+te)——>(V+ta)——>.
Sihasi moo isi ni natte simatta.
<1> <2> <7> <3> <12>

T<——(N+no)<——(n+ni)<——(V+te)<——(V)<——.
Nezumi no mue ni itte miru.
<5> <7> <8> <7>

T<——(N+wa)<——(Adv)<——(V+nai)<——.
Hati wa mattaku ugakanai.
<4> <4> <13>

These show a way of applying semantic elements according to probability after the sentence structure has been synthesised as one method of generating sentences. It is thought that it may be easier to obtain natural sentences with this, rather than with the method of creating sentences by directly taking the Probability of Adjacency of words and syllables. However, there are limits with the Probability of Transition for two phrases being adjacent, which we have atempted, if one is seeking a variety of sentences. One should be able to make more natural and varied sentences if one procedes by calculating the entropy of three phrases being adjacent using a computer and sentence structures of tertiary approximation are made.

Moreover, it is thought that grasping the various forms of Phrase Entropy using a computer for such things is important in the sense of obtaining reliable materials for syntactic theory.

2 See A. Tanaka, 'Zidoo Syooroku Syori ni okeru Kii-Waado no Seikaku', in 'Densi Keisanki ni yoru Kikugo Kenkyuu <V>', 'Report 49 of the National Language Research Institute', p.173 for the results of the vocabulary survey of 'Ki no Saki nite'.

6. Probability of Phrase Combination Forms

In this section we will consider several 'Phrase Combination Forms' which are often taken up in syntax using the Phrase Entropy in the two works which we have been treating.

First the results of seeing whether combinations of the pattern 'Zoo wa hana ga nagai', mentioned in Section 3 or combinations of the pattern 'Zoo no hana wa nagai' are more likely to be formed, are as in Table 12. Those marked 'Adjective-Type' in this table include phrases of patterns which begin with adjectives. Looking at Table 12, as far as one may infer from our data, sentence structures of the structure 'Zoo wa hana ga nagai'-type are far more likely to be formes than the 'Zoo wa hana ga nagai'-type.

If a similar method is used to examine which of the orders 'ni Case - o Case - Verb' and 'o Case - ni Case - Verb' mentioned in Section 4 as more likely to be formed, Table 13 is obtained.

'Verb-Type' is used in this table to include phrases of types beginning with a verb. As we understand from Table 13, in both cases, the Probability of Transition is always higher for 'ni Case - o Case - Verb' than for 'o Case - ni Case - Verb'. Consequently, this means that, within the scope of the data which we have obtained, the 'ni Case - o Case - Verb'-type is more likely to be formed.

Table 14 concerns the relation of sentence structures of the pattern 'mizu ga nomitai' and the pattern 'mizu of nomitai', which are often taken up in syntax. What has beeen marked 'Verb + tai'-type in this table includes phrases of patterns with the auxiliary verb 'tai' attached to the verb, such as 'nomitai', 'nomitaku' and 'nomitakatta'. In the results of Table 14, it can be seen that the 'mizu o nomitai'-type pattern is more likely to be formed than the 'mizu ga nomitai'-type, at least from the data in 'Sutoo Huzin' (in 'Ki no Saki nite', there was only one example of the combination '(Noun + o) - (Verb + takatta)',so no result could be obtained on this point.

Next, if the Probability of Transition of sentence structures of the 'tenki ga ii'-pattern and the 'tenki no ii'-pattern is calculated, based on data from 'Ki no Saki nite', Table 15 is obtained.

Table 12
The patterns 'Zoo wa hana ga nagai' and 'Zoo no hana wa nagai'

Data	(Preceeding Phrase)——→(Succeeding Phrase)	Probab.
Ki no Saki nite	(N+wa) —0.00164→ (N+ga) —0.00357→ (Adjective-type)	0.00059
	(N+no) —0.00750→ (N+wa) —0.00738→ (Adjective-type)	0.00553
Sutoo	(N+wa) —0.00200→ (N+ga) —0.00366→ (Adjective-type)	0.00073
Huzin	(N+wa) —0.00913→ (N+wa) —0.00437→ (Adjective-type)	0.00399

Data	(Preceeding Phrase)←——(Succeeding Phrase)	Probab.
Ki no Saki nite	(N+wa) ←0.00238— (N+ga) ←0.00411— (Adjective-type)	0.00098
	(N+no) ←0.00783— (N+wa) ←0.01233— (Adjective-type)	0.00910
Sutoo	(N+wa) ←0.00268— (N+ga) ←0.00685— (Adjective-type)	0.00184
Huzin	(N+wa) ←0.01494— (N+wa) ←0.01096— (Adjective-type)	0.01637

Table 13
The Probability of Transition of 'ni Case' and 'o Case' before the verb.

Tab 13

Data	(Preceeding Phrase)——→(Succeeding Phrase)	Probab.
Ki no Saki nite	(N+ni) —0.00459→ (N+o) —0.07609→ (Verb-type)	0.03490
	(N+o) —0.00217→ (N+ni) —0.04862→ (Verb-type)	0.01057
Sutoo	(N+ni) —0.00815→ (N+o) —0.08299→ (Verb-type)	0.06765
Huzin	(N+o) —0.00454→ (N+ni) —0.06881→ (Verb-type)	0.03123

Data	(Preceeding Phrase)←——(Succeeding Phrase)	Probab.
Ki no Saki nite	(N+ni) ←0.00543— (N+o) ←0.01304— (Verb-type)	0.00701
	(N+o) ←0.00183— (N+ni) ←0.01233— (Verb-type)	0.00181
Sutoo	(N+ni) ←0.00851— (N+o) ←0.01710— (Verb-type)	0.00184
Huzin	(N+o) ←0.00435— (N+ni) ←0.01480— (Verb-type)	0.00644

Table 14
The Patterns 'Mizu ga nomitai' and 'Mizu o nomitai' (Sutoo Huzin)

(Preceeding Phrase) ----- (Succeeding Phrase)	Probability
(N + ga) ——→ (V + Tai-type)	0.0024
(N + o) ——→ (V + Tai-type)	0.0057
(N + ga) ←—— (V + Tai-type)	0.0833
(N + o) ←—— (V + Tai-type)	0.2500

Table 15
The Patterns 'tenki ga ii hi' and 'tenki no ii hi' (Ki no Saki nite)

(Preceeding Phrase) ----- (Succeeding Phrase)	Trans.prob.
(N + ga) —0.00357→ (A) —0.05283→ (Noun-type)	0.01887
(N + no) —0.00167→ (A) —0.05283→ (Noun-type)	0.00881
(N + ga) ←0.00566— (A) ←0.00365— (Noun-type)	0.00201
(N + no) ←0.00377— (A) ←0.00365— (Noun-type)	0.00138

Moreover, if the Probability of Transition for the types 'ame ga huru hi' and 'ame no huru hi' is obtained from data in 'Sutoo Huzin', Table 16 is obtained.

Table 16
The patterns 'ame ga huru hi' and 'ame no huru hi' (Sutoo Huzin)

(Preceeding Phrase) ----- (Succeeding Phrase)	Probability
(N + ga) —0.00171→ (V) —0.09079→ (Noun-type)	0.01550
(N + no) —0.00200→ (V) —0.09079→ (Noun-type)	0.01820
(N + ga) ←0.00179— (V) ←0.00824— (Noun-type)	0.00148
(N + no) ←0.00460— (V) ←0.00824— (Noun-type)	0.00379

When we view these two tables, the result is that of the patterns 'tenki ga ii hi' and 'tenki no ii hi', the ga Case is most likely to occur and of the patterns 'ame ga huru hi' and 'ame no huru hi', no Case is slightly more likely to occur. Thus, in the sentence types 'noun + ga/no ~ (AP/VP) - (Noun-Type)', ga Case is likely when an adjective phrase is chosen and no Case is likely when a verb phrase is chosen. Furthermore, 'Noun-Types' in Tables 15 and 16 includes phrases beginning with a noun.

As the above results are based on data obtained from the two short stories we took up, of course general trends in Japanese sentence structure cannot be inferred using them. It is sufficient if the advantages of considering problems in sentence structure using the Probability of Transition are understood. It must be said that, when it is borne in mind that there is a sense in which a sentence is formed through a process of probability, there are occasions when it is quite dangerous to analyse mathematical trends in sentence structure relying just on the frquency count. As has been clear from various examples we have raised, the possibility of one phrase joining with another is quite different depending on whether the Preceding Phrase is made the Initial Condition and the direction is from the beginning of the sentence to the end or the Succeeding Phrase is made the Initial Condition and the direction is from the end of the sentence to the beginning. This means that, when a certain phenomenon in sentence structure occurs depending on the combination and relationship of one phrase with another, its weight as seen from the Preceding Phrase and that as seen from the

Table 17
Case where the Probability of Transition is Reversed

(Preceeding Phrase) ----- (Succeeding Phrase)	Probability
(N + ga) —0.00366→ (A) —0.06804→ (Noun-type)	0.02489
(N + no) —0.00301→ (A) —0.06804→ (Noun-type)	0.02046
(N + ga) ←0.00685— (A) ←0.00346— (Noun-type)	0.00237
(N + no) ←0.01233— (A) ←0.00346— (Noun-type)	0.00426

Succeeding Phrase are not necessarily the same. In fact, it has been shown that it is possible in some cases for the tendency obtained from the Preceding Phrase to be opposite to that obtained from the Succeeding Phrase.

As an illustration of this, is is a fact that U is likely to follow Q in English. However, O is likely to precede U, certainly no Q. In the case of problems in sentence structure, for example, a Substantive Phrase follows an Attributive Phrase with a probability of 100%, but 'Noun + No'-type phrases are most likely to occur before a Substantive Phrase, certainly not Attributive Phrases, as can be understood from Table 3.

In this survey, when the 'tenki ga ii / tenki no ii hi'-type sentence structures shown in Table 15 based on data from 'Ki no Saki nite', were examined in the case of 'Sutoo Huzin', Table 17 was obtained.

Thus, the result in this case is that when working from the Preceding Phrase, 'ga Case' is more likely to occur and when working from the Succeeding Phrase, 'no Case' is more likely. Consequently, at least within the range of the 'Sutoo Huzin' data, whether 'no Case' or 'ni Case' is likely to be selected in this type of sentence structure is a subtle point difficult to decide. This fact would surely not have been captured if this had been analysed with a mere frequency count.

It is more reliable to introduce the view point of process of probability rather than analysing with a mere frequency count when the connections and relationships of one phrase with another are being required or order and agreement etc. in phrases are being interpreted quantitatively in sentence structure. Moreover, it is thought that there has been a tendency in previous research in syntax and word order to procede with analyses using adjusted examples and deductive methods not based on actual utterance and sentences. It is necessary to consider methods of probability in analyses as weil as extracting the object of the research out of actual sentences.

Addendum

I would like to express my gratitude to Christopher Dillon of the School of Oriental and African Studies, the University of London, for translating this article.

Hřebíček, L.(ed.), Glottometrika 11, 1989

The entropy of phoneme frequencies. German and French

Ursula Rothe, Peter Zörnig, Bochum

1. This contribution contains additional calculations of the entropy of phoneme frequencies, the model of which has been set up and tested by Zörnig/Altmann (1984).

The total number of phonemes in a language has influence on the frequencies of the individual phonemes: the more phonemes a language has, the more degrees of freedom the speaker has to prefer one phoneme, the lower the repeat-rate (cf. Zörnig/Altmann (1983)) will be in general.

Thus the informational content of a phoneme is higher in languages with more phonemes than in those with fewer phonemes. The entropy of the phoneme frequencies is an information measure: if the entropy is low one can conclude that the informational content of the phoneme system is high and if the entropy high one can conclude that the informational content of the system is low.

The calculations are based on data from Hug (1979). In Hug, the ranked phoneme frequencies out of several texts in German[1] and French[2] have been detetermined. Then several theoretical distributions (i.e. binomial as basic model, Poisson and normal distribution additionally) were fitted to the data. Other theoretical phoneme frequency distributions have been considered by Altmann/Lehfeldt (1980), Orlov/Boroda/Nadarejsvili (1982) and Guiter/Arapov (1982).

Our aim is to show whether the observed entropies of HUG's examinations lie near the entropy curve of Zörnig/Altmann, to compare them with the data of earlier examinations in Zörnig/Altmann (1983 and 1984) and to look whether the new calculations confirm the adequacy of the model.

2. The collection of the ***French*** data in HUG is performed under several aspects.

The texts were first subdivided into "blocks" of 100 phonemes each. In each of the blocks the phoneme frequencies were counted, once separately for vowels and for consonants respectively and once mixed (global

counts). The separate counts were performed on 10.000 phonemes, the global counts on 5.000 phonemes. The "blocks" were used to compare the amounts of vowel and consonant in a parallel manner for special purposes, which are irrelevant for us. Here we only consider the lists of the global counts (vowels and consonants mixed) containing 5000 phonemes in total.

A second aspect is the phonetic norm used. Hug applies two different norms, namely the norm of Fouché (1936 and 1959) and the Dictionnaire du Français Contemporain (1971) for texts written in rhyme and the norm of Le Petit Robert (1973) for texts written in prose.

The first norm is less satisfactory for phonetician's purposes, but it allows a better treatment of final vowels (esp. / / in verse end position), of the difference between /e/ and /ɛ/ and of the "liaison" in archaic language.

Thus the French data are based on two different norms, but the phoneme number is constant in both (35 French phonemes).

In the analysis of the ***German*** data, Hug applies only one norm, namely that of Philipp (1970 and 1974). The texts are also divided in "blocks" like in French, the text of "Faust" is seperated into even and odd page numbers, but we neglected this separtion because it was irrelevant for our purposes.

3. The theoretical curve was computed according to the model of the entropy in Zörnig/Altmann (1984:42):

$$H_K = \mathrm{ld}\, e \ln \left[\sqrt{(B+K)(B+1)} \ln \frac{B+K}{B+1} \right] . \qquad (1)$$

K is the number of phonemes in the inventory, B is a parameter that has been estimated in Zörnig/Altmann. By minimization of different sums of squared deviations the values B = 0.27 and B = 0.61 were found which yield good fittings for the observed data.

The observed entropy is obtained by the formula:

$$H_K = \sum_{j=1}^{K} p_j \,\mathrm{ld}\, p_j \qquad (2)$$

where p_j is the relative frequency of the j-th phoneme in a language with K phonemes.

The results are presented in table 1 for B = 0.61 and in table 2 for B = 0.27. The counts of Zörnig/Altmann are recapitulated in the table, the boldfaced parts (line numbers 45 to 49 and 58 to 60) represent the new results obtained by Hug's examinations.

The "L" in the left columns indicates that the samples are drawn from dictionaries, no additional comment indicates that the material comes from texts.

Figure 1 and 2 show the run of the theortical entropy curve (for B = 0.61 and B = 0.27 respectively), the numbers 1-8 represent the observed values according to Hug's material. The numbers 1-8 refer to the keys of the texts used by Hug (key see annex).

5. As the new calculations show, the prediction of the theoretical entropy (H_K) by means of K and the formula (1) yields a good approximation to the observed entropies of Hug's data. The observed entropies for Hug's data obtained from French texts with K = 35 lie within the values of the former examinations of Zörnig/Altmann (1984) for K = 35. The data obtained from German texts with K = 40 lie within the values of the former examinations for K = 39 and K = 41.

It would be interesting to calculate the variance of the deviations of the observed entropies from the theoretical entropies and to set up a confidence interval for these deviations.

Notes

The texts examined by Hug are referred to by use of the following key:

1) = HUG1 = *Psyché*, Molière
2) = HUG2 = *Psyché*, Corneille
3) = HUG3 = *Le Crime de Sylvestre Bonnard*, Anatole France
4) = HUG4 = *Propos*, Alain
5) = HUG5 = *Le Canard enchainé*
6) = HUG6 = *Die Memoiren des Peterhans von Binningen*, Goetz
7) = HUG7 = *Nicht nur zur Weihnachtszeit*, Böll
8) = HUG8 = *Doktor Faustus*, Mann

Table 1. Observed and computed entropy of 71 samples from different languages (B = 0.61)

No.	Language	K	H_K	H_K
1	Hawaiian	13	3.37080	3.26797
2	Hawaiian L	13	3.23851	3.26797
3	Samoan	15	3.47585	3.45306
4	Hawaiian	18	3.56711	3.68292
5	Philipino	21	3.82032	3.87260
6	Philipino	21	3.72508	3.87260
7	Kaiwa	21	3.94776	3.87260
8	Sea-Dayak L	21	3.87869	3.87260
9	Estonian L	23	4.08615	3.98271
10	Swahili L	24	4.02396	4.03379
11	French L	24	3.96340	4.03379
12	Albanian L	25	4.21763	4.08253
13	Indonesian L	25	3.92894	4.08253
14	Chamorro	25	4.21472	4.08253
15	Dutch L	26	4.08584	4.12912
16	English L	26	4.21509	4.12912
17	Rumanian	27	4.25396	4.17376
18	Spanish L	27	4.02392	4.17376
19	Haussa L	27	3.92267	4.17376
20	Dutch L	28	4.04885	4.21658
21	Serbocroatian L	29	4.28710	4.25773
22	Bulgarian L	29	4.21932	4.25773
23	German L	29	4.17274	4.25773
24	Indonesian	29	4.09424	4.25773
25	Indonesian L	29	4.34865	4.25773
26	German L	30	4.14752	4.29732
27	Gujarati	30	4.49780	4.29732
28	Italian L	30	4.00675	4.29732
29	Italian	31	4.25122	4.33548
30	Ukrainian L	31	4.56502	4.33548
31	Russian L	31	4.43360	4.33548
32	American English	32	4.48737	4.37229
33	Hungarian	32	4.50842	4.37229
34	Hungarian L	32	4.54499	4.37229
35	Khasi	32	4.46886	4.37229
36	Latvian L	32	4.42823	4.37229
37	Russian L	32	4.45324	4.37229
38	German	33	4.44353	4.40786
39	Georgian	33	4.29313	4.40786
40	Georgian	33	4.31065	4.40786
41	Ostyak	33	4.37579	4.40786
42	Ostyak	33	4.39514	4.40786
43	Ostyak	34	4.40649	4.44224
44	Ostyak	34	4.39094	4.44224

Table 1. Continuation

No.	Language	K	H_K	H_K
45	**French (HUG1)**	**35**	**4.64775**	**4.47553**
46	**French (HUG2)**	**35**	**4.64845**	**4.47553**
47	**French (HUG3)**	**35**	**4.67931**	**4.47553**
48	**French (HUG4)**	**35**	**4.70838**	**4.47553**
49	**French (HUG5)**	**35**	**4.70994**	**4.47553**
50	Czech	35	4.70060	4.47553
51	Czech L	35	4.72275	4.47553
52	French	35	4.10179	4.47553
53	Marathi	38	4.51440	4.56942
54	Bengali	38	4.86326	4.56942
55	Hungarian	39	4.60281	4.59891
56	English	39	4.70980	4.59891
57	Armenian L	39	4.43397	4.59891
58	**German (HUG6)**	**40**	**4.75584**	**4.62757**
59	**German (HUG7)**	**40**	**4.75676**	**4.62757**
60	**German (HUG8)**	**40**	**4.76735**	**4.62757**
61	Russian	41	4.82569	4.65548
62	Polish	42	4.72524	4.68263
63	English	42	4.91803	4.68263
64	Gujarati L	43	4.66462	4.70909
65	English	44	4.90642	4.73489
66	Slovak	44	4.73183	4.73489
67	Swedish	45	4.84058	4.76005
68	Ukrainian	46	4.62795	4.78460
69	Hindi	52	4.58460	4.92067
70	Burmese L	68	5.23064	5.21339
71	Vietnamese	74	5.16055	5.30433

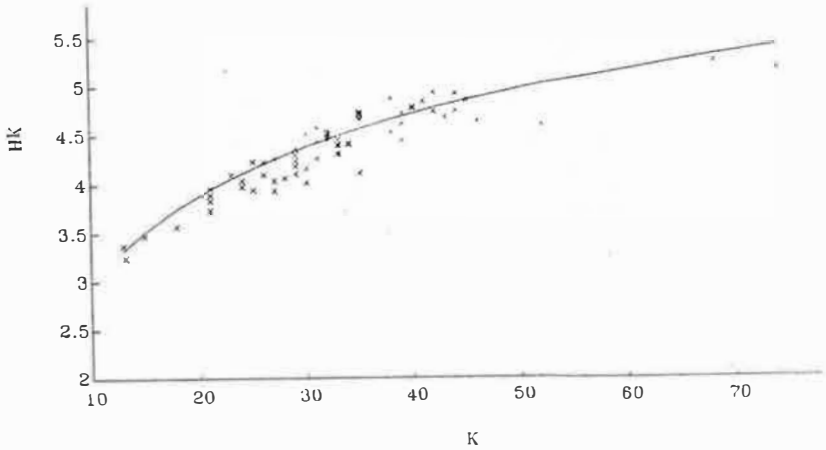


Figure 1. Entropy curve for 71 language data and observed entropies (cf. numbers) of Hug's material (B = 0.61)

Table 2. Observed and computed entropy of 71 samples from different languages (B = 0.27)

No.	language	K	H_K	H_K
1	Hawaiian	13	3.37080	3.32077
2	Hawaiian L	13	3.23851	3.32077
3	Samoan	15	3.47585	3.50949
4	Hawaiian	18	3.56711	3.74382
5	Pilipino	21	3.82032	3.93715
6	Pilipino	21	3.72508	3.93715
7	Kaiwa	21	3.94776	3.93715
8	Sea-Dayak L	21	3.87869	3.93715
9	Estonian L	23	4.08615	4.04935
10	Swahili L	24	4.02396	4.10139
11	French L	24	3.96340	4.10139
12	Albanian L	25	4.21763	4.15104
13	Indonesian L	25	3.92894	4.15104
14	Chamorro	25	4.21472	4.15104
15	Dutch L	26	4.08584	4.19851
16	English L	26	4.21509	4.19851
17	Rumanian	27	4.25396	4.24397
18	Spanish L	27	4.02392	4.24397
19	Haussa L	27	3.92267	4.24397
20	Dutch L	28	4.04885	4.28758
21	Serbocroatian L	29	4.28710	4.32949
22	Bulgarian L	29	4.21932	4.32949
23	German L	29	4.17274	4.32949
24	Indonesian	29	4.09424	4.32949
25	Indonesian L	29	4.34865	4.32949
26	German L	30	4.14752	4.36981
27	Gujarati	30	4.49780	4.36981
28	Italian L	30	4.00675	4.36981
29	Italian	31	4.25122	4.40866
30	Ukrainian L	31	4.56502	4.40866
31	Russian L	31	4.43360	4.40866
32	American English	32	4.48737	4.44614
33	Hungarian	32	4.50842	4.44614
34	Hungarian L	32	4.54499	4.44614
35	Khasi	32	4.46886	4.44614
36	Latvian L	32	4.42823	4.44614
37	Russian L	32	4.45324	4.44614
38	German	33	4.44353	4.48234
39	Georgian	33	4.29313	4.48234
40	Georgian	33	4.31065	4.48234
41	Ostyak	33	4.37579	4.48234
42	Ostyak	33	4.39514	4.48234
43	Ostyak	34	4.40649	4.51734
44	Ostyak	34	4.39094	4.51734

Table 2. Continuation

No.	language	K	H_K	H_k
45	**French (HUG1)**	**35**	**4.64775**	**4.55122**
46	**French (HUG2)**	**35**	**4.64845**	**4.55122**
47	**French (HUG3)**	**35**	**4.67931**	**4.55122**
48	**French (HUG4)**	**35**	**4.70838**	**4.55122**
49	**French (HUG5)**	**35**	**4.70994**	**4.55122**
50	Czech	35	4.70060	4.55122
51	Czech L	35	4.72275	4.55122
52	French	35	4.10179	4.55122
53	Marathi	38	4.51440	4.64677
54	Bengali	38	4.86326	4.64677
55	Hungarian	39	4.60281	4.67678
56	English	39	4.70980	4.67678
57	Armenian L	39	4.43397	4.67678
58	**German (HUG6)**	**40**	**4.75584**	**4.70565**
59	**German (HUG7)**	**40**	**4.75676**	**4.70565**
60	**German (HUG8)**	**40**	**4.76735**	**4.70565**
61	Russian	41	4.82569	4.73433
62	Polish	42	4.72524	4.76196
63	English.	42	4.91803	4.76196
64	Gujarati L	43	4.66462	4.78888
65	English	44	4.90642	4.81511
66	Slovak	44	4.73183	4.81511
67	Swedish	45	4.84058	4.84070
68	Ukrainian	46	4.62795	4.86567
69	Hindi	52	4.58460	5.00401
70	Burmese L	68	5.23064	5.30144
71	Vietnamese	74	5.16055	5.39380

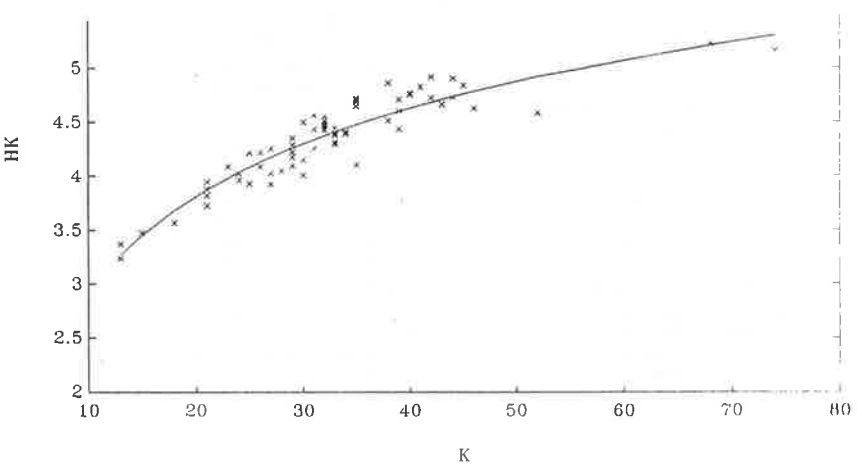


Figure 2. Entropy curve for 71 language data and observed entropies (cf. numbers) of Hug's material (B = 0.27)

References

Alain, E.Ch. (1967), Propos. In: Savin, M. (ed.), Bibliothèque de la Pléiade. Paris, Calmann-Lévy.

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

Bergmann, H. (1986), Einige Ergebnisse der Phonemstatistik. *Abhandlungen der Heidelberger Akademie der Wissenschaften, Philosophisch-historische Klasse* 5-19.

Böll, H. (1974), Nicht nur zur Weihnachtszeit. München, Deutscher Taschenbuch-Verlag.

Der Große Duden (1974),Bd. 6: Aussprachewörterbuch. Mannheim/Wien/ Zürich, Bibliographisches Institut.

Dictionnaire du Français Contemporain (1971), Paris, Larousse.

Fouché, P. (1936), Les diverses sortes de français du point de vue phonétique. *Français Moderne 4, 199-216.*

Fouché, P. (1959), Traité de prononciation française. Paris, Klincksiek.

France, A. (1967), Le Crime de Sylvestre Bonnard. Paris, Calmann-Lévy.

Goetz, C. (1974), Die Memoiren des Peterhans von Binningen. Frankfurt-Berlin, Ullstein-Bücher.

Guiter, H., Arapov, M.V. (eds.) (1982), Studies on Zipf's law. Bochum, Brockmeyer.

Hug, M. (1979), La distribution des phonèmes en français. Die Phonemverteilung im Deutschen. Genève, Slatkine.

Mann, Th. (1956), Doktor Faustus. Berlin-Darmstadt, Deutsche Buch-Gemeinschaft.

Molière, J.B., Corneille, P. (1962), Psyché. In: Oeuvres Complètes de Molière, par Robert Jouanny, tome II. Paris, Garnier.

Orlov, Yu.K., Boroda, M.G., Nadarejsvili, I.S. (1982), Sprache, Text, Kunst. Quantitative Analysen. Bochum, Brockmeyer.

Philipp, M. (1970), Phonologie de l'allemand. Paris, P.U.F.

Philipp, M. (1974), Phonologie des Deutschen. Stuttgart, Kohlhammer.

Zörnig, P., Altmann, G. (1983), The Repeat Rate of Phoneme Frequencies and the Zipf-Mandelbrot Law. *Glottometrika 5, 205-213.*

Zörnig, P., Altmann, G. (1984), The entropy of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 6, 41-47.*

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.* 194 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. 120 pp., pb ca. DM 34.80, 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. 350 pp., pb ca. DM 69.80, ISBN 3-88339-411-4
Vol. 10: KOCH, Walter A. (ed.), *Semiotics in the Individual Sciences.* Ca. 600 pp., pb ca. DM 94.80, 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.* Ca. 300 pp., pb ca. DM 69.80, 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.), *Reconstructing Languages and Cultures. Abstracts and Materials from the First International Interdisciplinary Symposium on Language and Prehistory Ann Arbor, 8.-12. Nov. 1988*. 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
***Vol. 24:** KOCH, Walter A., *The Wells of Tears. A Bio-Semiotic Essay on the Roots of Horror, Comic, and Pathos*. Ca. 100 pp., pb ca. 29.80, ISBN 3-88339-698-2 (3.89)

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.

March 1989

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: Monographien, 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*. 258 S., pb (paperback) DM 44.80, ISBN 3-88339-419-X (12/85)..
***Bd. 2:** ZHOU, Hengxiang, *Determination und Determinantien: Eine Untersuchung am Beispiel neuhochdeutscher Nominalsyntagmen*. 255 S., pb DM 49.80, ISBN 3-88339-412-2 (3/85).
***Bd. 3:** KOCH, Walter A., *Philosophie der Philologie und Semiotik*. 256 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. 210 S., pb ca. DM 39.80, ISBN 3-88339-415-7 (3.89)
***Bd. 5:** ESCHBACH, Achim (ed.), *Perspektiven des Verstehens*. 155 S., pb DM 34.80, ISBN 3-88339-414-9 (10/86).
***Bd. 6:** CANISIUS, Peter (ed.), *Perspektivität in Sprache und Text*. 242 S., pb DM 34.80, ISBN 3-88339-416-5 (7/87).
***Bd. 7:** EISMANN, Wolfgang, GRZYBEK, Peter (eds.), *Semiotische Studien zum Rätsel*. Ca. 293 S., pb ca. DM 69.80, ISBN 3-88339-417-3 (6/87).
Bd. 8: KOCH, Walter A. (ed.), *Semiotik in den Einzelwissenschaften*. Ca. 600 S., pb ca. DM 94.80, ISBN 3-88339-418-1
***Bd. 9:** SENNHOLZ, Klaus, *Grundzüge der Deixis*. 303 S., pb DM 64.80, ISBN 3-88339-462-9 (10/85).
***Bd. 10:** KOCH, Walter A., *Evolutionäre Kultursemiotik*. 321 S., illus., pb DM 64.80, ISBN 3-88339-463-7 (3/86).
***Bd. 11:** CANISIUS, Peter, *Monolog und Dialog*. 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*. 230 S., pb DM 39.80, ISBN 3-88339-487-4
***Bd. 13:** SCHMIDT, Ulrich, *Impersonalia, Diathesen und die deutsche Satzgliedstellung*. 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:** FLEISCHER, Michael, *Die Evolution der Literatur und Kultur*. Ca. 200 S., pb ca. DM 38.80, ISBN 3-88339-596-X (3.89)
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.), *Natürlichkeit der Sprache. Acta Colloquii*. Ca. 150 S., pb ca. DM 34.80, ISBN
Bd. 19: KOCH, Walter A. (ed.), *Natürlichkeit der Kultur. Acta Colloquii*. Ca. 150 S., pb ca. DM 34.80, ISBN
***Bd. 20:** KUGLER-KRUSE, Marianne, *Die Entwicklung visueller Zeichensysteme. Von der Geste zur Gebärdensprache*. 278 S., pb 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. 454 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.

März 1989