QUANTITATIVE LINGUISTICS

Volume 57

Peter Schmidt (ed.)

Glottometrika 15

Issues in General Linguistic Theory and The Theory of Word Length

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Glottometrika ... -
WVT Wissenschaftlicher Verlag Trier.
Früher mehrbd. begrenztes Werk. -
Bis 13 (1992) im Verlag Brockmeyer, Bochum
ISSN 0932-7991
Bd. 15. Issues in general linguistic theory and the theory of word length. - 1996

Issues in general linguistic theory and the theory of word length / Peter Schmidt (ed.). -
WVT Wissenschaftlicher Verlag Trier, 1996
(Glottometrika ... ; Bd. 15)
(Quantitative linguistics ; Vol. 57)
ISBN 3-88476-228-1
NE: Schmidt, Peter [Hrsg.]; 2. GT

Umschlag: Brigitta Disseldorf
(Marco Nottar, Agentur für Werbung und Design, Konz)

ISBN 3-88476-228-1
ISSN 0932-7991

WVT Wissenschaftlicher Verlag Trier
Bergstraße 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504

Contents

General

Word length distributions

Glottometrika 15, 1996, 1-21

Über eine Systemtheorie der Kultur
(Thesen zum Evolutionsproblem)

Michael Fleischer, Bochum

0. Problemstellung

Die Evolution ist ein universelles Prinzip und funktioniert in allen offenen Systemen. Was diese Systeme verbindet, ist die Auffassung, daß die Prinzipien der Evolution die Entwicklung von Systemen zu beschreiben erlauben. Seit einiger Zeit (vgl. Riedl, 1984, 1987; Wuketits, 1983; Fleischer, 1989a) spricht man auch in der Biologie (Ethologie) von der sozialen und kulturellen Evolution und vertritt die Meinung, daß soziale Systeme evoluierende offene Systeme darstellen.

Abstrahiert man nun von den Eigenarten sowohl des Untersuchungsgegenstandes als auch der angewandten Methoden jeder dieser Einzelwissenschaften, so ist doch bemerkenswert, daß sich dermaßen verschiedene Gegenstände auf ein einziges Entwicklungsprinzip zurückführen lassen. Zwei wissenschaftliche Strömungen der letzten Jahrzehnte haben in dieser Hinsicht einen besonderen Verdienst aufzuweisen, wenn es darum geht, isolierte Einzelwissenschaften übergreifende Tendenzen bzw. Entwicklungsprinzipien aufzudecken und unter einem einheitlichen Gesichtspunkt zu betrachten. Es sind einerseits die Arbeiten von Bertalanffy (z.B. 1945, 1957, 1968) auf dem Gebiet der Biologie und - breiter gesehen - der Systemtheorie, der darum bemüht ist, allgemeine Gesetzmäßigkeiten und Funktionsweisen von offenen Systemen, unabhängig davon, wo sie nun im einzelnen auftreten, zu erklären. Andererseits ist die systemtheoretisch orientierte Evolutionstheorie zu nennen (die Arbeiten von Riedl, Weiss u.a.), in der eine systemtheoretische Perspektive mit der Evolutionsforschung verbunden und in Einklang gebracht wurde, was zu weitreichenden neuen Erkenntnissen führte und was die Betrachtungsperspektive der Phänomene erweiterte.

0.1. Überblick

Es ist bei der Beschäftigung mit diesem Themenkreis bzw. Problemfeld unumgänglich, eine interdisziplinäre Betrachtungsperspektive und Analysestrategie zu

wählen, d.h. Erkenntnisse und Forschungsergebnisse aus verschiedenen, mitunter sehr unterschiedlichen Wissenschaften heranzuziehen, um eine Basis für die Beschäftigung mit der Kultur und der Literatur zu schaffen, auch wenn sie - streng genommen - mit den Geisteswissenschaften (noch) nichts zu tun haben. Diese Disziplinen sind jedoch zu Erkenntnissen gekommen, die wir, wenn schon nicht mitgestalten können, so doch zumindest aufnehmen sollten.

Es handelt sich dabei im einzelnen um die Thermodynamik, und genauer gesagt, um die Thermodynamik irreversibler Prozesse (bzw. um die Nicht-Gleichgewichts-Thermodynamik), es geht um die biologische Evolutionstheorie und die Systemtheorie, wie auch um die Semiotik allgemein und schließlich um die Kultursemiotik im einzelnen. Diese Aufzählung, wie befremdlich und anmaßend sie sich auch anhören mag, nennt eigentlich eine logisch und auch tatsächlich zusammenhängende Kette von Einzelwissenschaften, die darum bemüht sind, *einen* Gegenstand, *ein* Phänomen erklären zu wollen, nämlich die Entwicklung von Systemen, die Gesetzmäßigkeiten, denen sie offensichtlich gemeinsam unterliegen müssen, wenn so verschiedene Systeme auf ein und demselben Prinzip zu basieren scheinen, auf dem Prinzip der Evolution.

Nun taucht selbstverständlich sofort die Frage auf: Was hat denn die Literatur damit zu tun? **Es wird von der Behauptung ausgegangen, daß die Literatur und - breiter - die Kultur evoluierende Systeme sind, solche Systeme also, die einer Evolution unterliegen**. Diese Behauptung wird hier nicht bewiesen werden können, es soll jedoch - und lediglich - versucht werden, sie plausibel zu machen (eine detaillierte und breit angelegte Untersuchung dieses Problemfeldes ist in Fleischer, 1989a in Angriff genommen worden).

0.2. **Das Problem**

Es wird also von der Annahme ausgegangen, daß, wenn die gesamte Entwicklung der Materie ein evolutionärer Prozeß ist, auch die Entwicklung der Kultur - und damit die der Literatur - dem gleichen Mechanismus unterliegt. Eine gewisse Naivität bzw. Trivialität (und daher auch Angreifbarkeit) der Aussagen ist auf dieser - allgemeinen - Analyseebene nicht zu vermeiden, sie muß in Kauf genommen werden.

Die Prinzipien und Faktoren der Evolution haben die gesamte vorhandene Ordnung hergestellt. In der Phylogenese der Organismen sind auch Zeichensysteme entstanden, oder vorsichtiger formuliert, es sind Zeichen entwickelt worden, die zur Informationsübertragung und also zur Kommunikation dienen (siehe dazu detailliert Fleischer, 1990). Es haben sich offene Systeme entwickelt. Es wird angenommen, daß die Kultur und die Literatur offene Systeme sind, und daß sie

evolutionären Prinzipien unterliegen. Man weiß, daß die Literatur ein Zeichenphänomen ist.

Zeichen und offene Systeme unterliegen in der organismischen Welt der Evolution. Tiere, Menschen, soziale Systeme, Kultur, Literatur und Zeichen sind Produkte der Evolution. Kultur, Literatur, menschliche Zeichen sind Produkte des evolutiv entstandenen Menschen. Daraus wird gefolgert, daß es zumindest möglich sein könnte, daß auch diese Produkte dem Mechanismus der Evolution unterliegen und ihre Entwicklung eine evolutionäre ist. Diese Folgerung basiert zunächst einmal lediglich auf der folgenden Überzeugung: Es ist wahrscheinlich, daß sich die Kultur nach demselben Mechanismus entwickelt wie auch ihre Produzenten, und es ist unwahrscheinlich, daß gerade für die Entstehung der Kultur andere Mechanismen erfunden werden sollten, als die der Evolution (was nicht heißen soll, daß es die einzigen sein müssen). Zumal, wenn man bedenkt, daß die Evolution der Kultur **gleichzeitig** mit der Evolution des Menschen verläuft. Es ist ja keine statische Situation vorhanden, wonach der Mensch als »höchste« (d.h. vorläufig letzte) Stufe der Evolution nun dieser nicht mehr unterliegt und nach neuen Gesetzen die Kultur erfunden habe. Was voraussetze, daß diese neuen Gesetze schon am Beginn der menschlichen Kultur - sozusagen a priori - gegeben waren bzw. vorhanden sein mußten.

Das soziale System evoluiert, und somit ist es wahrscheinlich, daß die sich mit ihm entwickelnde Kultur ebenfalls der Evolution unterliegt, zumal die Kultur kein isoliertes Phänomen, sondern ein Produkt der menschlichen Gesellschaften darstellt, und solange schon vorhanden ist wie der Mensch. Dafür sprechen auch kulturähnliche Phänomene (Protokulturen), die in tierischen Gesellschaften auftreten (siehe Burkhard, 1977; Burkhardt et al., 1972; Fleischer, 1987), und die somit als Vorläufer der menschlichen Kultur zu betrachten sind.

Für die Annahme spricht - vorläufig - lediglich die Intuition, daß die Entwicklung der Kultur den Prinzipien der Evolution unterliegt bzw. daß dies wahrscheinlich ist. Es ist nicht mehr. Diese Annahme gilt es nachzuprüfen. Es wird hier - wie betont - kein Nachweis erbracht. Daher soll nur im Gedankenexperiment ein Weg gezeichnet werden, ohne daß die einzelnen Schritte und Behauptungen auch bewiesen werden können. Eine für ein Gedankenexperiment legitime Vorgehensweise.

Die Fragen, die sich darüber hinaus stellen, sind die folgenden: Ist die Evolution der Kultur mit der biologischen Evolution identisch, oder besitzt die Kultur auch besondere von der organismischen Evolution abweichende Merkmale? Es könnte ja sein, daß nur einige Mechanismen genutzt werden und andere - dann kulturspezifische - speziell für die Kultur erfunden wurden. Oder aber sind beide Fälle identisch, und die Kultur weist nur Speziallösungen auf, folgt aber generell dem gleichen Mechanismus? Die interessanteste Frage lautet aber: Wie sieht die

Entwicklung der Kultur im einzelnen aus, wenn man davon ausgeht, daß sie der Evolution unterliegt? Wie könnte die Evolution der Kultur aussehen, wenn man die Ausgangsannahme einmal akzeptiert?

1. Semiotik und Kommunikation

Die vorgeschlagene Betrachtungsperspektive geht auf die folgenden fast schon trivialen Feststellungen zurück: Kunstwerke sind etwas Hergestelltes, sie sind aus Zeichen bestehende Konstrukte, die semiotischen Regeln bzw. Gesetzmäßigkeiten unterliegen; sie sind somit miteinander vergleichbar. Zeichen entstehen in sozialen Systemen und werden zu kommunikativen Zwecken hergestellt und angewandt. Die in einem Sozium vorhandenen Zeichensysteme ergeben die Kultur, deren Materialisation die (in verschiedener Form) gespeicherten Nachrichten darstellen. Künstlerische Nachrichten sind nicht isoliert vorhanden, sondern resultieren aus Operationen innerhalb der Funktionsmechanismen des sozialen Systems und sind auf dieses System zurückführbar. Sie entstehen und funktionieren in sozialen Systemen und sind selber Systeme. Kunstwerke besitzen auf verschiedenen Ebenen Funktionen diverser Art.

1.1. Der Mensch und andere gesellschaftlich lebende Tiere stellen zu kommunikativen Zwecken Zeichen her. Alle Kunstwerke und alle Kulturwerke bestehen aus den gleichen triadischen, offenen, dynamischen Zeichen (siehe Fleischer, 1987, 1988). Damit ist gemeint, daß alle Nachrichten aus gleich generierten - eben triadischen - Zeichen bestehen, die ein Mittel, ein Objekt und einen Interpretanten aufweisen. Das ist die gemeinsame Grundlage. Nun ist aber feststellbar, daß aus diesen Zeichen verschiedene Nachrichten entstehen. Dies ist dadurch bedingt, daß Zeichen offene Gebilde sind und auf seiten des Interpretanten unterschiedliche Bedeutungen generieren können. Die Verschiedenheit der auf gleichen Zeichen basierenden Nachrichten beruht also darauf, daß die konkret angewandten Zeichen unterschiedliche Interpretanten und also Bedeutungen bekommen (zum Problem des Interpretanten und der Bedeutung siehe detailliert Fleischer, 1990:95-101).

Darüber hinaus unterscheiden sich Nachrichten untereinander auch dadurch, daß von Fall zu Fall verschiedene (nachrichtenspezifische) Verknüpfungsregeln der Zeichen und Verknüpfungsregeln nächsthöherer Einheiten (im Falle der natürlichen Sprache - Sätze u. dgl.) angewandt werden. Betrachtet man also die Kultur aus diesem Blickwinkel, so stellen sich alle Nachrichten als miteinander vergleichbar dar; sie weisen aber verschiedene Bedeutungen auf, abhängig davon, welche konkreten Funktionen sie wo erfüllen. Natürlich schwebt hier der Wunsch vor, in Analogie zur Biologie eine universelle DNA-äquivalente Struktur bzw.

solche Objekt- oder Systemeigenschaften zu finden, die es erlaubten, die Kultur als evolutionäres System zu betrachten. An dem Wunsch allein ist nichts auszusetzen. Ob dies aber so zu erklären ist?

1.2. Die sozial lebende Spezies Mensch baute (bedingt durch die Gesetzmäßigkeiten der biologischen Evolution) in phylogenetischer Hinsicht eine Zivilisation auf. Der Aufbau der Kultur basiert auf der **Zivilisation**, die hier - vereinfachend - folgendermaßen definiert wird: Die künstlichen Erzeugnisse des Menschen + Nutzung bestehender natürlicher Gegenstände *und* Gesetze.

Die Existenz der Kultur läßt sich bis auf die frühesten Phasen der phylogenetischen Entwicklung zurück nachweisen, und zwar universell in verschiedenen Populationen (siehe dazu Koch, 1986). Die Entstehung der Kultur ist ein allgemeines Phänomen aller sozialer (menschlicher) Verbände. Die Kultur wird hier als Äußerung zivilisatorischer Fähigkeiten betrachtet.

Es wird angenommen, daß die Kultur materialisiert ist, und daß eine nichtmaterialisierte Kultur nicht existiert. Die Materialisation der Kultur geschieht mit Hilfe von Zeichen. Die Herstellung von Zeichen ist keine ausschließlich menschliche Erfindung. Sie tritt in der biologischen Evolution von vornherein auf. Zeichen wurden zu kommunikativen Zwecken erfunden. Die DNA-Kette ist ein Zeichensystem, das mit informationsübertragender Funktion die Grundlage alles Lebendigen ist. Sozial lebende Tiere (darunter der Mensch) benutzen Zeichen zur Kommunikation. Während Tiere *ein* Zeichensystem benutzen, gebraucht der Mensch mehrere Zeichensysteme, dabei benutzt er Zeichen in (mindestens) zweierlei Funktion (der Funktionsbegriff wird hier nach Jachnow, 1981 verstanden):

(i) Es geht um die aktuelle Kommunikation innerhalb der Kultur. Alle Zeichen und alle Nachrichten entstehen zunächst mit dieser Funktion. Sie dienen zur Speicherung von Informationen, die zum gegebenen Zeitpunkt rezipiert werden.

(ii) Der Mensch hat darüber hinaus (nicht anders als alle anderen Organismen) Mechanismen entwickelt, abgespeicherte, d.h. in Zeichen und Zeichensystemen gespeicherte Informationen abzulagern und sie als informationellen, kommunikativen Bestand potentiell verfügbar zu halten. Es entsteht das Phänomen Kultur: Es existieren - d.h. sind gespeichert und abrufbar - Informationen und Bedeutungen, die über die jeweilige Aufnahme- und Speicherkapazität des Individuums hinausgehen, und in einem sozialen System, wenn auch nur potentiell, vorhanden und jederzeit nutzbar sind. Dadurch, daß ein Individuum nicht mehr imstande ist, die gesamte soziale Information aufzunehmen (was bei Tieren der Fall ist), bekommt die nun und so entstandene Kultur ihre **Eigendynamik**; die Inhalte und die Gesetzmäßigkeiten ihres Funktionierens werden vom Individuum unabhängig.

Dies ist nicht metaphorisch gemeint. Die Eigendynamik entsteht dadurch, daß jedes Individuum andere Bestandteile der Kultur - andere Zeichen also - aktualisiert und beliebige Zeichen aufnehmen oder aus den Speichern abrufen kann, so daß Mitglieder einer Kultur über verschiedene sich mehr oder weniger überlappende Informationen verfügen, die eine Eigendynamik der kulturellen Interaktion gewährleisten.

1.3. **Die Materialisation der Kultur** geschieht mit Hilfe von Zeichen. Konglomerate von Zeichen ergeben Zeichensysteme. Zeichensysteme müssen, falls es Systeme sind, Gesetzen der Systemtheorie unterliegen. Die Kultur wird durch gleich generierte Zeichen materialisiert; sie besteht jedoch aus unterschiedlich organisierten Informationen und Informationsträgern. Information in Form von Zeichen wird in Nachrichten erzeugt und als Nachricht gespeichert.

Die allgemeine Grundlage der Kultur ist ihr Zeichencharakter. **Kultur ist ein Zeichenphänomen. Kultur ist die Wirklichkeit der Zeichen.** Es wird also ein allgemeiner semiotischer Charakter der Kultur postuliert. Nicht in Zeichen materialisierte Bereiche der Kultur existieren nicht. Erst wenn eine Nachricht (in dieser oder jener - jedoch durch ein anderes Individuum wahrnehmbaren - Form) vorliegt, ist sie ein kulturelles Faktum. Die Integration, die Vernetzung dieser Nachrichten ergibt das Phänomen Kultur. Es wirkt hier die funktionelle Kausalität (siehe dazu Wuketits, 1985:77), d.h. die Vernetzung von Wirkung und Ursache.

Es werden immer mehr Zusammenhänge sichtbar, die im Hinblick auf das behandelte Problem auf das Vorhandensein von Systemeigenschaften (im Sinne der Systemtheorie) hinweisen. Man sollte daher fragen: Hat man es mit Systemen zu tun, und wenn ja, was für Systeme sind es?

Es besteht kaum Zweifel darüber, daß Zeichen - strenggenommen - Systemcharakter aufweisen; so wie auch kein Zweifel darüber besteht, daß die natürliche Sprache - als eines der Zeichensysteme - Systemcharakter besitzt. Darf man aber einen Schritt weitergehen und folgern, daß, wenn Zeichen Systeme sind, auch die Kultur, die ja auf Zeichenprozesse zurückzuführen ist, ein System ist? Es scheint, daß dies ein durchaus legitimer Schritt ist.

2. Das soziale System und die Kultur

Akzeptiert man das oben Gesagte, tauchen sofort neue Fragen auf: Was für ein System ist die Kultur, welche Grenzen hat sie, wie ist ihre innere Strukturierung? Es wurde bisher immer von *der* Kultur gesprochen, tatsächlich gibt es jedoch verschiedene Kulturen. Es scheint berechtigt zu sein, daß man von ***der*** **Kultur** als allgemeinem Mechanismus, als Suprasystem und daneben von verschiedenen und

unterschiedlichen Ausprägungen dieses Systems spricht. Darüber hinaus gibt es **Einzelkulturen** (Ebene I), unter denen die Nationalkulturen, so wie sie sich in der Geschichte ausgebildet haben, verstanden werden. Innerhalb dieser Nationalkulturen treten wiederum **Subkulturen** (Ebene II) diverser Art auf.

Es ist eine Tatsache, daß es zwischen den Subkulturen einer Einzelkultur, wie aber auch zwischen Einzelkulturen untereinander, ein breites Spektrum an Wechselwirkungen und Rückkopplungen gibt, also hier durchaus systemtheoretische Zusammenhänge zu vermuten sind. Das heißt aber auch, daß der hier benutzte Begriff 'die Kultur' einen Objektbereich charakterisiert. Denn wenn es zwischen den Einzelkulturen tatsächlich Wechselwirkungen gibt, dann kann man *die* Kultur als 0-te (höchstmögliche) Ebene definieren, so daß dieser Begriff die gesamte menschliche Kultur - nun in funktioneller Abgrenzung zum zivilisatorischen Bereich - bezeichnet. Kultur ist auf allen Ebenen jeweils als System zu charakterisieren (s. detailliert Fleischer, 1994).

2.1. Die **Abgrenzung zwischen Zivilisation und Kultur** ist relativ problematisch (wenn überhaupt nötig oder nützlich), sie kann auf keinen Fall mechanisch verstanden werden: etwa als zwei nebeneinander existierende Objektbereiche. Es ist eine funktionelle Differenzierung anzustreben. Man könnte vorläufig formulieren: Zivilisation unterscheidet sich von Kultur durch den Zeichengebrauch. Die Zivilisation ist der Wirklichkeitsbereich, in dem keine Zeichen benutzt werden, in dem keine Nachrichten produziert werden.

Es ist natürlich nicht an eine räumliche Differenzierung gedacht. Ein einfaches Beispiel: Die in einer Fabrikhalle stehende Maschine gehört zum Bereich Zivilisation; die gleiche Maschine, die auf einer Messe ausgestellt wird, besitzt Zeichencharakter und gehört somit zum Bereich der Kultur. Es geht also um eine funktionelle Differenzierung. Die Zivilisation stellt für die Kultur die Umwelt im Sinne der Systemdefinition dar, die besagt, daß eine Veränderung der Umwelt Einfluß auf den Zustand des Systems ausübt und umgekehrt, daß die Veränderungen der Systemeigenschaften den Zustand der Umwelt beeinflussen. Per analogiam ist *die* Kultur die Umwelt für jede Einzelkultur, und eine Einzelkultur ist die Umwelt für Subkulturen usw.

Dabei ist zu bemerken, daß hier kein einfaches (reduktionistisches oder holistisches) Stufenmodell gemeint ist, in dem jeder nächsthöhere Bereich die anderen umschließt, sondern vielmehr ein vernetztes Gebilde von Wechselwirkungen. Es kann also vorkommen, daß Veränderungen in der Zivilisation (also in der Umwelt der Kultur) direkten Einfluß auf eine bestimmte Subkultur haben, ohne daß sich dadurch die gegebene Einzelkultur verändert bzw. vorher verändert haben muß, und umgekehrt: Eine Subkultur kann auf die Zivilisation einwirken und sie verändern, diese Veränderung braucht sich erst später auf die Einzelkultur und noch

später, oder aber überhaupt nicht auf die Kultur auswirken. Die Ebenen dieses Gebildes werden (nach oben) immer komplexer; das heißt aber nicht, daß Wirkungen und Rückkopplungen nur schrittweise - kontinuierlich, Stufe für Stufe - möglich sind. Wechselwirkungen sind in allen Richtungen möglich, unabhängig vom Komplexitätsgrad des Systems. Die Komplexität des Gesamtsystems Kultur ist aber eine hierarchische und ansteigende. Diese Prinzipien schließen sich nicht aus. *Die* Kultur, Einzelkulturen und Subkulturen sind insofern homogene Bereiche, da sie auf die Benutzung von Zeichensystemen zurückgehen. Dagegen ist die Zivilisation ein Bereich der nichtzeichenhaften Phänomene (die jedoch, wie das obige Beispiel zeigt, grundsätzlich in Zeichen umfunktioniert werden können). Alle Ebenen der Kultur besitzen darüber hinaus eine Umwelt in Form der Zivilisation, die als höchstmögliche Umwelt des kulturellen Bereichs aufzufassen ist. Rückwirkungen finden auf allen Ebenen statt.

2.2. Es taucht nun die Frage auf: Wo ist in diesem Modell die soziale Komponente - die Gesellschaft - anzusiedeln? In dieser Frage kommt eine nicht adäquate Betrachtungsperspektive zum Vorschein. Das soziale System ist nämlich kein Glied in der Reihe Zivilisation → *die* Kultur → Einzelkultur usw., sondern deren **Grundlage**. Die menschliche Gesellschaft, die ebenfalls ein System ist und genau die gleiche Differenzierung wie die Kultur-Reihe aufweist, stellt die Grundlage, die Basis dieser Reihe dar. Das soziale System ist phylogenetisch, biologisch und soziologisch zu verstehen: der Mensch als Produkt der Evolution (siehe Humanethologie). Das soziale System produziert Kultur, es ist Träger und Produzent der Kultur.

Kultur (und Zivilisation) ist ein Objektbereich, wie das Verhalten der Tiere ein Objektbereich der tierischen Gesellschaften ist, und wie das Verhalten der Zelle der Objektbereich chemischer Prozesse in ihr ist. Das eine ist auf das andere nicht zurückführbar, und beide Bereiche sind aufeinander nicht reduzierbar, sie bedingen sich gegenseitig. Unter 'sozialen Systemen' werden also die menschlichen Gesellschaften (geographischer, nationaler, Stammes-Art usw.) verstanden; 'das soziale System' bezeichnet die Gesamtheit aller Gesellschaften. Auch dies sind offene Systeme. Die Organisationsformen der Gesellschaften besitzen Systemcharakter. Die Beschaffenheit sozialer Systeme bezieht sich direkt auf die biologische Evolution. Veränderungen und Entwicklungen in sozialen Systemen finden ihren Ausdruck in der Kultur (Stichwort: Meme). Insofern ist auch die Kultur ein Untersuchungsobjekt bei der Analyse des sozialen Systems.

Auch diese Relation ist nicht einseitig, es gibt Wechselwirkungen zwischen beiden Bereichen: Die Ausprägung des sozialen Systems äußert sich in der Kultur, die Entwicklungen (außer den biologischen) sind hier ablesbar (der biologische Zustand ist im sozialen System gegeben). Da aber die Kultur Eigendynamik

entwickelt und besitzt, wirkt auch sie auf die sozialen Entwicklungen zurück und kann sie teilweise bedingen. Eine Konsequenz daraus ist, daß die Kultur steuernde Funktionen ausüben kann; die Kultur als Bereich zeichenhafter Phänomene steuert die Entwicklung sozialer Systeme zumindest mit.

3. Kultur - Kunst - Literatur

Wie ist nun im Hinblick auf die hier vorgeschlagene systemtheoretische Fragestellung der Zusammenhang dieser drei Bereiche zu sehen? Es wurde gesagt: Kultur ist die Sphäre der zeichenhaften Tätigkeit des Menschen, sie umfaßt alle Phänomene, die mit Zeichen zu tun haben, und gliedert sich auf verschiedenen Ebenen in mehrere Subsysteme. Kunst und Literatur sind ebenfalls Gliederungsbereiche der Kultur, die jedoch mit den obigen nicht identisch sind. Die Kunst (darunter die Literatur) ist ein Bereich besonderen Zeichengebrauchs, sie benutzt zwar die gleichen Zeichen wie alle anderen Kulturbereiche, jedoch mit einer anderen Funktion. Es reicht, wenn man hier auf die ästhetische Funktion ganz allgemein hinweist. Die Kunst ist ein Phänomen, das auf jeder der oben unterschiedenen Ebenen (0, I, II) auftritt, sowohl also in Einzelkulturen als auch in Subkulturen usw. Das gleiche betrifft auch jede Kunstgattung (Literatur, Malerei, Film usf.).

3.1. Man hat es also mit einem äußerst komplexen Gebilde zu tun, das hierarchisch und integrativ auf vernetzter Kausalität basiert. Auf jeder Ebene der Kultur tritt das Phänomen 'Kunst' auf, das sich wiederum in verschiedene Kunstgattungen gliedert. Mit dem Anwachsen der Komplexität und dem Übergang zu nächsthöheren Ebenen wird auch das jeweilige Kunstphänomen komplexer, obwohl seine Gliederung die gleiche bleibt. Die Kunst an sich ist aber auf jeder Ebene etwas anderes, sie umfaßt jeweils mehr und mitunter andere, zusätzliche Elemente. Die Kunst einer bestimmten Subkultur (die wiederum in verschiedene Gattungen zerfällt) ist ein Bestandteil des nächsthöheren Bereichs - der Einzelkultur (einer Nationalkultur) -, der nun alle Subkulturen umfaßt; auf der nächsthöheren Ebene ist dann die gesamte Kunst anzusiedeln. Die literarischen Werke (beispielsweise) einer Subkultur stellen ein bestimmtes Subsystem dar, auf der Ebene der Einzelkultur hat man schon mit einem anderen Subsystem zu tun, das alle subkulturellen Subsysteme umfaßt; es sind also - so gesehen - auch mehr Texte vorhanden; usf. auf höheren Ebenen. Bei jedem Ebenen-Übergang ändert sich jedoch nicht nur die rein zahlenmäßige Größe des Subsystems, sondern ebenfalls sowohl die systeminternen als auch die kulturellen Funktionen und Bedeutungen.

Das heißt also, daß die Zusammensetzung des Phänomens 'Kunst' auf verschiedenen Ebenen verschieden ist und unterschiedliche Funktionen besitzt. Es sind aber immer Nachrichten, die aus (gleich generierten) Zeichen bestehen. Ihre ontologische Grundlage ist die gleiche, unterschiedlich ist die Komplexität des Subsystems, dessen Zusammensetzung wie auch seine Funktionen. Wir haben es mit einem hierarchischen, vernetzten Mechanismus zu tun, der sowohl im Bereich Kultur als auch im Bereich Kunst funktioniert. Ein bestimmtes System ist inhaltlich und funktionell mit dem jeweiligen Subsystem nicht identisch und nicht auf dieses reduzierbar. Jeder nächsthöhere Bereich besitzt eine eigene Existenzgrundlage und eigene Objekte, und er besitzt und erzeugt eigene Gesetzmäßigkeiten, funktioniert jedoch im Rahmen des für alle gleichen Prinzips zeichenhafter Organisation.

3.2. Was nun die einzelnen Elemente des Kunst-Systems betrifft - Literatur, Film usw. -, so funktioniert auch hier der gleiche Mechanismus. Auch die Literatur ist ein zeichenhaftes Phänomen, und auch sie weist die gleichen Gliederungen und Differenzierungen auf, wie sie für die Kunst beschrieben wurden. Auch die Literatur stellt auf verschiedenen Ebenen - sowohl inhaltlich als auch funktionell - ein immer komplexer werdendes Objekt dar, das sich auf einer gegebenen Kunstebene in den jeweiligen Kunstbereich als Obermenge und diese wiederum in die Kulturebene usf. einfügt. Keine Ebene und kein System ist isoliert. Es sind prinzipiell Wechselwirkungen zwischen allen (und über alle) Ebenen, zwischen Ebenen und deren Systemen, als auch zwischen Systemen untereinander, zwischen Systemen auf verschiedenen Ebenen, zwischen verschiedenen Systemen auf einer und auf verschiedenen Ebenen möglich. Es ist auch hier kein Stufenmodell impliziert, sondern eine funktionelle Kausalität. Verdeutlicht wird das oben Gesagte durch die Abbildung 1.

Der Bereich 'Kunst', der mit dem Index (0) bezeichnet wurde, ist mit jenen (1) und (2) nicht identisch, es sind verschiedene Subsysteme. Rein mechanistisch gesehen, ist Kunst $_0$ die Summe von Kunst'$_2$ und Kunst"$_2$ usf., wie auch die Summe von Kunst'$_1$ und Kunst"$_1$ usf. (das gleiche betrifft die Literatur). Der jeweils komplexere Bereich ist jedoch nicht die Summe der jeweils niedrigeren. Jede Ebene besitzt ihre eigenen Auswahl-, Differenzierungs-, Integrations-, und Ordnungs-Kriterien. Es kommt also immer wieder vor, daß bestimmte Elemente (welche, das regeln die genannten Kriterien) des jeweiligen Subsystems »verloren gehen« und in das Suprasystem nicht aufgenommen werden; es kommt auch vor, daß sie zwar aufgenommen, dann aber umfunktioniert werden. Das gleiche betrifft auch Relationen innerhalb der Ebenen. So kann es erklärt werden, daß ein und dasselbe Kunstwerk in verschiedenen (z.B.) National- oder Subkulturen unterschiedliche Bedeutungen oder Funktionen aufweist bzw. ausübt.

Man muß also feststellen, daß Kultur, Kunst und Literatur Phänomene sind, die miteinander im integrativen Zusammenhang stehen, ein vernetztes Allgemeinphänomen ergeben und in Abhängigkeit von der jeweiligen Ebene verschieden zusammengesetzte Systeme darstellen, die in objektbezogener und funktioneller Wechselwirkung zueinander stehen.

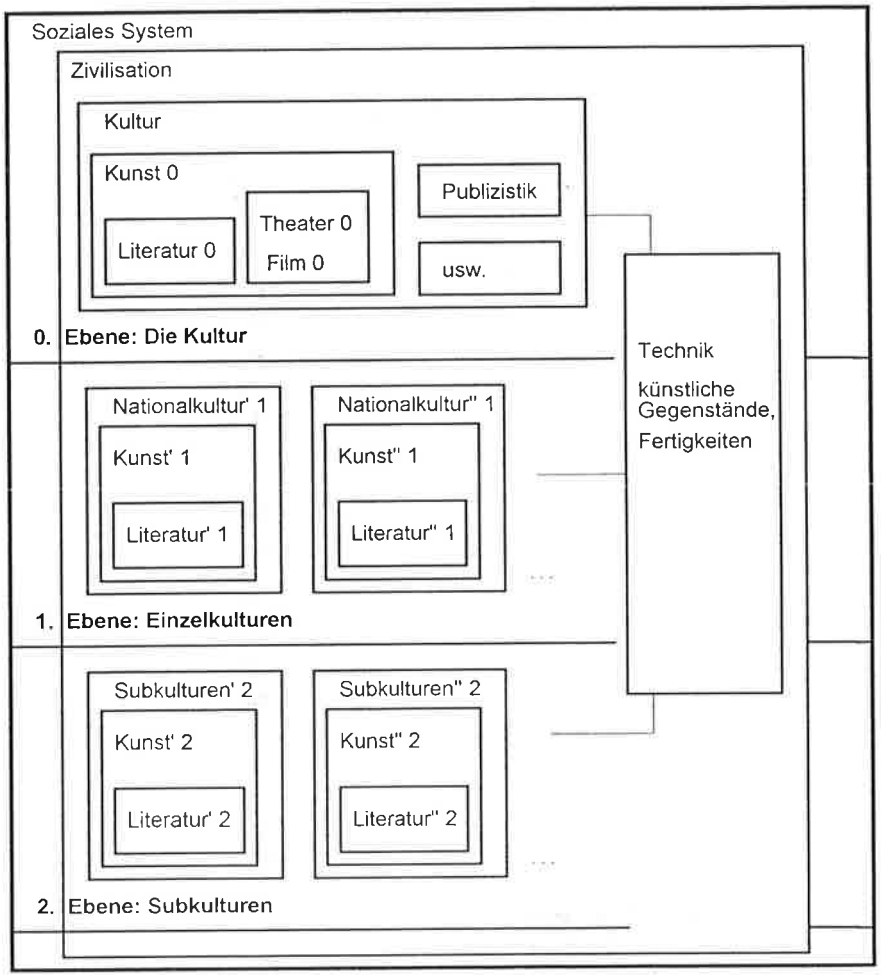


Abb. 1. Korrelation zwischen Kunst, Kultur, Zivilisation und sozialem System

3.3. Wie ist das Verhältnis dieser Bereiche zur Zivilisation und zum sozialen System? Jedes einzelne Phänomen (Literatur, Kunst, Kultur) bezieht sich seinerseits auf die Zivilisation, die auch die Umwelt dieser Systeme darstellt. Auf jeder Ebene ist die Zivilisation die Umwelt für die gegebene Kultur. Die Frage lautet: Ist sie das auch für die Kunst und für die Literatur, also für relativ untergeordnete Subsysteme? Bezieht sich also die Kunst auch auf den Bereich nichtzeichenhafter Phänomene?

Streng genommen »dürfte« sie dies nicht. Aus der Systemtheorie weiß man, daß das jeweils nächsthöhere System die Umwelt für das gegebene darstellt. Daher wäre zu fordern: Die Kunst sei die Umwelt der Literatur, die Kultur die Umwelt der Kunst usf. Man weiß aber, daß die Literatur sich auch auf die Zivilisation bezieht, dort entstandene Entwicklungen aufgreift u.dgl. Die Frage lautet: Geschieht das unmittelbar oder erst durch Vermittlung der Kultur bzw. der Kunst? Ein schwieriges Problem. Es ist eher anzunehmen, daß hier ein mittelbares Verhältnis auftritt. Die Literatur ist ein Zeichenphänomen, Zeichen besitzen Bedeutungen, Interpretanten, die sie zu Netzwerken verknüpfen. Das heißt, die Zivilisation weist als solche keine Bedeutungen auf, da sie ja kein zeichenhaftes Gebilde ist; sie kann also nicht direkt in Zeichensystemen auftreten. Erst wenn ein Bereich der Zivilisation durch diese oder jene Verfahren Zeichencharakter bekommt, d.h. in den Bereich der Kultur eingeht, kann er von der Kunst aufgenommen werden. Die Literatur kann (und sie tut es auch) Elemente der Zivilisation in ihren Texten aufnehmen. Die Literatur kann den Bereich der Zivilisation nicht von selbst semiotisieren, es bedarf der Vermittlung der Kultur. Und dies ist ein wichtiger Punkt.

Es scheint also, daß die Kunst und darunter die Literatur - als Systeme - nicht unmittelbar auf die Zivilisation als Umwelt zurückgreifen; für sie ist das nächsthöhere System - die Kultur also - ihre Umwelt. Erst wenn die Kultur neue Zeichen bzw. Zeichen-Verknüpfungsregeln einbringt bzw. zur Verfügung stellt, können sie von der Kunst in ihr System aufgenommen werden. Auch umgekehrt dürfte die Kunst nicht unmittelbar auf die Zivilisation wirken können.

Man sollte hier jedoch zwischen einfachen Einflüssen, kulturellen Einflüssen und Bedeutungen unterscheiden: Die Zivilisation wirkt dann auf die Kunst oder andere kulturelle Phänomene unmittelbar ein, wenn es sich um die materiellen Trägereigenschaften dieser Phänomene handelt. (Die Erfindung neuer Druckverfahren wirkt auf die Herstellung von Büchern. Und umgekehrt: Eine Flut von Autoren und Manuskripten kann einen Druck auf die Zivilisation ausüben, und es wird nach neuen, schnelleren Druckverfahren gesucht.) Dies sind die einfachen Einflüsse. Sobald wir von kulturellen Einflüssen der Zivilisation sprechen, haben wir es mit mittelbaren Beeinflussungen zu tun. Bei diesen bedarf es der Vermittlung von Zeichensystemen und der Vermittlung der Kultur, damit die Zivilisation

die Kunst oder andere Phänomene beeinflussen kann. Der zivilisatorische Gegenstand oder eine Fertigkeit allein können unmittelbar nicht wirken.

[Man nehme als Beispiel den Fall, in dem einem sog. Naturvolk Fotos von Mitgliedern der Gemeinschaft präsentiert werden (wir sehen die Fotos zunächst als Zivilisationsgegenstand an), die Betrachter können mit einem solchen Foto nichts anfangen, sie erkennen sich nicht darauf; eben weil es der Vermittlung des Zeichensystems, der Kultur bedarf, um Fotografien überhaupt zu verstehen. Ohne Kenntnis des Zeichensystems werden die Fotos nicht verstanden. Sie sind nur ein Zivilisationsgegenstand, können aber nicht wirken. Nun wird es aber interessant: Da der Mechanismus der Bedeutungsgenerierung, der kulturellen Umfunktionierung zivilisatorischer Gegenstände (wie in jeder Kultur) auch bei dem Naturvolk bekannt ist, kann es durchaus vorkommen, daß ein solcher Gegenstand mystifiziert wird und z.B. Kultcharakter bekommt. Der Mechanismus der Vermittlung ist bekannt, und er wird auch angewandt, indem versucht wird, den Gegenstand in die eigene Kultur zu integrieren. Kennt man das Zeichensystem oder die kulturelle Bedeutung nicht, substituiert man diese und fügt den neuen Gegenstand in die eigene Zeichenwelt ein, oder man ignoriert ihn.

Ein anderes Beispiel (bezogen auf eine Schallplatte): Man kann im Dokumentarfilm von Robert Flaherty "Nanook of the North" (aus dem Jahre 1922) beobachten, wie ein Eskimo zum ersten Mal eine Schallplatte zu sehen und zu hören bekommt und aus Unkenntnis des Zeichensystems mit diesem Gegenstand nichts anzufangen weiß und daher nach dem im Grammophon versteckten Menschen sucht. Im Falle von kulturellen Einflüssen bedarf es also einer Vermittlung.]

Das gleiche betrifft die Bedeutungen, auch sie entstehen im Hinblick auf zivilisatorische Gegenstände erst durch die Vermittlung von Zeichensystemen bzw. Kulturen. Die Zivilisation kann erst dann auf die Kultur wirken, wenn ihre Gegenstände oder Fertigkeiten Bedeutungen bekommen.

[Siehe dazu Duchamps ready-mades, die er von 1913 an »anfertigte«: die Gebrauchsgegenstände - z.B. das 'Pissoir' aus dem Jahre 1917 - sind als Kunstobjekte ausgestellt worden. Duchamp machte mit den Werken darauf aufmerksam, daß Zivilisationsgegenstände keine Bedeutung haben, aber sogleich eine bekommen, wenn Vermittlung von Zeichensystemen oder der Kultur auftritt.]

Wie verhält es sich dabei mit den sozialen Systemen, die ja hier als Träger der Kultur aufgefaßt werden? Die menschlichen Gesellschaften produzieren Kunst und Literatur, und insofern sind diese auch auf soziale Prozesse zurückzuführen bzw. resultieren aus ihnen. Andererseits entwickelt auch die Kunst Eigendynamik, die auf die Dynamik der Zeichensysteme allgemein zurückzuführen ist, und diese Eigendynamik beeinflußt wiederum die Entwicklungen bzw. den Zustand des sozialen Systems. Die soziologisch orientierten Kunsttheorien liefern hier genügend Belege (siehe z.B. Warning, 1975; Link & Link-Heer, 1978 - darin auch ausführliche Bibliographie). Es ist keine eindeutige und einseitige Relation, auch

hier gibt es Wechselwirkungen diverser Art, sowohl positive als auch negative Rückkopplungen.

3.4. Ein zusätzliches Problem ergibt sich, wenn man an die Zivilisation und das soziale System im Hinblick auf die oben unterschiedenen Ebenen denkt. Besitzt jede Ebene einen eigenen Zivilisations- und sozialen Bereich, oder bezieht sich jede Ebene auf ein und denselben sozialen Bereich? Es scheint (zumindest auf den ersten Blick), daß die Zivilisation ein homogeneres Gebilde darstellt als die Kultur. Besonders in der heutigen Zeit und in der westlichen Welt ist auf jeder Kulturebene das gleiche Zivilisations-Repertoire vorhanden, d.h. die Menge der Zivilisationselemente ist annähernd gleich (wenn nicht, kann sie ohne weiteres ergänzt werden: Eine in Amerika beispielsweise entstandene Gerätegeneration (Computer, Bodybuildinggeräte usw.) wird sich in kürzester Zeit im ganzen System ausbreiten). Es ist jedoch keine allgemeingültige Erscheinung. In früheren Epochen und auch heute - im Hinblick auf die Entwicklungsländer - war und ist es nicht immer der Fall. Auch der Zivilisationsbereich weist Differenzierungen auf und besteht nicht überall aus den gleichen Elementen. Dennoch kann man sagen, daß die Zivilisation allgemein homogener ist bzw. Differenzen rascher ausgleicht und zur Homogenität sozusagen »neigt«, als das bei den sozialen Systemen der Fall ist. Diese verändern sich langsamer, und die Homogenität ist nicht unbedingt ein angestrebter Zustand. Die Zivilisation ist homogener, und dieser Zustand des Ausgleichs im Hinblick auf zivilisatorische Gegenstände und Fertigkeiten wird angestrebt.

Anders verhält es sich im Hinblick auf die sozialen Systeme. Hier sind Differenzen ein konstitutiver Bestandteil des Systems selbst (wenn man die verschiedenen Ebenen isoliert betrachtet). So wird eine bestimmte Subkultur *und* deren Kunst von einer sozialen Gruppe (einem sozialen Subsystem also) getragen, die andere Mitglieder umfaßt als (a) andere Subkulturen (wobei hier Überlappungen seltener auftreten), als (b) eine bestimmte Einzelkultur oder bestimmte Einzelkulturen (wobei hier Überlappungen durchaus nicht selten sind - frankophil orientierte Alternativgruppen z.B.) und als (c) *die* Kultur. Die Komplexität des sozialen Systems steigt mit dem ansteigenden Ebenenniveau.

Dies ist auch nicht verwunderlich, denn wenn man davon ausgeht, daß die Kunst-Menge auf jeder Ebene quantitativ und funktionell unterschiedlich ist, dann muß man folglich auch annehmen, daß das soziale System, das der Träger der Kunst ist, auch eine quantitativ und funktionell unterschiedliche Menge darstellt. Das soziale System ist also auch eine ebenenartige und vernetzte Gesamtheit, die sich durch quantitativ ansteigende Komplexität und funktionelle Varianz auszeichnet. Dieser Tatsache ist in Figur (1) Rechnung getragen, indem auf jeder Ebene

soziale Subsysteme eingeführt worden sind, die zusammengenommen (integrativ, nicht summativ) das soziale System der Menschheit ergeben.

Die Zivilisation ist - vorsichtig formuliert - ein homogeneres Gebilde und umfaßt für jede Ebene praktisch die gleiche Menge. Sie kann sich geographisch von Einzelkultur zu Einzelkultur im Umfang unterscheiden, sie ist jedoch prinzipiell erweiterbar. Was noch nichts über deren Semiotisierung aussagt. Man kann sagen: Der Zivilisationsbereich von Subkulturen ist mit dem der Einzelkultur identisch, die Zivilisationsbereiche der Einzelkulturen sind mit dem *der* Kultur identisch usw. Im Hinblick auf die Kultur ist dies nicht möglich, weil man hier mit semiotischen Phänomenen zu tun hat, und die Gesetzmäßigkeiten von Zeichensystemen lassen keine mechanischen Aneinanderreihungen zu. Die Hinzufügung von neuen Elementen verändert das ganze System, zumal es noch mit verschiedenen sozialen Subsystemen verbunden ist, die ebenfalls verändert werden müssen. Da aber die Zivilisation kein Zeichenphänomen bildet, ist hier eine Homogenisierung einfacher. Die Zeichensysteme wirken auf die Benutzer zurück und steuern diese auch, weil sie bestimmte Eigengesetze und Eigendynamik aufweisen. Die Zivilisation scheint eine solche Rückwirkung nicht zustande zu bringen, da sie im Hinblick auf das soziale System untergeordnet ist.

4. Kultur als offenes System

In Anlehnung an das Weissche Systemkriterium (Weiss, 1970) wird die Kultur als System aufgefaßt. Offen bleibt noch die Frage, was für ein System sie ist, welche Merkmale sie besitzt, aus welchen Elementen sie besteht und wie ihre Zusammensetzung ist.

Die Kultur ist die Gesamtheit der Zeichensysteme und der in ihnen kodierten Nachrichten. Auf verschiedenen Ebenen zeichnet sich die Kultur durch unterschiedliche Komplexität aus. Die in einem sozialen Raum funktionierenden, d.h. angewandten Zeichensysteme und aktualisierten Nachrichten machen (entsprechend auf jeder Ebene) die Menge der **aktuellen Kultur** aus. Die Gesamtheit aller Nachrichten und Zeichensysteme bildet die **potentielle Kultur**. Die potentielle Kultur ist also eine Integration der aktuellen Kultur und der Tradition. Die Tradition an und für sich zerfällt wiederum in aktualisierte und potentielle Tradition, d.h. eine solche, deren Inhalte gespeichert und abrufbar sind, aber in der gegebenen Kultur zum gegebenen Zeitpunkt nicht funktionieren.

All diese Bereiche sind nicht als abgeschlossen aufzufassen, es sind jeweils offene Subsysteme, die untereinander ständigen Wechselwirkungen unterliegen; die Bereiche ändern ständig ihre Zusammensetzung, es sind dynamische Gebilde.

Die Kultur wird als offenes und dynamisches System aufgefaßt.

4.1. Aus der Thermodynamik weiß man, daß sich offene Systeme durch einen ständigen Austausch von Materie und Energie zwischen Umwelt und sich selbst auszeichnen (über Eigenschaften offener Systeme siehe detailliert Fleischer, 1989a:20-27). Als Umwelt für die Kultur ist oben die Zivilisation postuliert worden; die Umwelt der Einzelkulturen ist die Kultur und die Zivilisation. Wechselwirkungen zwischen ihnen können als vorhanden angesehen und brauchen hier nicht näher besprochen werden.

Ein Problem stellt die Materie- und die Energie-Auffassung im Hinblick auf die Kultur dar. Es wird folgende Lösung vorgeschlagen: Als Materie-Austausch kann man die materielle Beschaffenheit der Zeichen-Träger interpretieren. Die Zivilisation liefert ständig und ununterbrochen die materielle Basis der Kultur - die Materie der Zeichensysteme. Ohne Zufluß von Materie (Zivilisationsgegenstände, -fertigkeiten usw.) könnte die Kultur kein System aufbauen. Als Energie kann man den Informationsfluß zwischen Zivilisation und Kultur auffassen. Gestützt auf die (in Fleischer, 1989a:17-19, 21-23 dargestellten) Erkenntnisse, wonach man im Hinblick auf offene Systeme 'Information' und 'Energie' gleichsetzen kann, ist es möglich, Information, Informationsfluß und Informationsaufnahme als Energiezufluß des offenen Systems Kultur zu interpretieren.

4.2. Man könnte nun aber fragen: Welche Information denn, wenn doch die Zivilisation als Bereich, in dem keine Zeichen vorkommen, aufgefaßt wird? Das Problem ist - zugegeben - nicht einfach. Es werden von der Zivilisation - und das ist der Materie-Bereich - die materiellen Objekte der Zeichenträger geliefert. Darüber hinaus liefert die aus Gegenständen und Fertigkeiten bestehende Zivilisation durch ihr Vorhandensein auch Informationen, d.h. Kenntnisse über bestimmte Zustände oder Fertigkeiten, die von der Kultur aufgenommen, d.h. als Information betrachtet werden. Es ist also der Unterschied zwischen Zeichen und Information bzw. zwischen Bedeutung und Information, der die Sache schwierig macht. Zeichen besitzen Interpretanten. Diese werden (siehe Fleischer, 1990:95-101, 147-156) in Bedeutungs-Interpretanten und in Zeichen-Interpretanten gegliedert. Die erstgenannten legen die Zeichen als kulturelle Fakten fest, die letzteren verbinden die Zeichen miteinander und bilden insofern komplexe kulturelle Systeme - Nachrichten, Nachrichtensorten usw.

Darüber hinaus liefern Zeichen auch Informationen (d.h. beseitigen eine Unkenntnis), die noch nichts mit Bedeutung oder gar mit Interpretanten zu tun haben. Daneben gibt es Informationen, die nicht auf Zeichen aufgebaut sind und nicht aus Zeichen-Verknüpfungen entstehen (siehe näheres dazu Fleischer, 1990). Information ist also ein breiteres Phänomen; Information liefern nicht nur Zeichen.

Es geht hier also um das Problem: Bedeutung vs. Information. Etwas kann viel Information liefern, ohne eine Bedeutung zu haben. Und etwas kann Bedeutung haben, ohne Information zu liefern. Auf diesen Unterschied wird Bezug genommen, wenn die Information als Energiezufluß des Systems Kultur betrachtet wird.

4.3. Die Zivilisation liefert Gegebenheiten, Fertigkeiten u. dgl., die Information vermitteln. Es scheint nicht abwegig zu sein, die materiellen Zeichenträger und die Information als Materie- bzw. Energie-Zufluß in offenen Systemen zu begreifen.

Auf der anderen Seite muß - soll die Kultur ein offenes System sein - ein Energie- und Materie-Ausstoß stattfinden, der durch einen geringeren Ordnungsgrad gekennzeichnet ist. Es müßte sich zeigen lassen, daß die Kultur an die Umwelt Produkte abgibt, die einen höheren Entropiegrad besitzen. Als solche Zerfallsprodukte kann man ganz einfach den Müll, den eine Kultur produziert, auffassen, der in die Umwelt - in die Zivilisation - abgegeben wird. Es wäre doch kein zu einfaches Beispiel, wenn abgenutzte, zu zerstampfende Bücher oder Computerprogramme als Ausstoß-Produkte genannt werden, oder - wenn es um die Energie geht - nicht mehr »gebrauchbare« Informationen. All dies diente zur Aufrechterhaltung der Kultur, half dabei mit, den Ordnungsgrad des Systems Kultur zu erhöhen, und wird nun ausgeschieden.

Man kann also die Kultur als System betrachten, das aus einem Materie- und Energiefluß seine Ordnung aufbaut und aufrechterhält und entropiereiche Abfallprodukte ausscheidet. Dies betrifft die Relationen zwischen Kultur und Umwelt. Die Relationen zwischen einer Einzelkultur und *der* Kultur, zwischen einer Subkultur und der Einzelkultur sind - was die Ströme des Zuflusses und Abflusses betrifft - einfacher und brauchen hier nicht gesondert besprochen zu werden. Jede Einzelkultur oder Subkultur schöpft aus der übergeordneten Ebene Zustände höherer Ordnung, verbraucht sie, gibt die Abfallprodukte an die Zivilisation ab und (und das ist neu) reichert das jeweilige Suprasystem, dessen Bestandteil es ist, an. Es erhält somit und erhöht sein eigenes Ordnungsniveau. Ob diese Analogien tatsächlich zutreffen, bleibt natürlich eine andere Frage.

4.4. Es ist eine beobachtbare Tatsache, daß *die* Kultur (wie auch jede Einzelkultur usf.) sowohl relativ feste als auch relativ freie Bereiche besitzt. Die Kultur beinhaltet eine bestimmte Menge fester, invarianter Bestandteile, die alt und äußerst beständig sind, und darüber hinaus eine Menge jüngerer und instabiler akzessorischer Bestandteile; die Kultur ist kein homogenes System, sie besitzt eine komplexe dynamische innere Gliederung.

Aus der Thermodynamik weiß man (siehe Fleischer, 1989a:15-27), daß (laut zweitem Hauptsatz) die Entropie eines geschlossenen Systems ständig wächst, d.h. es kommt zum Ordnungsabbau; das System nimmt einen Zustand höchster Wahr-

scheinlichkeit ein bzw. strebt - allgemein - Zustände höherer Wahrscheinlichkeit an. In offenen Systemen wird der zweite Hauptsatz dermaßen »umgangen«, daß das offene System die Ordnung der Umwelt nutzt und selber Ordnung anbaut, dafür aber den entsprechenden Betrag an Unordnung - Entropie also - an die Umwelt wieder abgibt. Die Entropie der Umwelt steigt, die des Systems verringert sich. Dieser Vorgang scheint auch für das System der Kultur zuzutreffen.

Wenn die Kultur ein offenes System sein soll, dann muß es sich nachweisen lassen, daß sie ein System fern vom Gleichgewicht ist, und also durch einen stationären Zustand bzw. ein Fließgleichgewicht charakterisiert werden kann. Sie muß Schwankungen bzw. Instabilitäten, Selbstorganisation und steigende Ordnung aufweisen, sie muß ein dissipatives System sein. Wenn es sich nachweisen ließe.

Literatur

Bense, M. (1971). Systemtheoretische Erweiterungen des Zeichenbegriffs. *Zeitschrift für Literaturwissenschaft und Linguistik 1/2,* 91-96.

Bense, M. (1975). Semiotische Prozesse und Systeme. In: *Wissenschaftstheorie und Design.* Baden-Baden.

Bense, M., & Walther, E. (1973). *Wörterbuch der Semiotik.* Köln.

Bertalanffy, L.v. (1945). Zu einer allgemeinen Systemlehre. *Blätter für deutsche Philosophie 18(3/4),* 112-132.

Bertalanffy, L.v. (1950). The theory of open systems in physics and biology. *Science 111,* 23-29.

Bertalanffy, L.v. (1953). *Biophysik des Fließgleichgewichts.* Braunschweig.

Bertalanffy, L.v. (1957). Allgemeine Systemtheorie. Wege zu einer neuen Mathesis Universalis. *Deutsche Universitätszeitung Bd.12, H.5/6,* 8-12.

Bertalanffy, L.v. (1968). *General System Theory. Foundations, Development, Applications.* New York.

Bertalanffy, L.v. (1970). Gesetz oder Zufall: Systemtheorie und Selektion. In: Koestler, A., & Smythies, J.R. (Hrsg.), *Das neue Menschenbild.* Wien, 71-95.

Bertalanffy, L. v. (Hrsg.) (1972a). *Systemtheorie.* Berlin.

Bonner, J.T. (1983). *Kultur-Evolution bei Tieren.* Berlin und Hamburg.

Bühler, K. (1965). *Sprachtheorie. Die Darstellung der Sprache.* Stuttgart.

Burkhard, S. (1977). *Die Evolution der Sozialstrukturen.* Berlin.

Burkhardt, D., Schleidt, W., & Altner, H. (1972). *Signale in der Tierwelt. Vom Vorsprung der Natur.* München.

Dawkins, R. (1978). *Das egoistische Gen.* Berlin.

Dose, K., & Rauchfuss, H. (1975). *Chemische Evolution und der Ursprung lebender Systeme.* Stuttgart.

Dzwillo, M. (1978). *Prinzipien der Evolution. Phylogenetik und Systematik.* Stuttgart.

Ebeling, W. (1976). *Strukturbildung bei irreversiblen Prozessen.* Leipzig.

Ebeling, W., & Feistel, R. (1982). *Physik der Selbstorganisation und Evolution.* Berlin.

Ebert, R. (1974). Entropie und Struktur kosmischer Systeme. In: v. Weizsäcker, E. (Hrsg.), *Offene Systeme I. Beiträge zur Zeitstruktur von Information, Entropie und Evolution.* Stuttgart, 222-228.

Eigen, M. (1970). Selbstorganisation der Materie und die Evolution biologischer Makromoleküle. *Naturwissenschaftliche Rundschau 23,* 777-779.

Eigen, M. (1979). Sprache und Lernen auf molekularer Ebene. In: Peisl, A., &. Mohler, A. (Hrsg.), *Der Mensch und seine Sprache,* 181-218.

Eigen, M., Gardiner, W., Schuster, P., & Winkler-Oswatitsch, R. (1981). Ursprung der genetischen Information. *Spektrum der Wissenschaft 6,* 37-56.

Fleischer, M. (1987). *Hund und Mensch. Eine semiotische Analyse ihrer Kommunikation.* Tübingen.

Fleischer, M. (1989). *Die sowjetische Semiotik. Theoretische Grundlagen der Moskauer und Tartuer Schule.* Tübingen.

Fleischer, M. (1989a). *Die Evolution der Literatur und Kultur. Grundsatzfragen zum Entwicklungsproblem (ein systemtheoretisches Modell).* Bochum.

Fleischer, M. (1990). *Information und Bedeutung. Ein systemtheoretisches Modell des Kommunikationsprozesses (und das Problem des Verstehens).* Bochum.

Fleischer, M. (1991). *Die Semiotik des Spruches. Kulturelle Dimensionen moderner Sprüche.* Bochum.

Fleischer, M. (1994). *Die Wirklichkeit der Zeichen. Empirische Kultur- und Literaturwissenschaft (Systemtheoretische Grundlagen und Hypothesen).* Bochum.

Fleischer, M., & Sappok, Ch. (1988). *Die Populäre Literatur. Analysen literarischer Randbereiche an slavischem und deutschem Material.* Bochum.

Glansdorff, D., & Prigogine, I. (1971). *Thermodynamic Theory of Structure, Stability and Fluctuations.* London, New York.

Haken, H. (1983). *Synergetik.* Berlin.

Haken, H., & Graham, R. (1971). Synergetik - die Lehre vom Zusammenwirken. *Umschau, 71.Jhg., H.6,* 191-195.

Hall, A.D., & Fagen, R.E. (1956). Definition of system. *General Systems 1,* 18-28.

Hassenstein, B. (1972), Element und System - geschlossene und offene Systeme. In: Kurzrock, R. (Hrsg.), *Systemtheorie*. Berlin, 29-38.

Jachnow, H. (1981). Sprachliche Funktionen und ihr Hierarchiegefüge. In: Esser, J., & Hübler, A. (Hrsg.), *Form and Functions*. Tübingen, 11-24.

Kattmann, U. (1980). Fließgleichgewicht und Homöostase. Zur kybernetischen Beschreibung von Biosystemen. Teil I. *Der mathematische und naturwissenschaftliche Unterricht (NMU), 33. Jg., H.4,* 202-209.

Kattmann, U. (1980a). Das homöostatisch gesicherte Fließgleichgewicht. Zur kybernetischen Beschreibung von Biosystemen. Teil II. *Der mathematische und naturwissenschaftliche Unterricht (NMU) 33.Jg., H.5,* 283-289.

Koch, W.A. (1981a). Evolution des Kreativen: Symmetrie, Asymmetrie, Integration. In: Schnelle, H. (Hrsg.), *Sprache und Gehirn*. Roman Jakobson zu Ehren. Frankfurt, 158-173.

Koch, W.A. (1986). *Evolutionäre Kultursemiotik*. Bochum.

Link, J., & Link-Heer, U. (1978). *Literatursoziologisches Propädeutikum*. München.

Miller, J.G. (1978). *Living systems*. New York.

Peirce, Ch.S. (1931-1935). *Collected Papers*. Vols. I-VI. Cambridge. Vols. VII-VIII. 1958.

Peirce, Ch.S. (1967, 1970). *Schriften I. Zur Entstehung des Pragmatismus*. Frankfurt. *Schriften II. Zum Pragmatismus*. Frankfurt.

Popper, K. (1974). *Objektive Erkenntnis. Ein evolutionärer Entwurf*. Hamburg.

Posner, R. (1991). Kultur als Zeichensystem: Zur semiotischen Explikation kulturwissenschaftlicher Grundbegriffe. In: Assmann, A., & Harth, D., (Hrsg.), *Kultur als Lebenswelt und Monument*. Frankfurt, 36-74.

Prigogine, I. (1979). *Vom Sein zum Werden. Zeit und Komplexität in den Naturwissenschaften*. München.

Prigogine, I., & Glansdorff, D. (1971). *Thermodynamic Theory of Structure, Stability and Fluctuations*. London, New York.

Prigogine, I., & Nicolis, G. (1977). *Self-organization in non-equilibrium systems, from dissipative structures to order through fluctuations*. New York.

Prigogine, I., & Stengers, I. (1981). *Dialog mit der Natur. Neue Wege naturwissenschaftlichen Denkens*. München.

Riedl, R. (1975). *Die Ordnung des Lebendigen. Systembedingungen der Evolution*. Hamburg.

Riedl, R. (1976). *Die Strategie der Genesis. Naturgeschichte der realen Welt*. München.

Riedl, R. (1984). Evolution und evolutionäre Erkenntnis - Zur Übereinstimmung der Ordnung des Denkens und der Natur. In: Lorenz, K., & Wuketits, F., (Hrsg.), *Die Evolution des Denkens*. Zürich.

Riedl, R. (1987). *Kultur - Spätzündung der Evolution? Antworten auf Fragen an die Evolutions- und Erkenntnistheorie*. München.

Schuster, P. (1977). Selbstorganisationsprozesse in der Biologie und ihre Beziehung zum Ursprung des Lebens. *Der mathematische und naturwissenschaftliche Unterricht 30,* 324-335.

Warning, R., (Hrsg.) (1975). *Rezeptionsästhetik*. München.

Weiss, P. (1970). Das lebende System: Ein Beispiel für den Sichtendeterminismus. In: Koestler, A., & Smythies, J.R. (Hrsg.), *Das neue Menschenbild*. Wien, 13-70.

Weiss, P. (1978). Empirische Grundlagen des Systemdenkens. *Nova Acta Leopoldina (N.F.) 47 (226),* 325-334.

Wuketits, F. (1979). Gesetz und Freiheit in der Evolution der Organismen. *Umschau 79,* 268-275.

Wuketits, F. (1981). Die Systemtheorie der Evolution - Eine neue Sehweise der Entwicklung des Lebendigen. *Der mathematische und naturwissenschaftliche Unterricht (MNU) 34,* 1-7.

Wuketits, F. (1981a). *Biologie und Kausalität. Biologische Ansätze zur Kausalität, Determination und Freiheit*. Berlin.

Wuketits, F. (1983). *Die Evolution des Denkens*. München.

Wuketits, F. (1985). Die systemtheoretische Innovation der Evolutionslehre. In: Ott, J.A., Wagner, G.P., & Wuketits, F. (Hrsg.), *Evolution, Ordnung, Erkenntnis*. Berlin, 69-81.

Glottometrika 15, 1996, 22-37

From Language to Reality

Constantin Thiopoulos, Athens

1. Introduction

The representationalistic approach to meaning, according to which meaning is a denotational mapping between a language expression and a corresponding idea, does not take into consideration the semiotic processes that constitute this mapping. But if we look at language as a self-organizing system, it is exactly these processes that are of interest, because only by focussing on them can we explore the self-organization of meaning. This area, which lies on the borderline between semiotics, linguistic semantics and philosophy of knowledge, remains totally uninvestigated, the reason being that semantics is dominated by the representationalistic attitude of the rationalistic paradigm. Both in the Chomskyan approach to language and in model theory, meaning is seen as a fixed association of a language term to a given structure. In transformational grammar the meaning of a lexical entry is a set of semantic markers (see Katz, 1965, 1972) that represent objective concepts, and in model theory the meaning of a term is a denotational mapping to some given formal structure. This attitude does not capture at all the dynamic nature of meaning constitution and moreover blocks the way to a genetical interpretation of semantics, i.e. to an investigation of the constitution of meaning structures; instead, a synchronic meaning catalogization is presupposed.

Because of compositionality, i.e. that the meaning of a complex expression is a function of the meaning of the constituents, denotational semantics works only on hierarchical structures.[1] The main characteristic of semiotic systems is that they are circular structures (see Thiopoulos, 1992). This circularity is artificially[2]

[1] In semantics for transformational grammar the compositionality principle is immanent in the *projection rules* (Katz, 1965).

[2] The artificiality becomes apparent in metaphors. In order to explain metaphors transformational grammar posits *feature transfer* (Levin, 1977), which results in a shortcut of the hierarchy of the redundancy rules.

broken by assuming that language is a means for expressing internal ideas that are printed in our mind and which form something like logical atoms[3]. Again, this view does not capture the dynamic interplay between concepts, as self-organizing schemata for interpreting the world and language, i.e. it cannot explain language as a product of human interaction but has to presuppose it as a faculty given to us by God in order to be able to express our thoughts.

> *"Begriffe sind keine fertigen Gebilde, die uns ein deus ex machina in endgültiger Gestalt zuspielt, oder die wir aus der Realität in reiner Form extrahieren"* (Köhler & Altmann, 1993)

If we want to focus on the self-organizing dimension of meaning then we have to start from human interaction and understand how it is possible that two persons in a communicative situation can use one expression to refer to one thing instead of assuming that there is a preestablished harmony in their minds and that language plays the subordinate role of expressing this harmony. The following paper is an attempt to develop a semiolinguistic frame for the description of the semiotic processes, whose synergy constitutes the self-organization of meaning.

2. Language as a semiotic topos

The core of the rationalistic approach to language consists in viewing language as a vehicle for exchanging ideas about an external reality. Conforming to the mechanistic attitude of the deductive method, which forms the basis of scientific analysis, language is seen as a system for combining lexemes, conceived as primitive objects, into sentences.

> *"... car il n'y a que deux choses à apprendre en toutes de langues, à savoir la signification des mots, et la grammaire. Pour la signification des mots, il n'y promet rien de particulier; car il dit en la quatrième proposition: lingua illam interpretari ex dictionario, qui est ce qu'un jomme un pue versé aux langues peut faire, sans lui en toutes le langues communes. Et je m'assure, que vous donniez à M. Hardy un bon dictionnaire en chinois, ou en quelque autre langue que ce soit, et un*

[3] In semantics for transformational grammar (Katz, 1965, 1972) the circular structures of the lexicon are broken by assuming that the semantic markers are *not* words but concepts, which are eternal entities, a sort of platonic ideas (see also Peterson, 1973).

> *livre écrit en la même langue, qu'il enterprendra d'en tirer les sens. Ce qui empèche que tout le monde ne les pourrait pas faire, c'est la difficulté de la grammaire."* (Letter to Mersenne from November 1629 in Descartes, 1953).

Signification of lexemes and *grammar* are the two poles of language. Signification determines the inner aspect of language expressions, i.e. the corresponding ideas, and is universally identifiable, since it is a faculty of the human mind and therefore independent of a specific language. The signification of lexemes leads to the primitive ideas that form the logical atoms, which are combined according to the laws of reason into more complex ideas that are expressed in turn as sentences[4]. Since the laws of reason are immanent in the grammar of all languages, they constitute a universal grammar that incorporates the substantial aspects of the human mind. Besides these universal aspects every grammar of a specific language has its own rules, but these rules concern only the phonetic representation and not the underlying ideas. Standing in this tradition (see especially Chomsky, 1966), the Chomskyan paradigm is grounded on the difference between *deep structure*, conceived as the underlying semantic representation, and *surface structure*, conceived as the corresponding phonetic representation. Since, from this point of view, language is an expression of our thoughts, communication is possible exactly because language can be used as a medium for the exchange of ideas, i.e. the language user is a cognitive system that receives phonetic data through the receptors, sends these data to the brain, where the corresponding ideas are activated, and calculates, according to the laws of reason, other ideas that are in turn translated into phonetic output through the effectors. The guarantee for "understanding", i.e. that both partners in a communication act "talk about the same things", is the fact that the ideas of both of them are images of an objective reality, which is analyzed according to the universal laws of reason.

Recent results in neurology[5] demonstrated that the brain receives an already

[4] This is clearly stated in the principal work of the rationalistic philosophy of language (Arnauld & Lancelot, 1660/1662).

[5] Hubel (1960) has shown that the visual pathway cannot be seen as a channel in the sense of information theory. Maturana et al. (1960) discovered that there are special retina cells in the eye of the frog that react when a dark flying object activates them. This means that the behavior of the frog cannot be modelled after the receptors-brain-effectors schema, since the frog chases after the fly (or after a dark flying object which is not a fly) not according to a message from the brain but because of its neurological attunement. A similar anti-behaviorist position is presented by ethology (Lorenz, 1982).

highly selected structure of sense data and that therefore their manipulation does not take place only in the mind. This led to the *autopoietic* turn in cybernetics.

> *"But what was still more fundamental was the discovery that one had to close off the nervous system to account for its operation, and that perception should not be viewed as a grasping of an external reality, but rather as the specification of one, because no distinction was possible between perception and hallucination in the operation of the nervous system as a closed network... Two immediate consequences arose from this: ... The first consequence required that the question "How does the organism obtain information about its environment?" be changed to: "How does it happen that the organism has the structure that permits it to operate adequately in the medium in which it exists?" A semantic question had to be changed into a structural one. The second question required the actual attempt to describe the phenomena that take place in the organism during the occurrence of the phenomena of perception and cognition in a language that retained them as phenomena proper to a closed nervous system"* (Maturana & Varela, 1980:XV-XVI).

The ontological consequence is that reality is not an external objective sphere, since what the organism sees as reality depends on its attunement, but an intersubjectively constituted domain. A cognitive system is a closed system that is formed through continuous interaction with the environment and which constitutes in turn its environment according to the sedimentation of this interaction. This process of mutual perturbations between system and environment is called *structural coupling* (Maturana & Varela, 1980) and describes a predualistic level, on which the system cannot differentiate itself from the environment. Environmental reality[6] arises as a structural coupling between structurally coupled systems.

Phenomenology comes to similar results. Husserl's concept of *life-world*, as the immediate flow of unreflected life, and Heidegger's of *being-in-the-world* express this predualistic embedding in the environment. Exactly as cognitive systems are closed neural systems, they are also closed semiotic systems. It is closedness that differentiates them from the background and constitutes their self-sufficiency. The rationalistic approach to *meaning*, as signification of ideas, hides the domain of

[6] This corresponds to the *behavioral environment* of the *Gestalt* theory. *"We can explore directly only our behavioral environment, and indirectly merely, through the behavioral, the geographical one"* (Koffka, 1935:37).

meaningfulness[7], which characterizes the essential semioticality of experience: what is meaningful for a cognitive system has meaning *for it.*

> *"We may call this condition of reviewing reality and picking out the specific from the indifferent entities "knowledge" in its most elementary state. Knowledge, even at advanced stages of evolution, always involves the recognition of specificity or of meaningfulness for the knower"* (Prodi, 1989:95).

At the *first level of intersubjectivity* it is behavior that encloses semiologically environmental reality. As Merleau-Ponty[8] has shown, behavior cannot be analyzed in a rationalistic manner. At the *second level of intersubjectivity*, cultural reality is constituted through the use of language.

> *"By 'habitual thought' and 'thought world' I mean more than simply language, i.e. than the linguistic patterns themselves. I include all the analogical and suggestive value of the patterns (e.g., our 'imaginary space' and its distinct implications), and all the give-and-take between language and the culture as a whole, wherein is a vast amount that is not linguistic but yet shows the shaping influence of language"* (Whorf, 1965).

Language can therefore not be reduced to a mechanistic combination of signification and grammar, since it opens a semiotic perspective that constitutes the intersubjective cultural domain of all users of a specific language. Language is not a mirror of reality but reality a mirror of language.

> *"Human beings can talk about things because they generate the things they talk about by talking about them"* (Maturana, 1978:56).

[7] Bedeutsamkeit (Heidegger, 1927).

[8] In Merleau-Ponty (1942) the examination of experimental results from aphasic patients leads to the conclusion that aphasic diseases cannot be explained in a mechanistic way, since they neither can be physiologically localized nor explained in terms of reductionistic psychology; or they are rather modifications of the structure of behavior. This means that behavior cannot be analyzed in atoms. Behavior can only be understood as an incarnation of meaningfulness. This leads in Merleau-Ponty (1945) to the reexamination of perception not as a mental grasping of an objective reality but as an attitute towards the world (a *belief*, as Merleau-Ponty says), i.e. as the constitution of a reality.

If we want to penetrate into the core of communication we have to understand how "reality" is constituted by language and not vice versa. We have to start from that which is primarily given, and that is the *use of language* not language.

> *"Die erste und öffentliche Voraussetzung der Sprachwissenschaft will, daß es eine Sprache gibt. Und gerade das ist unsicher. So wenig wie aus dem Vorhandensein von Theologie folgt, daß es einen Gott gibt, noch aus geometrischen Lehrsätzen über den Kreis, daß solche Dinge in Wirklichkeit vorkommen, so wenig geht aus der gesamten Sprachwissenschaft die Gewähr hervor, daß es Sprache gibt. Zunächst gibt es tatsächlich keine, sondern nur das Sprechen: mein Sprechen, dein Sprechen, unser Sprechen von jetzt und hier heute und gestern usw. Unser Sprechen ist aber noch keine Sprache, sondern höchstens ein Gespräch. Und auch an diesem müßte man zweifeln wenn mein Sprechen nicht von einem andern gehört, verstanden und irgendwie beantwortet würde"* (Vossler, 1960:14).

The phenomenon that the other understands "mein Sprechen" points to an underlying structure[9], which is the sedimentation of the use of language, i.e. the system-driven finding of invariants of communicative situations. The mutual use of expressions results in the habitual sedimentation of metonymic connections, i.e. a cognitive system connects expressions that are used together in communicative situations. The sedimented *metonymic network* is an incarnation of *langue* and forms, as a way of *interpreting*[10] signs, a *semiotic topos*. Language, as the common core of a community of language users, is an intersystematic invariant reflecting the cultural reality of this community, i.e. it is the semiotic topos of all semiotic topoi.

3. On the semiotic foundation of categorization

From the reductionistic point of view categorization is based on internal schemata, which when applied to particular objects, considered as *tokens*, recognize these

[9] In terms of Glossematics: the *system* behind the *process* (Hjelmslev, 1961).

[10] *Interpretation* is used here in the sense of Peirce. The *final interpretant* leads thereby, as a new habitual sedimentation, to the modification of the metonymic network, something that has been stressed in Madison (1982) and Eco (1979).

objects as instantiations of a category, considered as the corresponding *type*, or not. Categorization can thus be modelled by a set of rules of the form:

if token *x* has the features *a, b*,... then it is of type *X*.

The gradual association of a given object to a certain category, which has been based on Rosch (1973), Rosch & Lloyd (1978) and which has been modelled by fuzzy sets, is an extension of the classical theory, in the sense that in order to identify a given token *x* as an instantiation of a type *X* it suffices if *x* satisfies some of the features of the corresponding rule. Some of the features are therefore seen as central and tokens satisfying only these features are *typical*, while tokens satisfying these features but also other features not included in the corresponding rule are seen as belonging to the same category, but to a lesser degree.

If we consider a cognitive system as an autopoietic system we have to understand categories as invariants of structural coupling. But what kind of invariants? If we take into consideration the extremely large amount of perturbances coming from the environment then it seems necessary that the system - in order to act efficiently - must build clusterings.

This grouping enables the system not only to handle a group as a single higher level entity, but also to "recognize" new perturbances as correlated to some of these clusterings and so to react faster in a changing environment, i.e. to be better adapted to it. A perturbance is not a stimulus, which - from a behavioristic point of view - serves as input, which is transferred through the nervous system to the brain, considered as a processor, and which in turn calculates the corresponding output, but a fluctuation imposed on a closed system that results in an activation. The activation of a term in a metonymic semiotic topos determines the *semiotic meaning* of this term, conceived as *valeur*, i.e as the positional value of this term in the system considered. Taking into consideration that the interdependence of the terms is a metonymical sedimentation of their mutual use, then this value is a representation of the use of the term from the system's point of view, it corresponds therefore to the Wittgensteinian conception of meaning.

More than one perturbances force the system to find a stabilization between the generated activations[11]. The stabilization results in the focussing of a substructure, which contains the terms that are relevant, i.e. metonymically connected, to all of

[11] This becomes apparent in cases where we cannot decide between possible stabilizations (Wittgenstein's duck-rabbit).

the activated ones. This substructure is the *situation*[12], which constitutes the actual state of structural coupling, i.e. the system recognizes being involved in. The states of structural coupling are (phenomenologically understood) situations. Situations are both subjective and objective. Each situation is *subjectively* determined because it can only be recognized as a modification of the "structures" of situations, which are sedimented by an autopoietic entity, as invariants of structural coupling that determine behavior, and *objectively* determined because only some aspects of the behavioral "objects" involved are focussed, so that they are independent of the actual situation, in the sense that they are not exhaustively captured.

New activations constrain even more what the system sees as relevant, leading thus to a more restricted situation and so on until finally a situation has been reached, which is self-sufficient, i.e. all possible activations on this substructure cannot constrain it anymore. This final substructure can be seen as a Gestalt emerging out of the metonymic network and constitutes - as an invariant of structural coupling - a pole of orientation. Semiotic categories are such orientation poles, because they allow one to handle an immense amount of stimuli. Each term that is "trapped" in such a *fix situation* remains in it, i.e. the system has to recognize it as a part of an assembly that forms an orientation pole[13]. Fix situations are therefore self-sufficient subsystems that differentiate themselves from the background of the metonymic semiotic topos, leading also to a restriction of the semiotic meaning of a term on the fix situation, on which this term is trapped, and

[12] Situation is understood therefore in the sense of Heidegger, as existential horizon: "Die Situation ist das je in der Entschlossenheit erschlossene Da, als welches das existierende Seiende da ist. Die Situation ist nicht ein vorhandener Rahmen, in dem das Dasein vorkommt, oder in den es sich auch nur selbst brächte" (Heidegger, 1927:299) corresponding to the Husserlian conception of *actual noematas* (Husserl, 1952:109), as a horizon in a noetic structure, and not in the dualistic sense of situation semantics (Barwise & Perry, 1983). In situation semantics, meaning is modelled as a relation between referred situation and interpretation, whereby the interpretation of a situation contains all situations that are connected with the referred one via intercontextual relations between types of situations. This approach captures the subjective dimension of meaning generation but remains trapped in the rationalistic paradigm by considering situations as logical atoms of an objective reality, which is referred by natural language expressions. See also Thiopoulos (1990).

[13] Fix situations are semiotic attractors which characterize the self-organization of the semiotic topos and lead therefore to a context-sensitive actualization of a substructure of a cognitive system, something that has been stated by cognitive grammar models. Langacker (1987) speaks of *functional assemblies* and Lakoff (1982) of *idealized conceptual models* without describing a way of finding this structure.

express the twofold way of how a sign is interwoven in a semiotic system.

> *"Comme tout signe, l'objet est au carrefour de deux coordonnées, de deux définitions. La première des coordonnées, c'est ce que j'appellerais une coordonnée symbolique: tout objet a, si l'on peut dire, une profondeur métaphorique, il renvoie à un signifié;... La deuxième coordonnée est ce que j'appellerais la coordonnée de classement, ou coordonnée taxinomique;..."* (Barthes, 1985:253).

The role of a sign in a semiotic topos is therefore determined through the dynamic interplay between its semiotic meaning and its belonging to one or more fix situations. Considering a fix situation as a semiotic (sub-)topos it is possible to restrict the semiotic meaning of a sign with respect to this fix situation, describing thus the context-sensitiveness that is characteristic for semiotic systems.

The interdependence of the semiotic categories is not hierarchic, as in the reductionistic approach, but analogical. Besides the habitual "automatism" of the metonymic network there is a *metaphoric* capability of language, which consists in establishing links of analogy between fix situations, out of a quest for unification and as *μεταφορα* of structuredness, by scrutinizing the metonymic structure of these situations.

> *"... le sujet opérateur suscite une nouvelle grandeur, la catégorie, qui est comme une réplique à la demande d'unité venant de la nécessité originelle"* (Greimas & Fontanille, 1991:42).

In addition to contextual contiguity that leads to the metonymic network, analogy between fix situations introduces functional contiguity and results in the emergence of a *metaphoric network* out of the metonymic one that forms a metaphoric semiotic topos superimposed on the metonymic one. While metonymic links establish semiotic meaning, metaphoric links establish semiotic significance[14] as the interrelation of a term to major orientation poles. An analogical interconnection of fix situations, as orientation poles, opens a reference perspective as a way of "categorizing" perturbances from the environment and can be structurally

[14] Significs has been introduced by Welby (1985) as this part of semiotics that investigates the significance of signs, in contrast to their sense and meaning.

compared to the hermeneutic experience gained by the study of a text[15].

> *"The ultimate goal of textual interpretation is the appropriation of the text in the reading experience. That is, it is that of actualizing the meaning of the text as it is addressed to a reader; the ultimate significance of the text is the heightened self-understanding that the reader acquires by means of his or her dialogical encounter with the text"* (Madison, 1988: 59).

From this point of view, the introduction of a metaphor is not a "category mistake" but an essential creative aspect of language consisting in the uncovering of hidden similarities pointing to suppressed "Weltsichten", reflecting thus the self-organization of categorization[16]. In contrast to the rationalistic view the two poles of language are therefore not signification and grammar but metonymy and metaphor[17].

4. The communication process

> *"Wir sagen, daß wir ein Gespräch 'führen', aber je eigentlicher ein Gespräch ist, desto weniger liegt die Führung desselben in dem Willen des einen oder anderen Partners. So ist das eigentliche Gespräch niemals das, das wir führen wollten. Vielmehr ist es im allgemeinen richtiger zu sagen, daß wir in ein Gespräch geraten, wenn nicht gar, daß wir uns in ein Gespräch verwickeln. Wie da ein Wort das andere gibt,*

[15] For explaining hermeneutic experience Gadamer (1960) introduces the analogy of learning a foreign language. The new way of "looking at things" is not identical to one of a native speaker of this language, because it can be only a modification of the world view imposed by the original language. This process results in a sort of "hermeneutical dialectics".

[16] Ricoeur (1975) has explicitly stated this creative role of metaphor. Turbayne (1962) distinguishes three stages in the introduction of a metaphor: *inappropriateness* (sort-crossing), *acceptance* (triumph), *commonplace*. It is in this sense that Ricoeur speaks of symbols as "dead metaphors" (Ricoeur, 1976) and Merleau-Ponty (1945) of *speaking* and *spoken* language.

[17] In "Two aspects of language and two types of aphasic disturbances", Jakobson (1971) posits metonymy and metaphor as the two poles of language equating metonymy with *combination* and metaphor with *selection*. Metonymy is here used in the sense of combination but metaphor in a somewhat different sense, since it does not presuppose a paradigmatic dimension, as is the case for selection.

wie das Gespräch seine Wendungen nimmt, seinen Fortgang und seinen Ausgang findet, das mag sehr wohl eine Art Führung haben, aber in dieser Führung sind die Partner des Gesprächs weit weniger die Führenden als die Geführten. Was bei einem Gespräch 'herauskommt', weiß keiner vorher. Die Verständigung oder ihr Mißlingen ist wie ein Geschehen, das sich an uns vollzogen hat. So können wir dann sagen, das etwas ein gutes Gespräch war, oder auch, daß es unter keinem günstigen Stern stand. All das bekundet, daß das Gespräch seinen eigenen Geist hat, und daß die Sprache, die in ihm geführt wird, ihre eigene Wahrheit in sich trägt, d.h. etwas 'entbirgt' und heraustreten läßt, was fortan ist." (Gadamer, 1972:361).

The rationalistic view that communication is based on the correspondence between internal ideas and external objects blocks the way for exploring the semiotic processes that take place in a communicative situation.

"Man muss die Sprache nicht sowohl wie ein totes Erzeugtes, sondern wie eine Erzeugung ansehen, mehr von demjenigen abstrahieren, was sie als Bezeichnung der Gegenstände und Vermittlung des Verständnisses wirkt, und dagegen sorgfältiger auf ihren mit der inneren Geistestätigkeit eng verwebten Ursprung und ihren gegenseitigen Einfluß zurückgehen" (Humboldt, 1788:44)[18].

The frame of a communicative situation is given at the first level of intersubjectivity as a Gestalt of the behavioral environment. A language expression can be used to refer to a "behavioral thing"[19] only if this is part of similar behavioral situations, as substructures of behavioral semiotic systems, i.e. the expression emerges as a Gestalt that differentiates itself from the background as a highlighted intersubjective invariant[20].

[18] Humboldt defends exactly the opposite position to Descartes: "Das Zerschlagen in Wörter und Regel ist nur ein totes Machwerk wissenschaftlicher Zergliederung" (46).

[19] *Zuhandenes* (Heidegger, 1927)

[20] This is a striking characteristic of the *symbolic semographs* of hieroglyphics, i.e. they express the intersubjectively "essential" aspects of the represented concept (see, for example, Budge, 1966).

"Den Bedeutungen wachsen Wörter zu. Nicht aber werden Wörterdinge mit Bedeutungen versehen" (Heidegger, 1927:161).

This establishment of intersubjective invariants in a community characterizes the culture of this community and constitutes, as a heritage[21], the cultural context that frames the world view of the successors, making it necessary to enlarge the concept of semiotics in order to capture not only semiotic systems but also their interactions.

"On entrevoit déjà que, non moins que les systèmes de signes, les relations entre ces systèmes constitueront l'objet de la sémiologie" (Benveniste, 1974:50).

This autonomization forms the mimetic behavioral structure of language that introduces the second level of intersubjectivity[22]. The linguistic patterns are formed in analogy to the behavioral structures. An introduction of a new expression can now take place not only in a communicative situation where the partners are present, but also in written form, i.e. in a text[23]. A neologism or the introduction of a polysemy can only be "successful" if it expresses a metaphorical invariant, i.e. if it serves as a marker of similarity. "Reality" is now not only behaviorally but also culturally determined. The hermeneutic transformation that the reader of a text undergoes results from the resonance of his and the writer's semiotic topoi mediated through the cultural context and is guided by the reader's exploration of contextual and functional contiguities leading to a modification both of the reader's metonymic and metaphoric systems, not only by the finding of existing similarities but also by the introduction of new ones, i.e. not only the metaphoric

[21] There is an interplay between biological and cultural heritage. On the one hand neurologically sedimented invariants of behavior lead also to a biological heritage; Eibl-Eibesfeldt (1989) gives for example a thorough ethological investigation of phylogenetically determined invariants in the interpretation of facial expressions. On the other hand cultural heritage emerges out of communicative situations that are influenced by phylogenetical mechanisms.

[22] This is what Gadamer (1972) calls "Erhebung zur Welt".

[23] The birth of the alphabet is traced back to the moment when humans began to interpret signs inscribed on vases, used as containers for objects representing, for example, numbers of cows owned by a certain person. In the beginning, for every object in the vase there was an inscription on the surface of the vase, so that it was possible to see what was in the vase. Later these inscriptions totally replaced the contents of the vase leading thus to their autonomization.

connections are altered but the conception of similarity itself changes[24]. This "Ent-decken" implies an ontological commitment sedimented as habit leading to a semiotic fixation of "reality"[25]. Because of the continuous interplay between the metaphoric and the metonymic dimension the introduction of new metaphors does not change only the metaphoric network, but as a result of this also the semiotic meaning of an expression in a fix situation, since this expression is now seen in the light of the metaphorically imported structure.

Referencing of expressions at the first level of intersubjectivity is possible because these expressions have been introduced as orientations towards behavioral invariants and consists not in the picking up of isolated entities, but in the focussing of a situational Gestalt[26]. The metaphorical transfer of invariance opens new horizons for referencing and integrates the existing ones into the second level of intersubjectivity[27]. A neologism can now be used as a new invariant because of the metaphorical interrelation to an already existing invariant. In this way behavioral invariants are transferred to cultural invariants and these to further cultural invariants and so on. There is therefore an opening of language towards the lived experience of the life-world.

The constitution of meaning, as it arises out of human interaction, and the process of reference can be understood as the result of the cooperation of the

[24] This difference has been expressed by Wheelwright (1954) by introducing *epiphor* and *diaphor* as two modes of metaphor.

[25] Turbayne (1962) characterizes this fixation by saying that we are "victims of metaphor".

[26] This focussing corresponds to the existential *Augen-blick* (Heidegger, 1927) and forms the basis of mythical understanding. Cassirer (1925) has shown that there are common roots in language and myth. In contrast to the rationalistic view that sees a system as a hierarchical structure of isolated entities and in which understanding is guided by the principle of subsumption, in both language and myth understanding is guided by the "catching" of experience-Gestalts. The meaning of an object is not given according to its position in the hierarchy of categories but it is the residuum of a lived experience, of which it is a part. Eliade (1949) has named this "mythical" meaning of an object *hierophany*, i.e. each object can become sacred by being embedded in a religious context and being viewed thus from a new perspective. By analogy, the semiotic meaning of an object could be called "semiophany": by being embedded in a semiotic system it aquires an associative meaning. (For an investigation into this mythical form of understanding see also Leach, 1976 and Levi-Strauss, 1956).

[27] The two levels are therefore not two different layers but form together a self-referential system, i.e. language is *verbal behavior*.

metonymic sedimentation, *semiotic categorization* and *intra- and intersystematic*[28] *metaphorical connection* processes, instead of postulating it as an extrasemiotic principle. Out of the use of language the metonymic sedimentation process determines the interconnectedness in the metonymic topos. The intrasystematic metaphorical connection constitutes an organic whole out of the fix situations that are the results of the semiotic categorization process. Finally the intersystematic metaphorical connection opens the referential field, as the structure of intersubjective invariants. These processes do not take place in isolation, so that they cooperate only through the exchange of mutual inputs and outputs, but form a synergetic complex, i.e. meaning can only be understood as a dynamic process that arises out of the interaction between them. Metonymic sedimentation builds an extraction out of the unreflected use of language. The intersystematic introduction of a metaphor produces a tension, since it questions the established "reality" by showing the way to new modes of analogical interrelations and triggers a process of reorganization of the metaphoric topos that leads to a new referential domain imposing a modification of the use of language, requiring thus an adaptation of the metonymic sedimentation process, which in turn influences the metaphoric interrelation process, since analogy detection and introduction operate by scrutinizing the metonymic structure. The resulting superprocess is thus determined through the mutual coadaptation of the metonymic and metaphoric processes and forms a sort of *semiotical dialectics*: the referential domain that constitutes the reality of a community of language users is the product of the synthesis of the metonymic and metaphoric poles of language.

References

Arnauld, A., & Lancelot, C. (1972). *Grammaire générale et raisonnée (1660) suivie de La logique ou l'art de penser (1662)*. Genève: Slatkine reprints.

Barthes, R. (1985). *L'aventure sémiologique*. Paris: Seuil.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge: MIT Press.

Benveniste, E. (1974). *Problèmes de linguistique générale. vol. II.* Paris: Gallimard.

Budge, W. (1966). *Egyptian language*. London: Routledge & Kegan.

[28] The intersystematic analogy detection corresponds thereby to the intrasystematic in the sense of Husserl's *Einfühlung* (Husserl, 1977), i.e. perceiving the other as *alter ego*.

Cassirer, E. (1925, 1976). Sprache und Mythos. In: *Wesen und Wirkung des Symbolbegriffs.* Darmstadt: Wissenschaftliche Buchgesellschaft.
Chomsky, N. (1966). *Cartesian Linguistics.* New York: Harper & Row.
Churchward, A. (1913, 1978). *The Signs and Symbols of Primordial Man.* Westport, Connecticut: Greenwood.
Descartes, R. (1953). *Oeuvres et lettres.* Paris: Gallimard.
Eco, U. (1979). *The role of the reader.* Bloomington: Indiana University Press.
Eibl-Eibesfeldt, I. (1989). *Liebe und Hass.* München: Piper.
Eliade, M. (1949). *Traité d'histoire des religions.* Paris: Payot.
Gadamer, H. G. (1972). *Wahrheit und Methode.* Tübingen: Mohr.
Heidegger, M. (1927, 1979). *Sein und Zeit.* Tübingen: Max Niemeyer.
Greimas, A. J., & Fontanille, J. (1991). *Sémiotique des passions.* Paris: Seuil.
Hjelmslev, L. (1961). *Prolegomena to a theory of language.* Madison: University of Wisconsin.
Hubel, D. H. (1963). The visual cortex of the brain. *Scientific American, November,* 1963.
Humboldt, W. von. (1788, 1968). Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts. In: *Werke Band 7.* Berlin: de Gruyter.
Husserl, E. (1976). Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. *Husserliana Band III/1.* Den Haag: Martinus Nijhoff.
Husserl, E. (1952). Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. *Husserliana Band IV.* Den Haag: Martinus Nijhoff.
Husserl, E. (1977). *Cartesianische Meditationen.* Hamburg: Felix Meiner.
Jacobson, R. (1971). *Word and Language: Selected Writings II.* The Hague: Mouton.
Katz, J. (1966). *The Philosophy of Language.* New York: Harper & Row.
Katz, J. (1972). *Semantic theory.* New York: Harper & Row.
Köhler, R., & Altmann, G. (1993). Begriffsdynamik. In: Beckmann, F., & Heyer, G. (eds.), *Theorie und Praxis des Lexikons.* Berlin: de Gruyter, 173-190.
Koffka, K. (1935). *Principles of gestalt psychology.* New York: Jovanovich.
Lakoff, G. (1982). *Categories and cognitive models.* Berkeley Cognitive Science Colloquium.
Langacker, R. W. (1987). *Foundations of cognitive grammar.* Stanford: Stanford University Press.
Leach, E. (1976). *Culture and Communication.* Cambridge: Cambridge University Press.

Levin, Samuel R. (1977). *The Semantics of Metaphor.* Baltimore: John Hopkins University Press.
Lévi-Strauss, C. (1956). The structural study of myth. *Journal of American Folklore 78,* 428-444.
Lorenz, K. (1982). *Vergleichende Verhaltensforschung.* München: DTV.
Madison, G. B. (1982). *Understanding: a phenomenological-pragmatic analysis.* Westport, Connecticut: Greenwood Press.
Madison, G. B. (1988). *The hermeneutics of postmodernity.* Bloomington: Indiana University Press.
Maturana, H. (1978). Biology of language. In Miller, G. A., & Lenneberg, E. (Eds.), *Psychology and biology of language and thought.* New York: Academic Press.
Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition.* Dordrecht: Reidel.
Maturana, H., Lettvin, J. Y., McCulloch, W. S., & Pitts, W. H. (1960). Anatomy and physiology of vision in the frog. *Journal of General Physiology, 43,* 129-175.
Merleau-Ponty, M. (1942). *La structure du comportement.* Paris: P.U.F.
Merleau-Ponty, M. (1945). *La phénoménologie de la perception.* Paris: Gallimard.
Peterson, Ph. L. (1973). *Concepts and language.* The Hague: Mouton.
Ricoeur, P. (1975). *La métaphore vive.* Paris: Éditions du Seuil.
Ricoeur, P. (1976). *Interpretation theory.* Texas: The Texas Christian University Press.
Rosch, E. (1973). Natural categories. *Cognitive Psychology 4,* 328-250.
Rosch, E., & Lloyd, B. B. (eds.) (1978). *Cognition and Categorization.* Hillsdale, N. J.: Erlbaum.
Thiopoulos, C. (1990). Meaning metamorphosis in the semiotic topos. *Theoretical Linguistics 16,* 255-279.
Thiopoulos, C. (1992). Towards a logic of semiotic systems. *Revue Mathématique, Informatique et Sciences Humaines 17,* 49-60.
Turbayne, C. M. (1962). *The myth of metaphor.* New Haven and London: Yale University Press.
Vossler, K. (1960). *Geist und Kultur in der Sprache.* München: Dobbeck.
Welby, V. (1911, 1985). *Significs and language.* Amsterdam: Benjamins.
Wheelwright, P. (1954). *The burning fountain.* Bloomington: Indiana University Press.
Whorf, L. (1956). *Language, thought, and reality.* Cambridge: MIT Press.

Glottometrika 15, 1996, 38-61

The Levels of Order in Language

Luděk Hřebíček, Prague
Gabriel Altmann, Bochum

Abstract

Five possible types of organization in or among phenomena are considered in connection with the perspectives of the investigation of linguistic systems. Our main concern is directed to the epistemological role of the respective type of order.

Introduction

Let it be stressed at the beginning that each typology or classification of things or parts of things or their arrangement in some observed phenomena is derived from our ability to observe such phenomena and to analyze them in terms of our concepts. Thus classification is something put onto reality from "outside". This is clearly shown by the history of empirical sciences and is nowadays being emphasized both by physics and by psychology, the two poles of science.

This standpoint does not represent any subjectivism. It merely means that nature answers only questions asked by human investigators, it does not reveal or manifest "truth" by itself. The real existence of an organization in and among phenomena is not denied by this attitude. On the contrary, it is a sort of realism in treating knowledge and the progress in knowledge as a steadily developing interaction between the observed objects and observing subjects in the Piagetian sense (cf. Furth, 1972; Piaget, 1973). The element of subjectivism which is undeniable in this standpoint is confronted with and controlled by scientific criticism. In science, criticism is not a free flight of philosophical ideas; science is the only domain in which our errors are criticised in a *systematic* way (cf. Spinner, 1974: 106). The pre-Galilean critical discussion has been replaced in modern science by the requirement to formulate scientific knowledge in the form of testable theories.

Scientific knowledge is selective (cf. e.g. Toulmin, 1974) and the means of selection depends on our ability to discern in a part of reality some regularity or organization, as well as on personal preferences of the researchers which may be

affected by tradition, education, fashion, etc. With the increasing ability to find different types of arrangement in things and with our increasing interest in abstract theories such as the theory of systems, probability theory, etc., or mathematics in general, the observed space becomes larger.

This also concerns linguistics. Unfortunately, this science more than other human sciences is handicapped in its theoretical development by its applicational goals. Knowledge of natural languages is reduced by the intention to formulate norms of communication and to learn foreign languages. It is generally accepted by the majority of linguists that linguistic science formulates rules for statements (utterances) both in the mother and foreign languages, rules for the transformation of utterances from one language to another, rules for the correct writing of words, sentences, letters, novels, etc. Linguistic "theories" are mostly formulated in a philosophical way, i.e. in a non-theoretical way. Theoretical linguistic knowledge rarely surpassed the borders of the above-mentioned utilitarianism and free sophistication.

The outstanding position of mathematics among sciences consists in the ability to seek and define abstract structures and formulate appropriate consequences. The best way to systematize our knowledge is the way indicated by mathematics. Notions like "system", "structure", "process", etc. are not understandable without mathematical treatment; in their current usage they are merely intuitively interpreted.

Provisionally, the following types of organization can be observed in different systems at the contemporary level of knowledge:

1. Nominal order
2. Rank order
3. Sequential order
4. Self-regulative order
5. The order of chaos.

1. Nominal order

On this level the universe of discourse is divided into disjoint classes which obtain names. As a matter of fact, this is the first attempt to bring some order to the phenomena at all. Both in phylogeny and ontogeny as well as in the beginning of science, this is identical with concept formation by means of categorization, abstraction, self-regulation, self-organization, etc., all of which represent processes, not states. Of course, this step is necessary but neither sufficient nor final for Man or for science. It is a trivial statement that there are no classes in reality,

merely individuals. Of course, there are objects displaying some kinds of similarity, which is a criterion for class building and inclusion; but in every case this similarity is a function of human concepts. Taking different features, two "similar" things can end up in two different classes. Even the selection of individual objects for observation represents an observer`s encroachment upon the observed reality.

In the humanities, especially in those parts working with qualitative concepts, the name, the class, is something eternal, something mystical, giving us power over things named (cf. Carnap, 1969). And there are so many names and class concepts to be created that at last, their sum is called "theory".

Nominal categorization represents a kind of one-dimensional static system, system being considered as a set of constituents and their internal and external relations. In many cases it is merely an inventory defined functionally or through the determination of membership, as in set theory, e.g. word classes, distinctive features, language types, sentence types, and grammars as sets of rules, text classes, NP - VP, prefixes, suffixes, infixes, etc. Textbooks of modern linguistics are mostly and merely collections of concepts and classes.

Typologies of this kind can, of course, have different epistemological status. We can distinguish the following kinds:

(I) If nothing follows from the established classes, then this is a pure, restricted *nominal classification* with a low epistemic value. This is the case of e.g. Sapir's language typology (cf. Sapir, 1921). In an extreme case one can restrict oneself to two classes and obtain a binary classification. This limiting case functions well in computers but is connected with the greatest loss of information. Binarism is well known in linguistics; the attempts to surmount it usually fail because of lack of a theoretical approach.

The epistemic value of this typology does not change even if it is performed with the frequently complicated methods of numerical taxonomy. One always obtains merely a decomposition or dissection of the universe of discourse.

If we carry out this approach as a thought experiment thoroughly to a limit then we can observe several disadvantageous results:

(a) If the number of taxa is small, e.g. 10, then a decomposition of the field into classes does not bring a better view or whatever one expects.

(b) If the number of taxa is large, e.g. 6000, then the following results can emerge:

(i) If we decide to build few classes then in any class there will be so many taxa that the knowledge of membership does not bring any practical or epistemic advantage. This classification will be compulsorily fuzzy.

(ii) If we let the number of classes increase without limit, then the classification will not be lucid and will yield little information (or anything else), no matter whether it is presented in form of hierarchies or clusters. This means that a purely nominal classification has a practical aim only if the number of taxa and the number of classes is medium (whatever this means).

(c) In numerical taxonomy, mathematicians have developed about 600 classification methods and this number increases from year to year by some dozens (cf. Bock, 1974). However, different methods yield different results for the same universe of discourse. Well, what is the epistemic value of a classification which is falsified by any other method? It is needless to mention that these methods can neatly classify even a number of random points.

(d) The choice of properties serving as a starting point of the classification is an important criterion for class building. Semantic properties generate word classes other than syntactic ones; the "ability to fly" does not imply the membership in the class of "birds", and on the contrary, "no ability to fly" does not imply an exclusion from this class. If we build polythetic classes then point (b) becomes relevant. Taking many properties into consideration may stabilize the classification but at the same time it leads to the generation of many classes or of fuzzy classes or of many fuzzy classes. Thus concepts like "natural class", "kind", "species" or "genus" become irrelevant in linguistic classifications. This means practically that for any universe of discourse we can generate as many equivalent classifications as there are possibilities of placing r balls in n urns, the balls being different and n taking all discrete values from 1 to r (cf. Feller, 1962:36ff.) or at least as many as there are classification methods. In the case where some of the properties are predictive, i.e. if they imply or exclude the presence of other properties used in the classification, the number of classes can be reduced but the epistemological status of the classification would not change because it refers merely to itself. Moreover, presence or absence are merely coarse binary reductions of information preferred in structuralism, in language typology, in "theories" of grammar, in dialectology, in sociolinguistics, etc., so that the prediction is very coarse and very limited.

(II) If from the classification we can infer something that has not been put into it, then we have at least a *predictive classification* which can be useful for practical purposes. It can even initiate the setting up of hypotheses since it points to some interrelations. Of this sort is e.g. Kretschmer's typology of body forms or Skalička's typology of languages containing the beginning of an explanative (see below) typology (cf. Skalička, 1966, 1979). However, predictivity is not the aim of classification and usually it is not present in it (cf. Bunge, 1983:331).

(III) *Polar typology* takes an intermediate position in this framework because here one can place an object between two poles. However, it requires an ordering criterion, uses comparative concepts and thus results from another type of order (see chapter 2). W.v. Humboldt (1963) showed a tendency towards this type of order.

(IV) An ideal typology - that can even be nominal - is, according to Hempel (1965), a classification following from a theory; i.e. the classes, the properties and the behavior of class members are consequences of the operation of laws. A classification of this kind can be called *explanatory* because the formation of classes can be explained by hidden mechanisms. A typology of this kind has a high epistemic value. The best known cases come from physics, chemistry and biology. In linguistics they are not present; anyway, since the introduction of the principle of self-regulation in language (cf. Köhler, 1986), they would be possible here.

The adoption of nominal order in the investigation of languages is important above all in applied linguistics (e.g. grammar); it characterizes a beginning phase of any scientific discipline. Except for some cases of explanatory typology in other sciences, it is a purely inductive epistemological procedure. Deductive building of nominal order (typologies, classification) is possible only in more mature sciences.

Though there are no classes in reality, we need them both for practical and for scientific purposes. In practical life they can facilitate concept formation, prediction, and orientation in reality and communication; however, in science they should constitute areas of validity of laws. Laws expressed by general statements hold for an entire universe of discourse (e.g. for all languages); their establishment can lead not only to the building of classes but at the same time laws are the only criterion helping us to decide whether we attained a certain small approximation to reality, i.e. they are the only justification for a given classification. Consequently, while all other typologies are important for some momentary practical purpose, only an explanatory typology obtained with the aid of scientific laws has a higher epistemic value and importance for a larger set of practical purposes.

Let us quote M. Bunge's opinion on this problem, which cannot be said better than in his own words (Bunge, 1983:330):

> "Now, mere perception is bound to lead to superficial partitions. Deep partitions call for hypotheses and, in particular, law statements, i.e. formulas about patterns. And, because law statements belong (by definition) to theories, if we want deep classifications we need theories, the deeper the better. Good examples of the power of theory to inspire deep classifications are contemporary biological systematics (based on the theory of evolution), the periodic table of the elements (based on the

> atomic theory), the classification of hadrons based on the quark model, and the classification of materials based on their constitutive relations or specific laws.
>
> Classing and theorizing are then mutually complementary activities."

2. Rank order

The individuals of reality are not ranked by themselves. Ranking is a procedure consisting in the creation of a comparative concept and of rules of its application to the individuals of the universe of discourse. Again, it can be made for practical purposes like description or classification or sorting of eggs into ordered weight-classes, which is identical with an arbitrary division of a continuous variable into disjoint intervals. This technique is usual not only in practical life; in science one uses it in order to obtain approximations or in order to test continuous functions on discrete data.

Thus one can order languages according to their degree of syntheticity and ascribe ranks to the degrees, or one can build intervals and ascribe them ranks. One can rank words according to their frequency of occurrence, polysemy, descriptivity, imagery, degree of abstractness, generality, etc.

Even though we know that things are not ordered in reality, there is a possibility that the mental activity of ranking itself abides by certain laws. Thus if we postulate the lawlikeliness of our ranking activity - and there is no other possibility - we are authorized to search for the pertinent laws. The knowledge of laws would allow us to interfere in nominal orders and to use rank-order laws as criteria of nominal class building.

This circumstance can easily be exemplified. Let us assume that in a certain language we have established nine word classes following the Latin prototype. Disregarding the practical usefulness for descriptive grammar, what is the criterion that this nominal class building is also theoretically relevant (or prolific or "correct")? Let us proceed from the fact that word classes arose through differentiation or diversification of some original amorphous "class". This diversification must in any case follow some laws embedded in language self-regulation. These laws control both the disintegration of language institutions and the disintegration of communication in general. It is for example impossible that in a language as many word classes arise as there are words. Consequently, there must exist at least a law restricting class formation. Thus, the establishment of classes in language is compelling because it optimizes memory effort, helps to create redundancy, enables us to formulate grammatical rules, etc.

Now, if from a class a new class emerges, then it is formed proportionally to

the "mother"-class, i.e. its "measure" (e.g. frequency of occurrence, size, etc.) is proportional to that of the "mother"-class. Without regard to the changes in language, the rank order of measures must abide by these laws, i.e. the feedback remains preserved even if some parameters or the proportionality itself change. In order to set up hypotheses of this kind, we put $y_x \sim y_{x-1}$ (where y_x is the measure of the class at rank x and "~" means "is proportional to"), define the proportionality $g(x)$ as a function of x and determine $g(x)$ with regard to all linguistic boundary conditions so that $0 < g(x) < 1$ (cf. e.g. Altmann, 1993). If it is no ranking hypothesis, then the range of $g(x)$ can be different.

The best known laws (or at least formulas) derived in this way are e.g. Zipf's law (cf. Zipf, 1949), the Zipf-Mandelbrot law (cf. Guiter & Arapov, 1982; Orlov, Boroda & Nadarejšvili, 1982), Martin's law (cf. Sambor & Hammerl, 1991), Menzerath's law (cf. Altmann & Schwibbe, 1989), some diversification laws (cf. Rothe, 1991) and even development laws (cf. Altmann, Buttlar, Rott & Strauß, 1983).

We assume that besides the intuitive, practical rank orders (e.g. "John is stronger than myself; Peter is stronger than John; Peter is - thank God - my friend") there are laws which not only order our knowledge during the development of our cognition but even force us to create and use new classes in language in a very special way. These laws also operate in the building of hierarchies and prevent hierarchies from getting too large, emergence of too many classes, classes containing too many elements, etc.

Cognition and ordering of knowledge are no doubt frequently associated with ranking so that capturing it would have a high epistemic value if it were connected with the formulation of laws.

It is easy to obtain an empirical (ranking) result if we consider only one isolated property. Problems arise immediately if we try to rank the objects according to several variables at once. However, even for these purposes mathematicians developed a number of multidimensional scaling methods that enable us to transform, reduce, rotate, etc. the data in such a way that we obtain a nice graphical representation for practical purposes. The epistemic value of these methods can be raised if they are able to stimulate us to set up hypotheses, otherwise they are merely means of description based on some ambiguous semantic conventions. Multidimensional scaling does not necessarily lead to a discrete order - it can also be continuous - but this does not mean an epistemic improvement.

The same approach to multiple ranking as above can lead to lawlike statements if we start with a feedback hypothesis and e.g. for a three-dimensional rank order, we set up the difference equation $y_{x,u,w} = f(x, u, w)y_{x-1,u,w} + g(x, u, w)y_{x,u-1,w} + h(x, u, w)y_{x,u,w-1}$. If the functions f, g and h can be determined on the basis of

linguistic boundary conditions or subsidiary conditions, we can obtain in this way epistemically valuable statements that can become part of a theory.

As a matter of fact, linguistic entities can always be *considered* as ranked (or otherwise ordered) in n dimensions; an isolated unidimensional ranking represents merely a marginal rank order which can, nevertheless, be examined separately. This is, unfortunately, mostly connected with problems, because the individual properties of linguistic entities are never fully independent of one another. A one-dimensional ranking of linguistic entities is always a strong simplification, but at the beginning some simplifications are necessary.

At this point we are unavoidably confronted with the fact that parts and properties of language are tuned to each other and form a system. Since systems in which "there is something going on" can survive only if they develop a kind of self-regulation, the step to a higher order is compelling.

3. Sequential order

While the concept of rank order can be characterized as a phenomenon based mostly on the idea of intensity, sequential order exposes the notion of process. There are two kinds of sequential order in language, lying at two different stages.

The first stage lies in the mental domain. From the multidimensional space of our knowledge, thought, imagination, emotions, memory, etc., we try to select what we want to transmit. This knowledge cannot be transmitted in its diffuse multidimensional form, in which it is present in the mind as a kind of Hjelmslevian substance; we must decide in which order of succession the individual parts should be transmitted. Text theoreticians call this process *linearization* (cf. Hermann & Hoppe-Graff, 1988). This process, evidently, does not have infinite degrees of freedom, since there are a number of conventions, genres, styles, etc., controlling the arrangement of information. In a scientific book the information is unfolded differently from a poem; we know in broad outline the sequential unfolding of a drama, we know much about the unfolding of a narrative (cf. Wildgen, 1993), etc. But we know very little about the mechanisms controlling the linearization itself. It is not sufficent to call a particular kind of linearization "Style X" since an ascription of a result of a linearization to a special class is an epistemological minimum, as was shown above.

The mental linearization can be displayed as in Figure 1. As can be seen, knowledge unit 1 is also repeated in the third message item and potentially also in other parts of the message. Thus repetitions give rise to numerous patterns that can be modelled mathematically (cf. Altmann, 1988), and they are valid for the text as a whole (cf. Hřebíček, 1991, 1992).

The second kind of sequential ordering is realized in coding, i.e. within the sentence and between sentences (= within the text). The former is the main domain of grammar and is unfortunately stamped by its applicational aim. When coding a particular fragment of the given information, we use particular strategies (grammars) guaranteeing that the given part of the information will be transmitted with sufficient redundancy in order to be understood. For example the statement "railway-station you can where me tell" can be well understood in a special situation even if the hearer needs a longer decoding time because of reduced redundancy. Using all necessary grammatical rules, the contents of this statement would not change, but the redundancy would increase. Thus grammatical rules are (important) redundancy building means, but not theoretical entities.

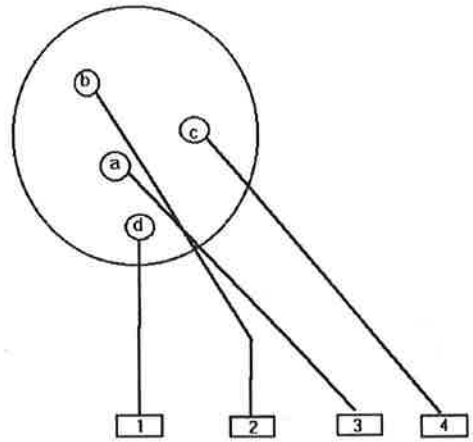


Fig. 1. Mental linearization

From the theoretical point of view, the concrete form of a grammar is not important; knowledge of it merely enables us to describe it or to use it effectively. Any intuitive or explicit knowledge of grammar (= sequential order within the sentence) does not yield any deeper insight into the mechanisms controlling it. Whether we represent such a sequence as SOV or S → NP + VP or draw dependential, stratificational, functionalistic, etc. pictures or capture them by means of formulas, we do not gain much from the epistemological point of view. Neither can we expect that from grammatical research theoretical approaches will arise, since "theories" of grammar merely touch the surface of language.

In order to illustrate the problem of theoretical capturing of the sequential order, let us use an example from physics. If at the bottom of a dish the temperature of water is slightly raised, the molecules transmit the heat to their neighbours. But if we raise the temperature steadily this means of heat transmission does not do - the molecules themselves begin to move and get into a process creating at the surface the so-called Bénard cells, i.e. small hexagonal cells in the center of which the warm molecules come to the surface and at the edges of which they go down again (cf. Haken & Wunderlin, 1991:7-8). As can be expected, this is the most effective means of heat exchange for water molecules - though the molecules cannot know it. At a critical point where the instability through increase of heat cannot be eliminated any more, the molecules organize themselves and create new effective heat conduction patterns.

There is a strong analogy between this case and the stored knowledge that must be suddenly transmitted in a linear way though it is not organized linearly in the mind. The first linearization is independent of language, but nevertheless it displays very characteristic patterns (analogous to Bénard cells) that are not part of this knowledge. It is a kind of linear self-organization of knowledge into information. The "within-sentence"-part of the second linearization (= grammatical linearization) organizes parts of the information itself by means of rules in order to give it sufficient redundancy and to convey a portion of the information which, for the sake of efficiency, is not expressed lexically. It depends on language and is not necessarily lawlike. Rules are part of this information. The "between-sentence" or "within-text" part of the second linearization (= textual linearization) also creates patterns analogous to Bénard cells. It organizes the language material in a lawlike manner which is neither part of the knowledge, nor part of the information, nor part of the rules, nor part of the individual entities in the sequence. Here, the character of the language material is merely a boundary condition. Thus we have four kinds of sequential order resulting from linearization:

(i) Mental ordering of knowledge to be transmitted, which is independent of language;
(ii) Coding by means of rules in order to transmit a message effectively, which is particular to every language (= grammatical ordering);
(iii) Regular or spontaneous (random or chaotic) rise of sequential patterns, which is common to all languages.
(iv) Conscious creation of sequential patterns, e.g. in poetry.

The epistemic value of rules - which are the only contents of grammatical research in individual languages - is very problematic. No string of substitutions generating a sentence - a string having the form of a sequence of rules - can be confirmed by

a native speaker (informant) who ratifies the generated sentence. Two or more mutually differing strings producing one and the same sentence can be confirmed in this way, which is evidently unacceptable.

Let us suppose that two subsequent sentences are generated by a set of rules, which is a purely hypothetical supposition, as no such set was formulated even only for two sentences. What does the statement pronounced by the informant mean then? Does his/her ratification or rejection also concern the correctness of the collocation of the sentences?

Even an individual sentence cannot be *explained* with the help of substitutional rules. The process of generation of each observed sentence is a result of stochastic proceedings which cannot be expressed in the form of substitutional rules. Occam's razor cannot be applied to these rules as they are formulated in advance and the decision concerning a remnant in the rules is always formulated in relation to the other rules but not to laws arranging language units into levels and sequences. The "rule theory" cannot introduce into its structure the principle of rejection. Thus the structures generated are nothing but some abstract entities. Of course, they have relations to language reality - however, these relations must remain inexplicit.

The ostentatious refusal of quantitative approaches to language, so faithfully approved by his followers, has been a tragic mistake of Noam Chomsky. His "theory" remains without connections with that epistemology which expresses the principles of other empirical sciences. We can say in short that the generative rules, even when they generate something that is proclaimed to be "correct", cannot be rejected as their correctness was inserted into the formulation of the rules. The rules themselves are the criterion of correctness.

Rules represent a non-cognitive tool of the language users. However, any linguist should know the reasons for the usage better than the normal users of a language. Linguists must produce explanations and not rules of correctness. The requirement for sciences to formulate testable theories is not a challenge to create sets of rules.

Let us introduce an analogy taken from music, for which we must give the following explication: The intonation values of notes are defined by musical clefs. On a violin as well on a viola it can be played in fingerings classified as 'first position', 'second position', etc. A position one higher is played with the first finger in that place on the instrument where in the next lower position the second finger is used. Suppose there is a violinist who is not able to read the notes for the viola. However, he can immediately play a composition for viola on a viola if he abides by the following rule:

Play the notes for viola and on the viola in the first position as you would play the same notes on a violin (in treble clef) in the third position.

The violinist can immediately play on the viola the melody in a *correct* way, but from the notes he cannot ascertain their intonational values. His cognition remains empty.

Sequences are directly observable in texts - they are, so to say, surface phenomena. But there is a huge difference between the sequential order controlling the concatenation or repetition of properties of language entities (in a nomological way) and the rule-governed order of a particular grammar. Traditional linguistics is primarily interested in the latter phenomenon because it seems to be easily captured and described, no hypotheses about background mechanisms need to be set up, no theory seems to be necessary. Usually, a set of rules controlling the grammatical correctness of a sentence in a particular language is called "theory". Now, the rules prescribe that a certain slot in the sentence must (should, can) be filled by a set of entities, e.g. words. These entities in turn form classes, so that rule-governed order leads to the establishment of a nominal order in a particular language. Thus the establishment of rule-governed order (syntax, syntagmatics) and nominal order (paradigmatics), if restricted to a particular language, are the most elementary linguistic activities with minimal epistemic impact.

Epistemically more prolific than a description of grammar would be the investigation of *laws* of linearization which *must be common to all languages.* A start has been made by the famous mathematician Markov who used a text of Pushkin and discovered what one calls today Markov chains, but several examinations have shown that they could be adequate (if at all) in a very restricted domain of language. Unfortunately, in sequential concatenation of language units there exist not only dependencies on immediate or non-immediate predecessors (to a different extent), but also on the successors. For example the German article (*der*, *die*, *das*) depends on the successor-noun; in coarticulation the variant of a consonant depends on the next vowel even if it lies five places ahead; concord is a typical non-Markovian phenomenon, etc. Moreover, certain properties of language units depend even on the constructs which they are parts of. Thus language sequences are much more complicated than simple Markov-chains.

Nevertheless, we cannot commit an error if we assume that processes are operating here which control the *properties* of all units in the sequence. These processes generate *sequences of properties* of, say, phonemes, syllables, morphemes, words, syntagms, sentences, etc., which are not even taken into account in usual grammars. Moreover, these processes are, perhaps, linked with one another in different control cycles at different levels. It would be just the investigation of these processes that would have a high epistemic value. It could be done by applying the theory of time series, theory of runs (cf. Grotjahn, 1980), theory of repetitions (cf. Altmann, 1988), theory of distance patterns (cf. Zörnig,

1984 a,b, 1987; Orlov, Boroda & Nadarejšvili, 1982), the theory of fractals (cf. e.g. Barnsley, 1988), etc.

From this rather negative attitude against the role of individual grammars in the theory of language it must not be concluded that grammatical processes are absolutely "stray" and have nothing to do with the nomological mechanisms of language. There have been different approaches to the "explanation" of grammatical phenomena both from the universalistic and the typological points of view (cf. e.g. Givón, 1979; Hawkins, 1988). Though they do not represent real explanations, being merely interpretations of possibilities, all of them are more or less reasonable, saying that rules arise under special conditions and restrictions. However, if we know a sufficient condition, we do not have an explanation since the same linguistic phenomenon can arise under different conditions, as is well known in linguistics. If we know the necessary condition, we do not have a genuine explanation, since a special condition in language can give rise to different phenomena (functional equivalents) and we would be forced to find out why it leads exactly to the given phenomenon, i.e. to look for further conditions. Thus, in the end, in order to explain a grammatical phenomenon we must put it into a nomological net (cf. Salmon, 1984) or into a self-regulation scheme (cf. e.g. Köhler, 1990), which cannot be done on the basis of the knowledge of rules of only one language.

4. Self-regulation and self-organization

Every thing abides by laws (Bunge, 1977:17). Any aspect of any thing can be captured by a set of laws. In other sciences this has been generally known for a long time. At first a law was considered as a deterministic dependence with firm causality but later on this rigidity was relaxed (cf. e.g. Bunge, 1959). In linguistics one already used this concept in the past century (sound laws) but these statements were merely empirical generalizations with a very restricted domain of validity. The interest in language rules, especially in the USA, suppressed any initiative in theoretical investigations, though this direction has been advanced here since the 1920's by G.K. Zipf (cf. e.g. 1929, 1932, 1935, 1949).

If we search for lawlike order we must take into consideration hidden mechanisms generating observable phenomena. These mechanisms are joined in control cycles or create order spontaneously and have (in language) a stochastic character. If we consider merely an isolated dependence on the OP-level, we are inclined to speak about causality, but if our horizon widens and we are able to capture a control cycle in its totality, or to discover a self-organizing mechanism, the role of causality in language loses much of its importance (cf. Riedl, 1982). The situation

can be seen in Figure 2. Here properties A, B, C, and D are joined in a control cycle that has been derived deductively or set up axiomatically, and that must be tested on empirical interrelations between a, b, c, and d in individual languages.

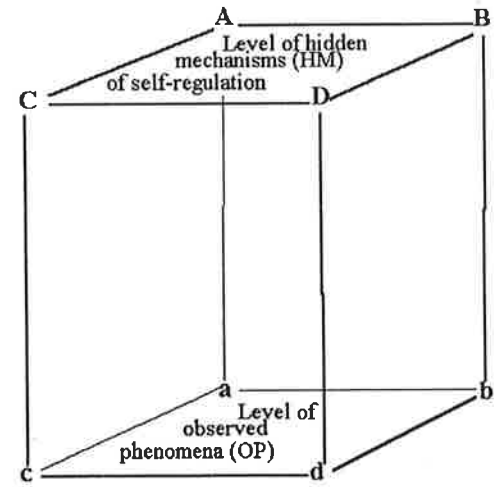


Fig. 2. Observed phenomena and hidden mechanisms

In linguistics one often confuses laws, which are properties of the reality, with law statements, which are conceptual constructions. In philosophy, one distinguishes them sometimes as law_1 and law_2 or "objective law" and "scientific law". For definition and kinds of laws cf. e.g. Bunge (1961, 1967).

Properties of linguistic entities display five kinds of interrelations that can be relevant for a theory of language:

(1) The above-mentioned *sequential interrelations* of neighbouring entities. A part of them is of a grammatical nature having, however, no importance for a theory. Much more important are time series, oscillations and fractals behind which we assume the existence of self-organizing mechanisms. The speakers of a language care for the grammatical correctness of their sentences, but seldom or never for e.g. a regular pattern of sequences of sentence lengths. However, all properties of all linguistic entities organize themselves in the course of their realization on the time axis, even if "only" in a form described in fractal theory (cf. e.g. Mandelbrot, 1982 or Feder, 1988). Linguistics is at the threshold of this research.

(2) The relations between a property of an entity and its other properties. There is a relation e.g. between the polysemy and the length of a word, or between length and frequency of occurrence of a word. Relations of this kind belong to the *self-regulating* part of the language system. The properties are in a steady-state and belong to different control cycles. This discovery was made by G.K. Zipf (not by F. de Saussure) and developed by R. Köhler (1986, 1990, 1992).

(3) The *hierarchical relation* between a construct and its components through all levels of language. The best known relation of this kind is that between the respective sizes of constructs and components controlled by Menzerath's law. Here there is a hierarchical dependence: Size of an aggregate → length of sentences → length of clauses → length of words → length of syllables → duration of sounds (cf. e.g. Hřebíček, 1992; Altmann & Schwibbe, 1989, and the references therein). Many of them have been investigated, and all corroborated the hypothesis that "the longer a construct the shorter its components". This is a perfect analogy of the "allometric law" with a negative exponent. This order arises in texts spontaneously through self-organization. The part "word length → syllable length" is not spontaneous but developed and stabilized in the course of the history of language and displays a very constant character. It has, of course, different parameters in the lexicon and in texts where inflectional affixes can be attached to the word or where *ad hoc* compounds arise. The part "syllable length → sound duration" exists merely in speech.

(4) If we realize that a property is controlled in a text both sequentially and hierarchically, then it seems to be a wonder that there exists a further self-organisatory force controlling the *frequency distribution* of the given property in a text. Irrespective of the succession of the degrees of a property in text, e.g. succession of sentences of different length, irrespective of the (hierarchical) modification of the properties of entities at the lower levels, e.g. the clause lengths by sentence lengths and in turn word lengths by clause lengths, etc., in the completed text both sentence length and clause length (and all other properties) display very "respectable" frequency distributions that can be modelled starting from simple assumptions. Even *classes* of entities display "respectable" rank-order distributions.

Peculiarly enough, it is not possible to operate here with ready, "eternal" formulas. Rather we assume that in the course of text generation, which can be very variable, or in the course of language development, which is more steady, a property finds an attractor, converges to it, oscillates around it or moves chaotically within an interval in the neighbourhood of the attractor. Thus it is possible that in different languages or different texts we must set up different models for the behavior of the same property or we must take many subsidiary conditions into

account. The door to this part of self-organization in language is already open; a lot of research has been done hitherto.

(5) The properties of an entity can be *productive* themselves in two ways. (a) Either they evoke something in the entities at a higher level, e.g. long sentences make the text more difficult or (b) they "develop propensities", e.g. short stems or words are more frequently parts of compounds than long ones. This is also a kind of self-organization following rank-order or frequency distribution laws.

5. The order of chaos

The last type of organization we intend to present here concerns the chaotic way of organization. This type has its observable correlates in the phenomenon of turbulence and of free movement of particles in a volume, as well as those having indiscernible borders between their structural constituents. This type of organization was introduced into science for the first time by meteorologists.

It is the consequence of the long-standing education and training directed to language users that language and all its discernible items appear to be a system of sets of certain firmly postulated units. This fixedness looks like a pure and simple consequence of a rich collection of conventions. Language units - phonemes, morphemes, lexemes, etc. - are understood as semantically centred items standing somewhere among their allophones, allomorphs and interpreted lexemes, respectively. Higher units and structural parts of text, however, resist such intuitive handling grounded on the contrastive treatment of the *emic* and *etic* units. In such approaches a semantic invariant is postulated. This is not acceptable on two grounds:

1. This approach cannot be applied to larger units, for example to sentences and text paragraphs.

2. It cannot be applied to the semantic level of communication in a natural language; this level seems to be the highest linguistic level (or better: a cluster of the highest linguistic levels); any linguist operating on the semantic level loses the opportunity of transition from the sphere of the language expression to the spaces of meaning. Thus the principle of transition is substituted by movement in circles which cannot offer an instrument for explanatory activity.

When the properties of language units are treated as parameters of a space structured in a controllable way, then phonemes, morphemes, lexemes or other units can be considered as attractors occurring in a part of the measurable space of a given language where the respective *etic* counterparts occur. When a part of any language system changes in a way, the change reveals itself as a loss of stability and of the self-regulating qualities in the respective piece of the language space.

The respective units are then reorganized into a new attractor or attractors and the system passes from one attractor to another one. This change brings about a new shape of structuring of the communication means, i.e. it represents a self-organization.

The same idea can be applied when the relation between language and speaker is supposed. This relation is evidently a generator of the above-mentioned changes of attractors in the language system. Each language user represents a certain oscillation around the position of the respective language in the abstract communication space. An idiolect in relation to another one reveals a chaotic type of organization. The same is valid for dialects, each taken as one item; their changes are also based on the self-regulative principles and are applied in the horizontal and vertical stratification of languages.

A sequence of arbitrary units appears to be chaotic if certain of their properties - for example, their length - are observed as they follow one after another in a text. E.g. the sequences of lengths of words evidently form fractals. The same holds for other properties, as, for example, in the case of the verb-adjective ratio forming a wild fractal, or the type-token ratio forming a relatively simple fractal approximated by the curve $y = x^b$ which is an attractor (cf. Köhler & Galle, 1993). The style of a text can doubtlessly be modelled as a fractal structure encompassing different attractors.

Generally, many language units are usually treated as precisely ordered items; the other units, however, appear to be phenomena resisting a simple-minded rational treatment. Sciences examine many systems in which determination is combined with chaos, and this property, once treated as an unacceptable paradox, nowadays is found to represent mutually compatible qualities.

The human ability to produce and discern different meanings is not in a parallel position to the order of units in language constructs. The transition from the level of language expressions to the level of meanings is a passing from light to darkness. The types of order applied in linguistics are devices for casting light upon the order of meanings. In linguistics the opposite direction of transition is usual, namely that going from semantics to the language expressions; however, it is based on an unacceptable supposition that the complete knowledge of meanings and semantic systems is at our disposal. Natural languages represent one, and possibly the most important, look-through into the system of meanings and concepts. The transition from the intuitive ideas concerning meanings to the principles of the language arrangements is useful for practical purposes (e.g. for language teaching), but not for the explanation of language systems and construction of linguistic theories.

The point of transition from language expressions to meanings can be characterized as a turning point placed between determinism and chaos. Nonetheless, each language unit, regardless of its appurtenance to a certain linguistic level, represents such a turning point. In the fractal theory such a point is called *bifurcation*. And any larger language construct, such as text, represents a chaotic pattern of bifurcations on different language levels. The Cartesian approach, which consists in itemization of a complex phenomenon into smaller parts appearing to be more lucid, is the actual approach of classical linguistics. Thus more complicated parts of the language systems remain unexplained and the limit of the sentence is not surpassed by linguistics oriented in this way.

The theory of fractals by B. Mandelbrot (1977) is formulated as a new type of geometry. Nevertheless, the theory has already been used many times for explanations of complicated types of systems occurring in abstract spaces. The main principle of this theory can be characterized as seeking transformations which are invariant in relation to certain types of changes; these changes can be exemplified by the change of the measure of observation. Any branch of a tree can be taken as a bearer of a branch or branches, so that the whole tree itself resembles a branch. This is a kind of iteration characterizing the whole system. Another often quoted example concerns the length of a coast: its length depends on the length of a stick used for the measurement; as the length of the stick approaches zero, the length of the coast becomes infinite. At the same time, one obtains similar pictures of the coast regardless of the distances from which it is observed.

In linguistics there are known problems consisting in seeking similarities between units of different levels, e.g. the similarity between the morphological structure of a verb and the syntactic structure of a sentence. Such attempts are based on a vague concept of structure and can hardly lead to a refutable theory. The notion of self-similarity is rationally constructed and provides a clearer insight into the chaotic parts of the language subsystems. It must be stressed that quantitative linguistics reached this methodological level independently of the development in geometrical inquiry. On the other hand, it must be stressed that mathematical theory represents a rich source of methods applicable to the chaotic parts of the language systems.

Menzerath's law (cf. Altmann, 1980; Altmann & Schwibbe, 1988) was derived and formulated as a language law characterizing the dynamics of language levels. Its formula is

$$(1) \quad y = Ax^b$$

with x denoting a language construct, y the respective constituent, A equalling y with $x = 1$, and b a constant. The law states that the longer the language construct the shorter its constituent (on average) ("longer" and "shorter" the terms "more complicated" and "less complicated" or similar pairs can be understood). The law

has been confirmed for different levels. It is valid also for a supra-sentence level with units called *text aggregates* containing all sentences of a given text, in which a given (semantically interpreted) lexical unit occurs.

The parameter b expresses the complexity of the system of two respective levels, i.e. the level of the supposed construct and its constituents. Let it be stressed that not only neighbouring levels - for example, phonemes and morphemes - are subordinated to the law, since it is possible that between two known levels there exist also some hitherto unknown linguistic level or levels.

In the fractal theory a parameter analogical to b in (1) always obtains a positive value, while according to Menzerath's law b must be negative. The analogical parameter in fractal theory represents an expression for the complexity of the respective system or its dimension. Parameter b in (1) can be transposed into such a measure. In equation (1) the relation is exposed from a construct and goes to constituents. In fractal theory the consideration starts with a constituent and leads to its construct. When in (1) the function and its argument are interchanged, so that x is taken as the function of argument y, from

$$(2)\quad y = Ax^{-b}$$

the expression

$$(3)\quad x = \left(\frac{A}{y}\right)^{1/b}$$

can be obtained where b and the entire power are positive. It can be deduced that b is the inverted value of the self-similarity dimension in the respective language subsystem.

Instead of (3) the following equation can be obtained from (1):

$$(4)\quad b = \left(\frac{A}{y}\right)^{1/\ln x}.$$

When b is understood as a constant - and this is the case of (1) - it appears to be quite unrealistic that the right-hand side of (4) represents a constant expression. On the contrary, it is rational to suppose that this structure is a variable and is ordered in a way corresponding to Menzerath's law. Thus we prove the assumption that Menzerath's law represents the iterator of the language system. In accordance with this assumption, from (4) the following expression is to be deduced:

$$(5)\quad Ax^{b} = \left(\frac{A}{y}\right)^{1/\ln x}.$$

Then

$$(6)\quad y = Ax^{-(\ln A + b \ln x)}.$$

This equation has as its perfect analogy the equation expressing the so-called Zipf-Alekseev function (or distribution, after norming):

$$(7)\quad f_x = f_1 x^{-(a + b \ln x)}$$

with the following correspondences between the quantities: $f_x \equiv y$, $f_1 \equiv A$, $\ln A \equiv a$. Formula (7) was derived in a slightly different way in Hřebíček (1994) and confirmed by the observed values of the distribution of word associations (cf. Altmann, 1992). Thus it also seems to be confirmed that word associations are items which can be handled as 'potential texts' or 'aggregates' *sui generis*.

On the other hand, text aggregates were found in different texts of several languages (see Hřebíček, 1989, 1992, and Schwarz, 1992) and confirmed as fractal structures subordinated to Menzerath's law. Thus the Menzerathian structure reveals itself to be a language autoiterator, an iterator applied to itself in the way indicated in (5) and demonstrated in Figure 3.

The supposed iteration encompasses the gradual structure:

$$y^{y^{y\ldots}}$$

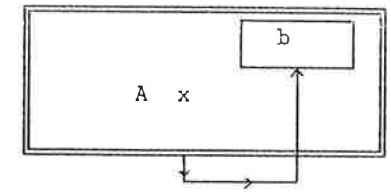


Fig. 3 The scheme of iteration contained in equation (5)

where y indicates formants on different levels. These relations require further investigation and a full development of the language fractal theory. Generally, A appears to be a point attractor to which different b's are directed, as is evident from Figure 4, where the courses of different y's are presented for b = {-0.1, -0.9, -1.5} with fixed A = 20.

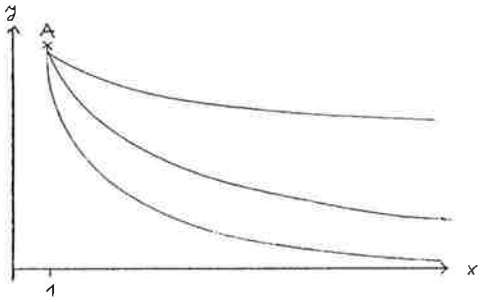


Fig. 4. Three different curves y with a fixed value of A and three different b's

Now it seems to be likely that between the sphere of the language expression and the language meaning there is not only one point or space of bifurcation. This transition occurs between each language level and semantics, and the fractal structure of language is itself a fractal.

Chaos can be treated as a type of dynamic system in which no point of its trajectory is visited twice by the system. Nevertheless, in the language space there are sections where the points of the trajectory are close to each other. These parts are called attractors, and the contrastive items are repellers. It should be taken into consideration that natural languages in many respects appear to be systems with dissipation.

The application of the paradigms of chaos opens new prospects for linguistics.

References

Altmann, G. (1988). *Wiederholungen in Texten.* Bochum: Brockmeyer.
Altmann, G. (1992). Two models for word association data. *Glottometrika 13,* 105-120.
Altmann, G. (1993). Phoneme counts. *Glottometrika 14,* 54-68.

Altmann, G., Buttlar, H.v., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical Linguistics.* Bochum: Brockmeyer, 104-115.
Altmann, G., & Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.
Barnsley, M. (1988). *Fractals everywhere.* Boston: Academic Press.
Bock, H.H. (1974). *Automatische Klassifikation.* Göttingen: Vandenhoeck & Ruprecht.
Bunge, M. (1959). *Causality.* Cambridge, Mass.: Harvard University Press.
Bunge, M. (1961). Kinds and criteria of scientific laws. *Philosophy of Science 28,* 260-281.
Bunge, M. (1967). *Scientific Research.* Berlin: Springer.
Bunge, M. (1977). *Treatise on basic philosophy. Vol. 3. Ontology I: The furniture of the world.* Dordrecht: Reidel.
Bunge, M. (1983). *Treatise on basic philosophy. Vol. 5. Epistemology and methodology 1: Exploring the world.* Dordrecht: Reidel.
Carnap, R. (1969). *Einführung in die Philosophie der Naturwissenschaften.* München: Nymphenburger.
Feder, J. (1988). *Fractals.* New York: Plenum.
Feller, W. (1962). *An introduction to probability theory and its applications, Vol. 1.* New York: Wiley.
Furth, H.G. (1972). *Intelligenz und Erkennen. Die Grundlagen der genetischen Erkenntnistheorie Piagets.* Frankfurt: Suhrkamp.
Givón, T. (1979). *On understanding grammar.* New York: Academic Press.
Grotjahn, R. (1980). The theory of runs as an instrument for research in quantitative linguistics. *Glottometrika 2,* 11-43.
Guiter, H., & Arapov, M.V. (eds.) (1982). *Studies on Zipf's law.* Bochum: Brockmeyer.
Haken, H., & Wunderlin, A. (1991). *Die Selbststrukturierung der Materie. Synergetik in der unbelebten Welt.* Braunschweig: Vieweg.
Hawkins, J.A. (ed.) (1988). *Explaining language universals.* Oxford: Blackwell.
Hempel, C.G. (1965). Typological methods in the natural and the social sciences. In: Hempel, C.G., *Aspects of scientific explanation.* New York: The Free Press, 155-171.
Hermann, T., & Hoppe-Graff, S. (1988). Textproduktion. In: Mandl, H., & Spada, H. (eds.), *Wissenspsychologie.* München-Weinheim: Psychologie Verlags Union: 283-298.
Hřebíček, L. (1989). Menzerath-Altmann's law on the semantic level. *Glottometrika 11,* 47-56.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures.* Bochum: Brockmeyer.
Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., & Rieger, B. (eds.), *Contributions to Quantitative Linguistics.* Dordrecht-Boston-London: Kluwer, 36-41.
Hřebíček, L. (1994). Word associations and text. In this volume.
Humboldt, W.v. (1963). *Werke Bd. 3.* Darmstadt: Wiss. Buchgesellschaft, 144-756.
Köhler, R. (1986). *Zur sprachlichen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Köhler, R. (1990). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11,* 1-18.
Köhler, R. (1992). Elemente der synergetischen Linguistik. *Glottometrika 12,* 179-187.
Köhler, R., & Galle, M. (1993). Dynamic aspects of text characteristics. In: Hřebíček, L., & Altmann, G. (eds.), *Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag, 46-53.
Mandelbrot, B. (1982). *The fractal geometry of nature.* New York: Freeman.
Orlov, Ju.K., Boroda, M.G., & Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.
Piaget, J. (1973). *Einführung in die genetische Erkenntnistheorie.* Frankfurt: Suhrkamp.
Riedl, R. (1982). *Evolution und Erkenntnis.* München: Piper.
Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Rottmann.
Salmon, W.C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, N.Y.: Princeton U.P.
Sambor, J., & Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze.* Lüdenscheid: RAM.
Sapir, E. (1921). *Language.* New York: Harcourt, Brace & Co.
Schwarz, C. (1992). *Zur Verteilung von Aggregaten in Texten.* Bochum, Msc.
Skalička, V. (1966). K voprosu o tipologii. *Voprosy jazykoznanija 1966/4,* 157-163.
Skalička, V. (1979). *Typologische Studien.* Braunschweig: Vieweg.
Spinner, H. (1974). *Pluralismus als Erkenntnismodell.* Frankfurt: Suhrkamp.
Toulmin, S. (1974). Ist die Unterscheidung zwischen Normalwissenschaft und revolutionärer Wissenschaft stichhaltig? In: Lakatos, I., Musgrave, A. (eds.) *Kritik und Erkenntnisfortschritt.* Braunschweig: Vieweg, 39-47.

Wildgen, W. (1993). The distribution of imaginistic information in oral narratives. In: Hřebíček, L., & Altmann, G. (eds.), *Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag, 170-194.
Zipf, G.K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 11,* 1-95.
Zipf, G.K. (1932). Selected studies of the principle of relative frequency in language. Cambridge, Mass.: Harvard University Press.
Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin.
Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge: Addison-Wesley.
Zörnig, P. (1984a). The distribution of distances between like elements in a sequence I. *Glottometrika 6,* 1-15.
Zörnig, P. (1984b). The distribution of distances between like elements in a sequence II. *Glottometrika 7,* 1-14.
Zörnig, P. (1987). A theory of distance between like elements in a sequence. *Glottometrika 8,* 1-22.

Glottometrika 15, 1995, 62-76

"Language Forces" and Synergetic Modelling of Language Phenomena

G. Altmann, Bochum
R. Köhler, Trier

"Language Forces" and Order Parameters

The modelling of dynamical language phenomena as well as the formulation of language and text laws meet with the difficulty of finding appropriate concepts for describing the dependence of an entity on another one. In analogy to the corresponding physical concept, G.K. Zipf (1949) postulated the existence of "forces" in the dynamics of language. According to Zipf, there are two opposed forces, the force of the speaker and that of the hearer - or, the diversification and the unification force. On the basis of these concepts Zipf set up the first linguistic model with an explanatory claim, which gave a plausible picture of the interaction of linguistic processes in many different domains of analysis - even if the logical structure of his argumentation remained imperfect.

From a modern point of view, in particular within the framework of a systems theoretical approach to language, Zipf's "forces" can be reinterpreted as order parameters or other system requirements. In fact, many requirements (cf. Köhler, 1986) and processes in the dynamics of language can be grouped in pairs of opposing effects ("competitive" versus "cooperative" processes in synergetic linguistics) and assigned to the interests of the (idealised) speakers or hearers, respectively.

A number of the effects of these "forces", requirements, or processes can easily be observed in different domains of language. These effects consist of e.g., regularities, equilibration and steady state phenomena, trends, functional dependences, and frequency distributions. The most regular phenomena among them, i.e. those which can approximately be described by means of deterministic methods such as formal grammars, algebraic rules, and set theoretical structures, form the object of qualitative/formal linguistics. Both types of phenomena, the (almost) deterministic and the indeterministic/stochastic ones, demand explanation, i.e. the formulation of universal laws which can be used to deduce and predict the observed structures and processes.

Let us illustrate some spheres of operation of polarised processes (or Zipf's forces) by examples from different linguistic levels or subsystems:

(i) *Phonetics*. The speaker tends to utter the linguistic material with least effort. This behaviour causes abbreviation, assimilation, sandhi and other phenomena; the hearer, however, demands maximal phonetic distinctions, words of sufficient length etc. in order to decode the message with least effort. In synergetic linguistics, this instance of a pair of speaker/hearer forces corresponds to the requirement of minimal sign production effort, and of minimal decoding effort, respectively.

(ii) *Unit length optimisation*. The speaker shortens frequent morphs, words, etc., thereby maximising economy of effort. Now since abbreviation leads in the extreme case to a phonetic reduction of words to a minimum, and *eo ipso* to an excessive amount of homonyms, the hearer hinders the speaker in doing so or, as an equivalent alternative, forces him to develop new means of discrimination (tone, accent, composition etc.). These competitive processes result in an optimal length-frequency-distribution.

(iii) *Optimisation of inventories*. The inventories of units on all levels of linguistic analysis show similar dynamics. The corresponding processes result in the reduction or increase of the number of units (phonemes, morphemes, words, etc.) according to the speakers' (minimisation of encoding effort) or hearers' (minimisation of decoding effort) needs, respectively. At the same time, the functional relations change as a result of the unification and diversification processes. Lexical unification, as an example, corresponds to the speakers' tendency of avoiding encoding effort and results in the increase of polysemy in the lexicon, whereas lexical diversification, which corresponds to the hearers' need of minimising decoding effort, leads to reduction of polysemy. Similarly, the (idealised) point of view of the speakers' side prefers a small inventory of polyfunctional units on all levels, saving memory and selection effort. On the other side, from the hearers' point of view one-to-one relations between form and functions are preferred.

(iv) *Code optimisation*. Whenever new code is generated, i.e. by creating new words and constructs as well as by producing texts using the existing inventories (lexicon, grammar), from the speaker's point of view, an optimum would be reached if as much information as possible would be expressed by as few linguistic material as possible. The hearer, however, needs a minimum of redundancy. As a result, at all levels a certain information/redundancy ratio emerges as a compromise. Consequently, not a single linguistic inventory (syllable, morph, word, ... inventory) does exhaust its combinatorial potential: only a small proportion of e.g., the possible phoneme combinations constitutes valid syllables, words, etc. At the text level, we find a certain ratio and certain patterns of information/redundancy chunks (given/new, topic/comment, theme/rheme pairs) and the well-known

frequency and repetition spectra of units, such as the frequency distribution of words (of the type of Zipf's law) and the Type Token Ratio (TTR).

(v) *Variants*. We can find variants in all language units and patterns, continuously created by the speakers as a consequence of their economical behaviour, of their wish to express themselves originally, of mistakes, etc. The increasing variety of forms and rules is controlled and partially eliminated by the hearers' force of unification.

Besides these pairs of forces of the Zipf type there are, of course, other classes of factors which contribute, by means of corresponding processes, to the rise and the forming of language structures. There are opposing pairs of requirements or control variables which cannot be assigned to speaker or hearer interests, e.g. the competition of the requirements of context economy and context specifity (cf. Köhler, 1986:63f.), which reflect both speaker and hearer needs, and the competitive pair of stability and adaptation requirements (ibid.:150) on the highest control level of the system.

Other control variables don't come up merely in pairs but also in larger groups of competitive elements. Within this class there are the functional equivalents, which are - more or less, with respect to a particular requirement - exchangable structures, methods, or elements.

There are countless examples of this type of competition in language; one of them is the set of methods for coding new meanings, which consists of four elements: lexicon, syntax, morphology, and prosody.

Moreover, there is a general competition in systems with respect to hierarchical relations: subsystems have a drive for autonomy, whereas supersystems must force their subsystems to integrate and even cooperate. In synergetics (cf. Haken, 1978), the term "order parameter" has been introduced for control variables of a supersystem which "enslave" processes in subsystems.

Word length, for example, is the order parameter that enslaves the length of the words' syllables (cf. Altmann, 1980); sentence length is the order parameter that enslaves the length of the clauses (cf. Heups, 1983; Teupenhayn & Altmann, 1984).

Other requirements are opposed with respect to their direction but not with respect to processes or results, such as the specification and the despecification requirements, which are both functionally met e.g. by adding more specific resp. more general expressions to the lexicon and hence both have an increasing influence on the variable lexicon size - which means that they cooperate, in a technical sense. Many requirements cooperate with respect to their direction and their effect, as can be seen in the case of the process of minimising memory effort. This process is a result of both, hearer and speaker interests and its effects are, regardless of their role in communication, minimisation or reduction of inventories.

In all cases, the requirements - or changes of their weight - result in changes in the system and form one of the sources of the system's dynamics. The corresponding processes can be mathematically modelled by means of stochastic processes.

In most cases, the approaches are based on equilibrium assumptions, and their solutions yield functions or probability distributions, which can get the status of laws if sufficient theoretical and empirical corroboration exists.

There are several alternatives how to approach a process model. One is the construction of a suitable process with discrete or continuous time and find a solution for $t \to \infty$. In many cases, a direct approach is preferable: equating the rates at which an entity enters and leaves states/classes, or deriving functions or probability distributions from requirements and boundary conditions. There is no general reason for preferring one method over another.

While with stochastic processes we usually strive for a time-dependent explanation, in the latter cases we have a time independent processual explanation relating the operating forces to the changes in the "enslaved" quantities.

Discrete approach

A well-known case of this kind is the approach of Orlov (cf. Orlov, Boroda & Nadarejšvili, 1982), starting from the assumption that the flow of information in a text depends on its planned or intended length. If the author has decided on how long his text should become then he (unconsciously) organizes the repetition of words (and other entities) correspondingly, i.e. he abides by the regularity of the rank-frequency distribution of the Zipf-Mandelbrot law. Orlov's assumption is quite plausible, since the "intended length", which he calls "Zipf's Number", represents an order parameter in the synergetic sense, controlling one of the text processes, namely the process of word repetitions.

Laws, whether behavioral or "mental", are not conscious - one cannot decide to act in agreement or disagreement with them. They are properties of our language or information processing devices, of our articulatory or auditive subsystems, of the physical medium or consequences of communicative and other social needs.

Many of these laws exhibit a surprisingly stable nature - though stochastic, they show recurrent patterns even in unpredictable sequences of units (texts), which are endowed with an enourmous number of degrees of freedom. Many of them can be represented by means of probability distributions. In order to derive them one can go one of two ways:

(1) One can set up psycholinguistic hypotheses about the neurolinguistic,

psychological, aesthetic, communication theoretical etc. mechanisms in the brain or other subsystems connected with language or

(2) one can set up linguistic hypotheses about the manner of operation of Zipfian and other forces upon the relative rate of change of probability or upon the differences between the underlying probabilities. This can be done e.g. by setting up a stochastic process or starting directly from a steady-state assumption and find a balance between the neighbouring probability classes.

The second alternative can be illustrated for the discrete case as follows. Let x be a random variable and P_x its (discrete) probability function, in tabular form:

x	P_x
0	P_0
1	P_1
2	P_2
3	P_3
⋮	⋮

We then postulate that there is some mechanism connected with text production which is affected by e.g. the difference between the probabilities of subsequent classes ($P_x - P_{x-1}$) or by their proportionality $P_x \sim P_{x-1}$ and which optimises these values according to the flows of information and redundancy, or to criteria still unknown at present. It should be clear that mechanisms and processes of this kind can be neither accessed nor controlled consciously. Among such processes there are e.g. Zipf's forces and also processes which correspond partly to Bühler's functions (cf. Altmann, 1987b): While the speaker forms a text in order to describe (Darstellung) something or make the hearer do something (Appell), he also forms it with respect to some extent of expressivity (Ausdruck) and, at the same time, takes in account, to some extent, the hearer's needs. All these, and more, factors have to be controlled in order to find an optimal compromise, yielding a text of an appropriate composition of information/redundancy, expressivity/descriptivity, production economy/comprehensibility and many more property pairs. If we assume that optimisation takes place with respect to the differences between adjacent probabilities we can set up the following equation:

$$(1) \qquad D = \frac{P_x - P_{x-1}}{[x - (x-1)]P_{x-1}} = \frac{\Delta P_{x-1}}{P_{x-1}}$$

[because $x - (x-1) = 1$] which is a function of x, i.e. $D = f(x)$.

It is an analogy to the continuous case

$$(2) \qquad D = \frac{dy}{y} = f(x)dx$$

(see below). The assumed form of D leads to a hypothesis referring to the given property under the effect of the respective "forces"; one can substitute in (1) or (2) the known or assumed factors and disturbances.

In the second case we have the proportionality

$$P_x \sim P_{x-1}$$

yielding

$$(3) \qquad P_x = g(x)P_{x-1}$$

which is identical to (1) since $g(x) = 1 + f(x)$.

This approach allows us a quite ample, linguistically well-interpretable modelling, encompassing parts of the systems of Pearson, Katz and Ord (cf. Pearson, 1895; Katz, 1965; Johnson & Kotz, 1969; Ord, 1967, 1972) and permitting several extensions. Let us denote by

S = "force" of the speaker, factor of the speaker, expression impulse
H = "force" of the hearer, factor of the hearer, signalling restriction and their (constant) interactions

(4) *S+H, S-H, H-S, SH, S/H, H/S, S-H/S etc.*

If we examine a linguistic variable in isolation, then the approach can include three other kinds of factors entering the formulas:

(1) an "enslaving" constant representing a fixed basic order parameter,
(2) in texts there are disturbances originating in style, kind of text, theme etc. that are difficult to track down,
(3) influences of other levels or variables with which the given variable is joined in the self-regulation cycle of language (collateral or hierarchic influences) but which mostly exert merely a constant pressure if the given variable is examined in isolation.

These four kinds of factors or forces:

(i) Zipf's forces (drive for unification and diversification), Bühler's functions

(signalling, expression), norms of the language community;
(ii) Fundamental quantities, disturbance quantities (random fluctuations, language interference, substratum etc.);
(iii) Effect of language self-regulation (cooperation and competition of subsystems, collateral effects);
(iv) Level quantities, hierarchic enslaving will be denoted by lower-case letters *a, b, c, ...* and used in difference equations.

Let us consider the simplest case of a discrete model, namely

$$(5) \qquad D = \frac{-Sx}{Hx} = -a, \qquad 0 < a < 1$$

where a is one of the constant interactions in (4), here $a = S/H$. In this case the relative difference of "neighbouring" probabilities is constant (signalling a rather undisturbed "norm").

This means

$$\frac{\Delta P_{x-1}}{P_{x-1}} = -a$$

or

$$(6) \qquad \begin{aligned} P_x - P_{x-1} &= -aP_{x-1} \\ P_x &= (1-a)P_{x-1} \end{aligned}$$

where the last formula is a reformulation in terms of (3), with the solution

$$(7) \qquad P_x = (1-a)^x P_0$$

From the condition $\Sigma P_x = 1$ follows

$$(8) \qquad P_x = a(1-a)^x, \qquad x = 0,1,2,...$$

With the usual notation $a = p$, $1-a = q$ we obtain the geometric distribution

$$(9) \qquad P_x = pq^x, \qquad x = 0,1,...$$

in which p can be considered as e.g. an interaction of Zipfian forces.

A distribution of this kind was obtained by Brainerd (1976) when analysing the text as a Markov chain. It is identical with the Stephan-Mishler-Horvath rank-participation law (cf. Stephan & Mishler, 1952; Horvath, 1965; Kadane & Lewis, 1969; Tsai, 1977), controlling the participations in a polylog and is the first rank-order distribution of phonemes assumed by Sigurd (1968); cf. also Altmann (1994).

The model

$$(10) \qquad D = \frac{k - ax}{cx}, \qquad c > a > 0, \qquad k = b + a - c$$

with $r = b/(c-a)$, $q = (c-a)/c$, or analogically

$$P_x = \frac{a + bx}{cx} P_{x-1}$$

with $a/b = r-1$, $b/c = q$ leads to the negative binomial distribution

$$(11) \qquad P_x = \binom{r + x - 1}{x} p^r q^x, \qquad x = 0,1,...$$

obtained by Grotjahn (1982) as a model for the distribution of word length (cf. also Altmann, 1987a,c; Altmann & Best, 1996) and used for several diversification phenomena (cf. Rothe, 1989).

A more general approach using linear functions can be formulated as follows: The locally operating speaker (Sx) and a globally operating fundamental or disturbance factor (b) form a quantity that is controlled by two other quantities, namely the locally operating hearer (Hx) and the globally operating level factor (d), i.e.

$$D = \frac{Sx + b}{Hx + d}.$$

Instead of S and H we shall use a and c, which can themselves consist of an

interaction; i.e., generally

$$D = \frac{ax + b}{cx + d}.$$

From this approach one can obtain a number of probability distributions according to whether the parameters (*a, b, c, d*) are positive, negative, or zero. The advantage of this kind of modelling is its lucidity, interpretability and extendability. The Zipfian forces are here linear and additive, the other factors being additive and constant. If the number of factors is more than four then they either merge into the constants or a new linear multiplicative function is added, i.e.

$$D = \frac{(ax + b)(hx + k)}{ax + d} \tag{13}$$

or

$$D = \frac{ax + b}{(cx + d)(hx + k)} \tag{14}$$

etc. The equations are easy to solve; they automatically yield recursion formulas for the computation of moments and probabilities, and the combination of the parameters with other variables without difficulty yields convolutions, compound and generalized distributions. A number of examples can be found in this volume.

Continuous approach

In the case of the continuous quantities used e.g. in phonetics, or historical linguistics, the approach is analogous. The solution of the differential equation

$$D = \frac{dy}{y} = f(x)dx$$

yields both curves and probability distributions that were already frequently used. In order to demonstrate some known theoretical cases let us adduce some examples.

The approach leading in the analogous discrete case to the geometrical

distribution, namely

$$D = -\frac{b}{d}dx \tag{15}$$

yields the exponential function

$$f(x) = Ke^{-ax} \tag{16}$$

with $b/d = a$, and the exponential distribution if $K = 1/a$ and $x > 0$. Here only global forces operate, e.g. the norm (b) modified by the level (d) of the variable. The approach

$$D = \frac{b}{cx}dx \tag{17}$$

which can be written as

$$\frac{dy}{y} = a\frac{dx}{x} \tag{18}$$

or $d(\ln y) = a\, d(\ln x)$ with $a = b/c$ leads to the power curve known in linguistics as Menzerath's law (the allometric law in biology)

$$f(x) = kx^a \tag{19}$$

which plays an important role both in controlling the subsystems by the system (e.g. syllable length by word length, see Hřebíček, 1994) and in the mutual cooperation of subsystems (cf. Altmann, 1980; Gerlach, 1982; Köhler, 1982; Heups, 1983; Rothe, 1983; Teupenhayn & Altmann, 1984; Fickermann, Markner-Jäger & Rothe, 1984; Schwibbe, 1984; Sambor, 1984; Altmann, Schwibbe, Kaumanns, Köhler & Wilde, 1989; Altmann & Best, 1996).

One can interpret (18) in a very elementary and plausible way: the relative rate of change of the variable y is proportional to the relative rate of change of the variable x. The proportionality constant a can be interpreted as follows: the constant quantity b of the subsystem representing e.g. a norm or arising on combinatorial grounds is locally restricted (reduced) by the demands of the community (c). The approach

(20) $$D = \frac{b - ax}{cx}\,dx$$

leads in the continuous case to the gamma-distribution

(21) $$f(x) = Kx^{b/c}\, e^{-ax/c}$$

with K as norming constant and $x > 0$. The simple linear approach

(22) $$D = (b - ax)dx$$

yields a normal distribution in the form

(23) $$f(x) = K \exp\left(\frac{-a\,(x - \frac{b}{a})^2}{2}\right)$$

with K as norming constant, $a = 1/\sigma^2$, $b = \mu/\sigma^2$, $x > -\infty$, etc.

Applications

Let us illustrate this approach in a concrete case. The most material properties of text units are their lengths. The occurrence of these lengths in a text underlies a rhythm depending on the above-mentioned factors. The length abides by certain laws that can be expressed in form of probability distributions.

If we measure the length of a unit in terms of the number of its immediate constituents then the model is

(24) $$D = \frac{b - ax}{cx}$$

where the constant and global influence (pressure) of the text type (b) is locally "softened" by the (expression) force of the author (ax) and both of them are controlled by the hearer, by the requirement of efficiency of signalling (cx). This approach yields the negative binomial distribution, as shown above, and should hold in the hierarchy

Hřebíček's aggregate
Sentence
Clause
Word
Syllable
Sound

for the neighbouring levels. For the relations sentence/clause and word/syllable numerous corroborations exist (cf. Grotjahn, 1982; Altmann, 1987a; Altmann & Best, 1996).

Yet, if we measure the length of a hyperunit in terms of numbers of units of a non-neighbouring level, e.g. sentence length in number of words or word length in number of sounds, then there is a hierarchically conditioned disturbance or shift originating in the intervening level so that the approach must be enriched by a constant (d)

(25) $$D = \frac{b - ax}{cx + d}.$$

This leads, under special conditions, to the hyper-Pascal distribution that has been well corroborated as the distribution of sentence length (measured in terms of number of words) (cf. Altmann, 1987a).

It is not known how this approach must be modified in case there is more than one intervening level between the construct and its constituents (systems and subsystems).

In the continuous domain this approach yields e.g. the distribution of vowel length in speech measured in milliseconds. If the phonemically fixed length (F) is modified by the expression drive or fluctuation or diversification tendency of the speaker (Sx), by the accent (Bx) and inversely by the environment of the vowel (C/x) and at the same time controlled by the unification force of the community (Hx), then we obtain

(26) $$D = \frac{F + Sx + Bx + C/x}{Hx}\,dx$$

with the solution

(27) $$f(x) = Kx^a e^{bx + c/x}, \quad 0 < x < R$$

with K as norming constant, where in the concrete case at least two parameters are negative (cf. Geršić & Altmann, 1987).

The above approach allows us to develop more complex models as well. So the probability of the change impulse for altering a sound can be derived from the assumptions

$$D = \left(\frac{a}{1 - x} - \frac{b}{x} \right) dx \tag{28}$$

where x is the (normed) effort of the articulation, a is the factor of the speaker and b that of the hearer. This yields the distribution

$$f(x) = K(1 - x)^{-a} x^{-b}, \qquad 0 < x < 1 \tag{29}$$

(cf. Job & Altmann, 1985).

Conclusion

The principle of this kind of model-building is quite comprehensible and in most cases easily applicable.

The constant (norming) factor, the modifying (e.g. diversifying) factors and the damping factors absorbing the fluctuation, controlling the interactions and 'caring' for the efficiency of communication should be adequately inserted in (2) or (3), in order to yield an appropriate model.

Modelling in this way has the advantage of a lucid interpretation, easy expandability, generalization, deducibility, adjustability and systematization. It gives the Bühlerian and Zipfian concept formation a mathematical foundation and thereby renders them useful in theory-building. It is, of course, not applicable without restrictions, because language "forces" can enter into other relations, too.

References

Altmann, G. (1987a). Verteilungen der Satzlängen. *Glottometrika 9,* 147-169.

Altmann, G. (1987b). Bühler or Zipf? A re-interpretation. In: Koch, W.A. (ed.), *Aspekte einer Kultursemiotik.* Bochum: Brockmeyer, 1-6.

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

Altmann, G. (1991). Word class diversification of Arabic word roots. In: Rothe, U. (ed.), Diversification processes in language: grammar. Hagen: Rottmann, 57-59.

Altmann, G. (1994). Phoneme counts. *Glottometrika 14,* 54-68.

Altmann, G., & Best, K.-H (1996). Zur Länge der Wörter in deutschen Texten. In this volume.

Altmann, G., Schwibbe, M.H., Kaumanns, W., Köhler, R. &, Wilde, J. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Brainerd, B. (1976). On the Markov nature of text. *Linguistics 176,* 5-30.

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

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

Geršić, S., & Altmann, G. (1988). Ein Modell für die Verteilung der Vokallänge. *Glottometrika 9,* 49-58.

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

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

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

Horvath, W.J. (1965). A mathematical model of participation in small group discussions. *Behavioral Science 10,* 164-166.

Hřebíček, L. (1994). *Text levels.* Trier: WVT.

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

Johnson, N.L., & Kotz, S. (1969). *Discrete distributions.* Boston: Houghton Mifflin.

Kadane, J.B., & Lewis, G.H. (1969). The distribution of participation in group discussions: An empirical and theoretical reappraisal. *American Sociological Review 34,* 710-723.

Katz, L. (1965). Unified treatment of a broad class of discrete probability distributions. In: Patil, G.P. (ed.), *Classical and contagious distributions.* Calcutta Statistical Publishing Society, 175-182.
Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. *Glottometrika 4,* 103-113.
Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Köhler, R. (1990). Elemente der synergetischen Linguistik. In: *Glottometrika 12,* 39-46.
Ord, I.K. (1967). On a system of discrete distributions. *Biometrika 54,* 649-656.
Ord, I.K. (1972). *Families of frequency distributions.* London: Griffin.
Orlov, Ju.K., Boroda, M.G., & Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.
Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variations in homogeneous material. *Philosophical Transactions of the Royal Society of London, Series A, 186,* 343-414.
Rothe, U. (1983). Wortlänge und Bedeutungsmenge. Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5,* 101-112.
Rothe, U. (ed.) (1989). *Diversification processes in language: grammar.* Hagen: Rottmann.
Sambor, J. (1984). Menzerath's law and the polysemy of words. *Glottometrika 6,* 94-114.
Schwibbe, M.H. (1984). Text- und wortstatistische Untersuchungen zur Validität der Menzerathschen Regel. *Glottometrika 6,* 152-176.
Sigurd, B. (1968). Rank-frequency distribution for phonemes. *Phonetica 18,* 1-15.
Stephan, F.F., & Mishler, E.C. (1952). The distribution of participation in small groups: An exponential approximation. *American Sociological Review 17,* 598-608.
Teupenhayn, R., & Altmann, G. (1984). Clause length and Menzerath's law. *Glottometrika 6,* 127-138.
Tsai, Y. (1977). Hierarchical structure of participation in natural groups. *Behavioural Science 22,* 38-40.

Glottometrika 15, 1996, 77-95

A Dynamic Model of Narrative Reorganization

Wolfgang Wildgen, Bremen

Introduction

The programme of synergetics, which "deals with systems composed of many parts which, by their cooperation, can produce patterns of coherent action on macroscopic scales" (Haken, 1990:3), can be applied to all levels of language and its subsystems.

The following is a partial list for illustrative purposes:

- linguistic societies; their contacts, conflicts, and convergence/divergence,
- systems of language users in interaction, and communication,
- dialogues, monologues and their parts,
- the interaction of sentential constituents and their contribution to the meaning of a sentence,
- the lexicon and its subdomains,
- morphological cooperation of stems and affixes,
- the muscular coordination in articulation, etc.

As language is a highly interactive and cooperative system, any dynamic (i.e. realistic) model of language can apply synergetic concepts and techniques. Section 1 considers some restrictions and introduces a scale of dynamic tools (continuous vs. discrete, deterministic vs. stochastic). It will become apparent that in many cases it is necessary to first build a qualitative model in order to prepare a proper synergetic one. The model we present is dynamic (and discrete), but the probabilistic aspects are not yet considered (the appropriate empirical data are lacking), therefore it is rather a preparatory stage for a more complete synergetic model. The cooperative and the partially stochastic character of language are very

obvious in the domain of sociolinguistics.[1] In a prior study on language shift in Bremen (Wildgen, 1986), I showed that the synergetic models developed by Weidlich & Haag (1983) can be transferred to the study of language shift and language maintenance. In this paper I shall rather consider the next domain on the scale of levels sketched above, i.e. the interaction of language users.

A set of language users doing narrative 'work' on a given text transform the text. The cumulative effect of consecutive transformations produces specific patterns which show:

- rather general patterns of reorganization,
- the probabilistic nature of these changes at the micro-level, i.e. every series of retellings produces different products (these are neither determined strictly by the basic text nor by the length of the series).

The qualitative pattern of these changes will be considered in a preparatory analysis. It is first necessary to introduce some basic notions of dynamic systems theory and to elaborate an instrument for the proper representation of narrative content.

1 Some basic notions of dynamic systems theory

The mathematical theory of dynamic systems is a huge field and the basis of many different models, so that possible choices from this field must be specified. A rough architecture of dynamic systems is summarized below for this purpose (cf. Gilmore, 1980: chapter 1).

For the description of the dynamics of a well-known physical system (e.g. the solar system, falling bodies, etc.) a system of differential equations can be used. For computational purposes these equations may be approached by a system of difference equations (only a discrete grid on the space parameters and a fixed length of time intervals are necessary). Thus, *continuous* and *discrete* dynamic systems can be analyzed side by side. The continuous systems allow for generalizations (general theorems, the search for invariants), the discrete systems are easier to calculate. The strategy for finding good models consists in searching for the simplest system which can still represent important processes and features of the 'real' system. If the dynamic system (the system of differential or difference equations) is gradually simplified two very simple models are finally arrived at.

[1] It is also obvious in lexicology; this explains that lexicology has become one of the central areas of 'synergetic linguistics', cf. Köhler & Altmann (1986).

discrete dynamic systems	**continuous dynamic systems**
- systems of unit vectors	- catastrophes (i.e. basic types of stable processes)
- two-dimensional cellular automata	- bifurcations (generalized catastrophes)

The use of continuous dynamic systems in semantics has been developed in Wildgen (1979, 1981, 1982, 1985). In this article discrete modelling will be used (cf. Wildgen, 1990b). If a probability distribution and stochastic flows (attractors, repellors) are introduced, a synergetic model results. This treatment is thus preparation for a more complex synergetic model.

The question of time and dynamics is one which has puzzled philosophers since antiquity. The answers which have been finally found are still interesting as natural steps towards the solution of the problem (the process is not finished yet).

Three phases can be distinguished (cf. Thom, 1990:314-331):

(a) The mathematics of *time*. In ancient Greece, musical (and celestial) ratios, i.e. harmonic proportions and rhythms were taken as basic phenomena. The rational numbers (e.g. 1/2, 1/3, 1/5 etc.) were the corresponding arithmetic concepts. The basic idea was, however, that of a continuous flow with discrete subdivisions.

(b) The mathematics of simple *geometric* objects like triangles, squares, circles etc. Conceptually this field introduced two (or three)-dimensional abstract entities. The rational numbers became insufficient; irrationals like the square root of 2, 3, etc. had to be considered.

(c) The third and last phase is directly related to the concept of *motion*. Since Kepler (1571 - 1630) and Galilei (1564 - 1642) the solution of this conceptual problem has become the basis for the explanation of celestial and terrestrial kinetics. The differential calculus introduced by Leibniz and Newton opened the way for a systematic solution of the problem of motion in the framework of modern mathematical physics.

These three basic concepts, presented in their historical context, are still relevant. A concept of time, space (different types of spaces will be considered), and motion are needed. The discrete notions of time interval, spatial domain and unit of motion corresponding to the more basic, continuous concepts must be determined. Galilei and, even more radically, Descartes considered only quantitative processes, changes in "extended" matter. We have also applied the basic notion of kinetics and dynamics to qualitative processes and changes (thus we have tried to reintegrate parts of the Aristotelian heritage discredited by modern dynamics since

Galilei).[2] The theoretical background for this extension is provided by modern qualitative dynamics.

The relation between discrete and continuous mathematics is not only a philosophical question. Theoretical dynamics (in physics, chemistry and biology) make use of (continuous) differential equations, whereas concrete calculations are made with the aid of computers, which operate on the basis of discrete algorithms. Thus the coexistence of continuity and discreteness is a very general feature. Section 2 introduces the notion of a two-dimensional cellular automaton which will be used as the formal basis for the model of narrative reorganization.

2 Cellular automata and dynamic systems

Much of formal linguistic modelling relies on a special kind of one-dimensional abstract automata, which may be further specified as a context-dependent, a context-free, or a finite state automaton. If instead of *one* dimension *two* dimensions are considered, we make the step from serial to parallel machines. The latter seem to be a better approximation of real dynamic systems (e.g. the brain).

A *model* of the real system may be based on (continuous) differential equations but can be radically simplified by replacing it by a model with a grid of discrete steps and which is self-similar, i.e. every piece of the system is identical to all the others and the same rules apply to every piece of the system. This allows very quick and highly frequent applications of the same rules to all pieces of the system.[3] This type of mathematical model is called a cellular automaton (CA):

Toffoli characterized the CA as follows:

> "In the *cellular-automaton* model of a dynamical system, the "universe" is a uniform checkerboard, with each square or *cell* containing a few bits of data; time advances in discrete *steps* and the "laws of the universe" are just a small look-up table, through which at each time step

[2] The relation to Aristotelian thought is further elaborated in Thom (1988: chapters 6 to 8). The subtitle of this book is: "Aristotelian physics and catastrophe theory".

[3] T. Toffoli (1984:121) argues that continuous models in physics first "*stylize* physics into differential equations, then (...) force these equations into the mold of discrete space and time and truncate the resulting power series, so as to arrive at *finite difference equations*, and finally, in order to commit the latter to algorithms (...) project real-valued variables onto finite computer words ("round-off"). At the end of the chain we find the computer again a physical system."

> each cell determines its new state from that of its neighbors; this leads to laws which are *local* and *uniform*. Such a simple underlying mechanism is sufficient to support a whole hierarchy of structures, phenomena and properties" (Toffoli, 1984:119).

In comparison to a continuous dynamic model the following transformations are necessary:

a) continuous space and time are replaced by a discrete grid,
b) the system/state at each point remains a continuous variable of the same kind (e.g. real, complex, vector) as in the original equation, and
c) derivatives are replaced by differences between state-variables that are continuous in space and time (Toffoli, 1984:121).

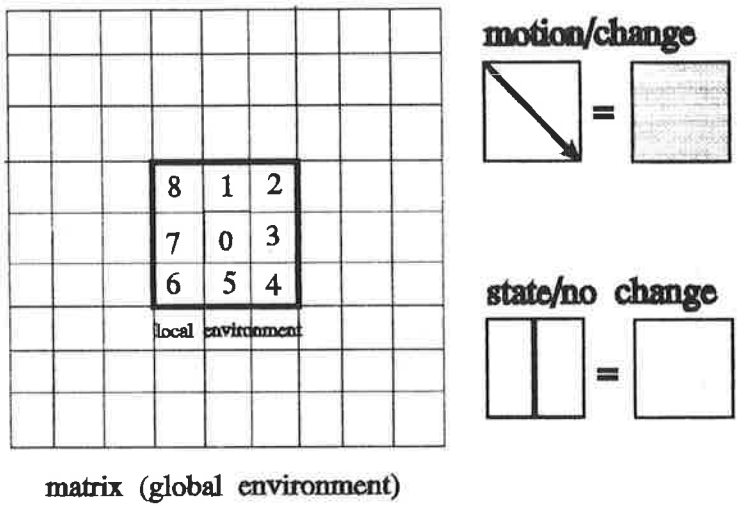


Fig. 1. The basic matrix of the cellular automaton

If, for example, the starting point is a very primitive vocabulary of narrative moves, zero-motion and steady motion, two basic values are possible as every cell

of the system may have either the value 1 or 0. The zero-vectors can be represented graphically as a blank cell and the (positive) unit-vector as a shaded cell. In order to illustrate what a cellular automaton can describe, a very primitivetive model of expansion/reduction of a narrative can be constructed on this basis. The starting point is a matrix and a local environment defined by the neighbouring cells which touch the central cell with at least one point.

Imagine a game where several narrative units (clauses containing an event or state) are given and every participant has to complete sequences of events and to eliminate narrative units without proper followers. The rules of this game can be stated in terms of a CA restricted in terms of specific environments. In the example the set (8,4) of units on the diagonal from upper left to lower right will be subjected to special restrictions.

Rules of the 'narrative' game:

1: zero-vectors (the cell is 0) are not affected by the game,
2: if a cell is 1 and its relevant neighbours (in the set (8,4)) are 0, make it zero
3: if a cell is 1 and one of its neighbours (8,4) is 1, change the other neighbour to 1 also.

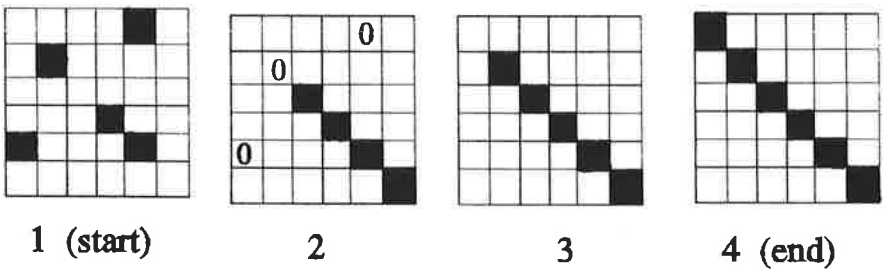


Fig. 2. Four stages of the 'narrative' game

Figure 2 shows different phases of the game with a random starting condition. All units in (2) without a positive neighbour are eliminated and the diagonal sequence is completed until the border line of the matrix is reached. This example suggests that the temporal evolution of a narrative plot (e.g. if a story is retold again and again) can be expressed with the help of a cellular automaton. If a basic representation of words, phrases, and sentences exists, a cellular automaton can

model the transformation S of these patterns (if they are memorized, reproduced, learned, etc.).

3 A basic vocabulary of imaginistic units

3.1 The set of monovalent imaginistic[4] units

All units occupy a unit cell, which is a quadratic field with a vector length on t and $r = 1$. The only possible intervals for t are (0,1) or (0, 1/2, 1); for r the following choices are possible (trivalent units have a double range):

a: r: (0, + 1)
b: r: (-1/2, +1/2)
c: r: (-1, 0).

This leads to the vocabulary of basic imaginistic units represented in Figure 3. As a zero vector on r means no change (semantically), only vectors with a non- zero component on r receive an arrow. The content of these units is described below (cf. Wildgen, forthcoming for a semantic interpretation).

Description:

ad a: Unit interval on r: (0,1)

Unit 1: Continuous, positive motion on r.
Unit 2: Transition (cf. the zero vector at $t = 0,5$) from a continuous positive motion to a stable state.
Unit 3: Transition to the opposite direction.
Unit 4: Transition between two partial, positive motions.

ad b: Unit interval on r: (-1/2, +1/2)

Unit 5: No motion on r, a stable state.
Unit 6: Transition from a state to a (positive) motion.
Unit 7: Transition from a state to a (negative) motion .

[4] We use the term 'imaginistic' instead of 'imagined' in order to point to the more abstract use of the term, which is also compatible with other perceptual channels and with mental categorization (compare Kosslyn, 1980).

Unit 8: Transition between two independent stable states[5]

ad c: Unit interval on *r*: (-1, 0)

Unit 9: Continuous, negative motion on *r*.
Unit 10: Transition from a continuous negative motion to a stable state.
Unit 11: Transition from a negative motion to a positive one.
Unit 12: Transition between two partial, negative motions.

The set of 12 basic imaginistic units is exhaustive for the specific stage of differentiation of this space-time matrix.

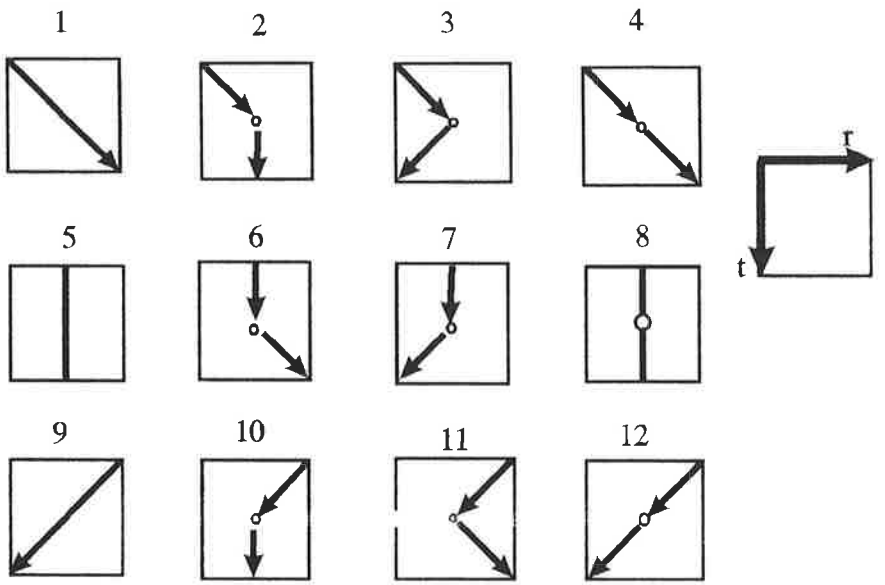


Fig. 3. The set of monovalent imaginistic units

[5] As no antagonistic or protagonistic forces appear in these states, they are treated as zero-vectors on *r*. The implicit change is given by a redefinition of the parameter *r* inside the unit at the transition point: $t = 0.5$, $r = 0.5$.

3.2 The set of bivalent imaginistic units.

The valence of an imaginistic unit is the maximum number of coexisting centres in one unit (at a time interval of that unit). The restriction of valence to the number 3 is practically motivated by the marginality of simple clauses with more than three basic (obligatory) noun phrases with different functional roles (deep cases, thematic roles, etc.)[6]. In order to restrict the set of bi- and trivalent imaginistic units one principle is assumed, which is very general, as it holds for the vast majority of 'real' dynamic systems; they are systematically imperfect, i. e., not in perfect stability and order. This is the *principle of broken symmetry*.

Principle of broken symmetry (principle 1)

If a unit (local system) has two or more sub-units (sub-vectors), at least one of these is dynamically inferior (weak).

The above list has 12 units whose *r*-vectors have length 0, 1/2 or 1. The different units fall into three classes:

r-vector = 0	units 3, 5, 8, 11
r-vector = +/-1/2	units 2, 6, 7, 10
r-vector = +/-1	units 1, 4, 9, 12

In order to construct a bivalent unit, two monovalent units are integrated into one compound bidimensional picture. Thus further two-dimensional units can be formed, which have vectors pointing in different directions and whose points of transition have two inputs or two outputs. Figure 4 gives the list of bivalent units which result from this construction (cf. Wildgen, 1990a and mainly Wildgen, 1994, for a more explicit account).

All other combinations are either trivial (they have one of the twelve basic units as results) or they violate principle 1.

[6] Many of the apparently higher valences turn out to be reducible by the coordination of identical functional roles or by the splitting of one role into several variants of it. In Wildgen (1985) we presented mathematical arguments which restrict the basic set to four roles (= attractors) maximum.

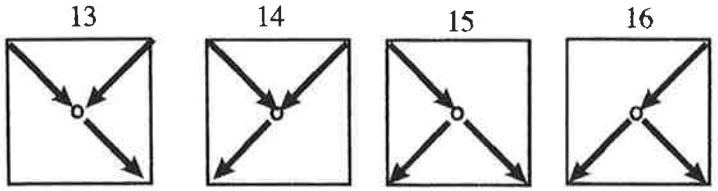


Fig. 4. The set of bivalent imaginistic units

3.3 The set of trivalent imaginistic units

Starting with bivalent units, they are combined in order to allow for a trivalent interaction and the coordinates are readjusted in an appropriate manner. The original bivalent units must be complementary, i.e. positive/negative. This criterion is valid for the pairs 15/13 and 16/14. They lead to four different elementary sequences as Figure 5 shows.

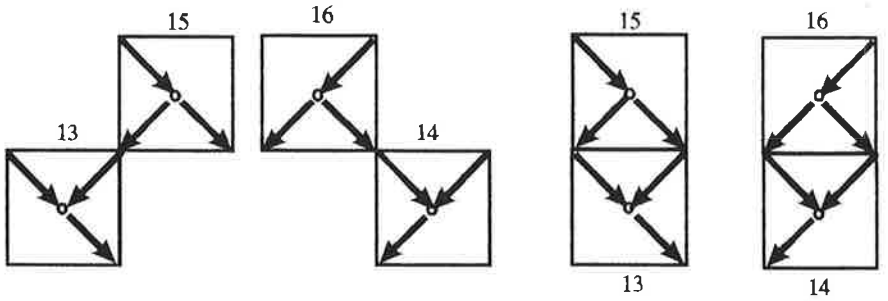


Fig. 5. Four significant combinations of bivalent units

The pairs 15/13 and 16/14 on the left are strongly adjacent if the vectors which relate both units can be added to each other (positively); the pairs on the right are weakly adjacent as the vectors in contact are at an angle of 90°. However, the two-dimensional combination of the bivalent pictures is insufficient to represent trivalent units. They rather represent a sequence of events (and thus of narrative units). Looking more closely at these units, it becomes apparent that an internal order of processes exists in all trivalent units. First, a process of emission (positive bifurcation) takes place and later a complementary process of capture (negative bifurcation) closes the overall movement, resulting in the two units being adjacent. In the representation both units have been transformed to 1/2 of their size. The

underlying strong units (in No. 13, 15 it is No. 4, in No. 15, 14 it is No. 12) are extended by adding the units No. 1 and No. 9. This doubling means that the new units have a scale on the *r*-dimension of: (-2, 0), (0, +2) and on the *t*-dimension of: (0, 2). However, in the narrative sequence they are considered as one unit and the space reserved in the matrix does not mirror any gain in importance or functional weight. Two groups of trivalent units can be distinguished.

a) double-spaced on *r* : (-2, +0) or (0, +2)
b) double-spaced on *t* : (0, 2)

This results in the four units: 17, 18, 19, 20, which are shown in Figure 6.

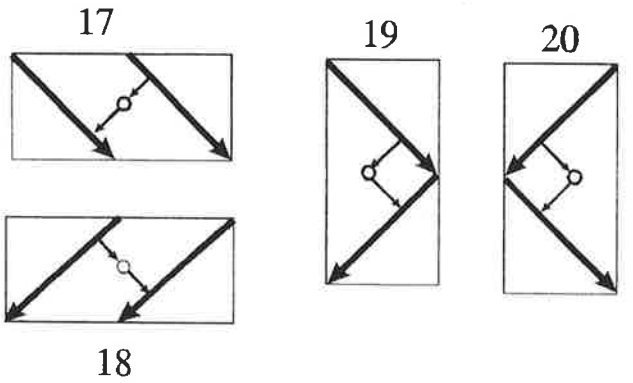


Fig. 6. The set of trivalent imaginistic units

The units 19 and 20 are arranged vertically because the *r*-vectors in the units join in the vertical direction. The exact construction is given in Wildgen (1994, chap. 7).

4 The dynamics of two series of reorganization in story-telling

Narratives have their own life. They are first told after the event has happened and, if the event is relevant enough, they are retold several times by the person who had the experience and may be retold by a member of his audience. In a classical experiment Bartlett gave different stories as the starting point of a sequence of retelling. In Stadler & Wildgen (1987) we repeated his experiment with a German version of one of his stories in order to observe the processes of self-organization

in the sequence of retold versions of the story. A summary of the results follows. The original version (in Bartlett, 1967:65) was:

The War of the Ghosts

One night two young men from Egulac went down to the river to hunt seals, and while they were there it became foggy and calm. Then they heard war-cries, and they thought: "Maybe this is a war-party". They escaped to the shore, and hid behind a log. Now canoes came up, and they heard the noise of paddles, and saw one canoe coming up to them. There were five men in the canoe, and they said:

"What do you think? We wish to take you along. We are going up the river to make war on the people".

One of the young men said: "I have no arrows". "Arrows are in the canoe", they said.

"I will not go along. I might be killed. My relatives do not know where I have gone. But you", he said, turning to the other, "may go with them." So one of the young men went, but the other returned home.

And the warriors went up the river to a town on the other side of Kalama. The people came down to the water, and they began to fight, and many were killed. But presently the young man heard one of the warriors say: "Quick, let us go home: the Indian has been hit". Now he thought: "Oh, they are ghosts". But he did not feel sick, but they say he had been shot.

So the canoes went back to Egulac, and the young man went ashore to his house, and made fire. And he told everybody and said: "Behold I accompanied the ghosts, and we went to fight. Many of our fellows were killed, and many of those who attacked us were killed. They said I was hit, and I did not feel sick".

He told it all, and then he became quiet. When the sun rose he fell down. Something black came out of his mouth. His face became contorted. The people jumped up and cried.

He was dead.

Information which is important for the balance of forces in the story is the number and type of participants (protagonists and antagonists). The participants are:

A: *young men* (normally the protagonists);	number:	2
later:		1
B: *warriors* (normally the antagonists);	number:	5

Figure 7 summarizes the content of the first four episodes and translates the summaries into imaginistic units.

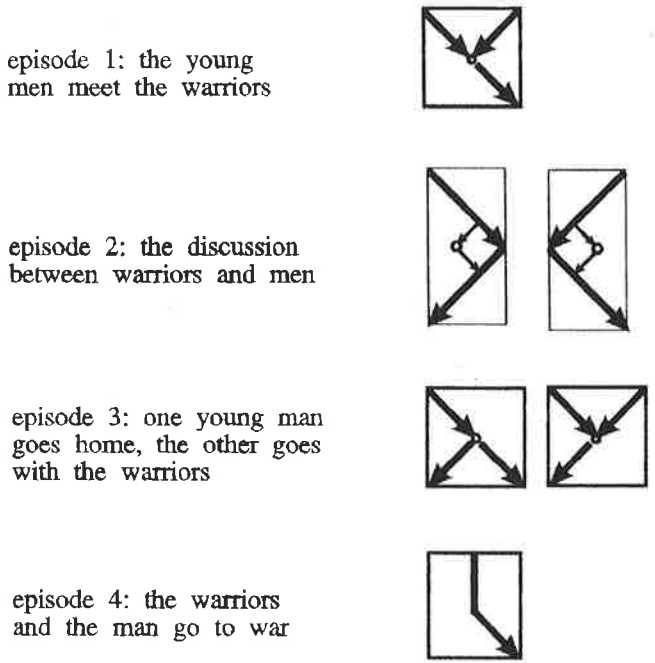


Fig. 7 The first four episodes and their imaginistic representation

In order to show the elaboration of the episode, the content of the first episode in the original story is fully reproduced. The central event is the point where both lines of protagonistic and antagonistic movement meet, i.e. in t_7 (s. Fig. 8). "They saw one canoe coming up to them". This central domain may be analysed as two units: t_{8a}: they saw one canoe + t_{8b} : the canoe came up to them. A list of simple propositions extracts the content of the first episode:

t_1: two young men went down to the river
t_2: it became foggy and calm (change of quality)
t_3: they heard war cries (perceptual process)
t_4: they thought: "---" (mental process)
t_5: they hid behind a log
t_6: canoes came up (antagonistic motion)
t_7: they heard the noise (perceptual process)
t_{8a}: they saw one canoe (perceptual process)
t_{8b}: the canoe came to them (antagonistic move against the protagonists)

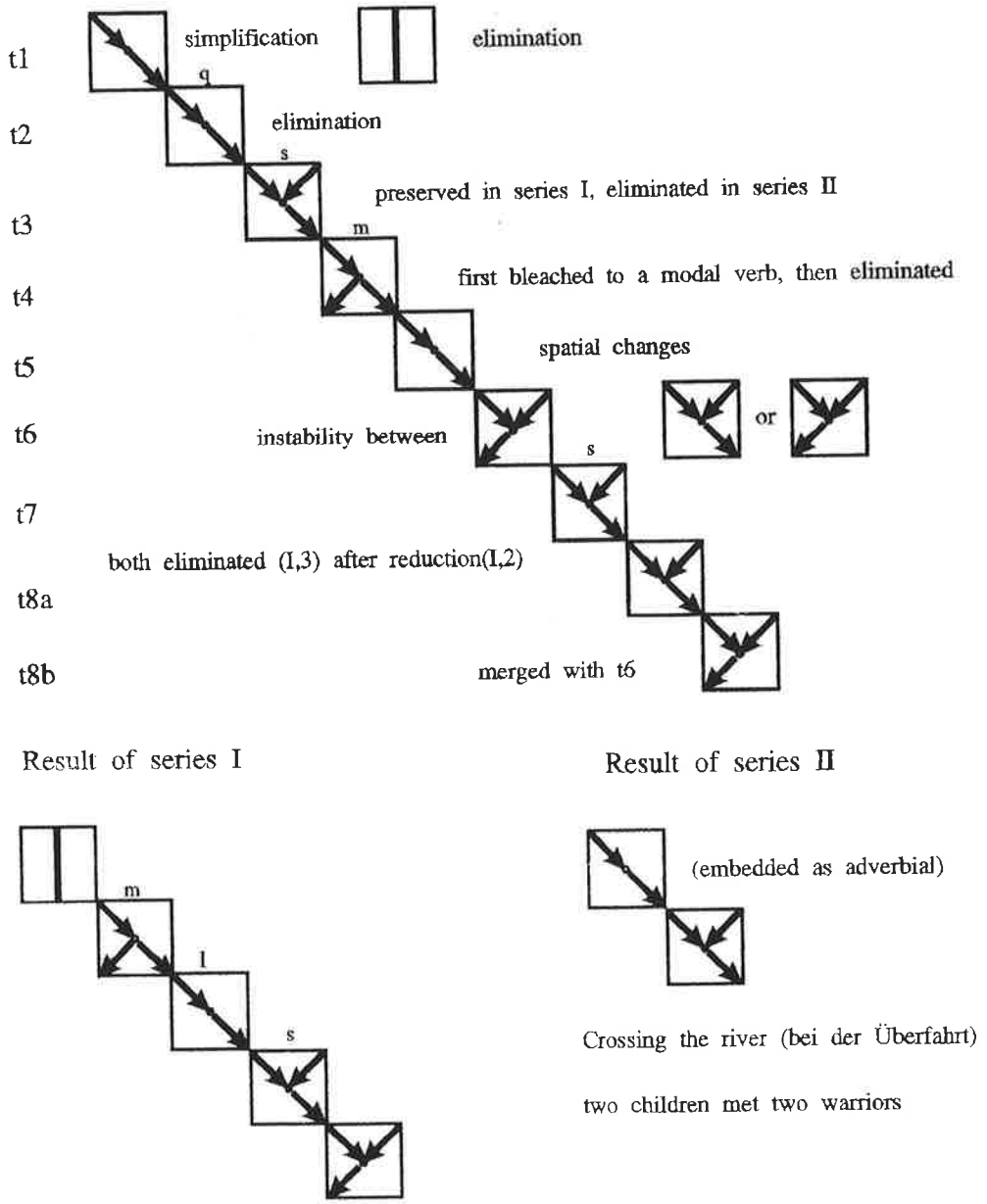


Fig. 8. Analysis of the first episode and modifications in the two series I and II

Figure 8 shows the analysis of the first episode in terms of the imaginistic grammar (left) and the kind of modifications appearing in the two series of retellings (I and II). The other episodes may be described in the same manner.

The sequence of retellings shows not only the diminution or replacement of details but also of episodes. Here we shall consider the reorganization of the topology of the episodic structure together with changes in the force profile. The first series of retellings contains the following reorganizations (cf. Table 1):

Table 1

story number	content	force profile: young men	warriors
1		2	5
2	elimination of episode 3	2	5
3	elimination of episode 4	3	5
4		3	N (passive)
5	a new unit is introduced	3	1
6	all participants go to war	3	1
7		3	1

The central interaction between the participants in episodes 3 - 4 has dramatically changed; the force profile is inverted ($2 < 5$ vs. $3 > 1$) and the 'capture' is replaced by an agreement. With the inverted force profile, as shown in Table 1, we can say that the protagonists ($N = 3$) take the warriors as their helpers (as secondary protagonists) with them (capture). The whole configuration is thus symmetrically shifted.

The second series follows a different pattern of reorganization, as shown in Table 2.

Table 2

story number	content	force profile: young men	warriors
1		2	*N* (some warriors)
2	the discussion between young men is introduced	2	*N*
4	persuasion instead of capture	2	2 (symmetry)
7	the young men ask the warriors for help	2	2
10	the young men become 'children'	2	2

In story 2, the communicative exchange initiated by the warriors is changed into a discussion between the young men. In story 4, the *physical* capture is changed into a *mental* capture (persuasion). These are slides on the interpretation scale. The qualitative change of the protagonists from young men to children (No. 10) is a consequence of the symmetry of forces established in 4 and its reversal in 7. Thus, the understanding of the seemingly stable 'dramatis personae' is shifted depending on the holistic configuration in the story.

These shifts in the first part have consequences for the result of the story. In the original story, the young man who goes with the warriors finally dies. In the series, it is the warrior (the helper of the 'children') who finally dies. In general, the reorganization of stories may be seen as a process with different branches and with chance phenomena governing the transitions.

Going beyond the rather coarse classification underlying the imaginistic grammar, focus can be put on the transformation of the imagined scene in episode 1. In the original story the young men go down to the river and hide behind a log on the shore. The "five men in the canoe" come up to them on the river. We can label the displacement of the protagonists (the young men) as A and the displacement of the antagonists as B. The central event is just the contact between the two separate views A and B: (A, B). If the spatial positions of A, B and (A, B) are compared, interesting transformations can be observed.

In the first series the displacement A no longer sticks to the shore, the protagonists cross the river (No. 3) and finally meet the antagonists on the shore at the other side (No. 7).

At the onset of the second series, the protagonists are seen as being in a canoe on the river where they meet the warriors. Later on in the series (No. 6) they take

off from the shore and meet the antagonists on the river and finally (No. 11) both parties start from the opposite shores and meet in the middle of the river. If we compare the results of the two series we see that an inhomogeneous scenario (the protagonists are on the shore, the warriors are on the river) is transformed into an homogeneous one:

- series I: The meeting (A, B) occurs on the shore.
- series II: The meeting (A, B) occurs on the river (two canoes meet on the river).

This result corresponds to the gestalt principles governing serial reproductions of visual patterns (cf. Stadler & Wildgen, 1987:117f. and Stadler & Kruse, 1990).

Conclusion

Even such an extremely complex cognitive accomplishment as the retelling of a story shows the effect of very basic principles of 'gestalt', or, in synergetic terms, of "slaving parameters". In the analysis of the imaginistic sequence in an episode of the story (cf. Fig. 7) we have shown that two basic operations occur:

(a) The non-localistic entities are eliminated: t_2, t_3, t_4, t_7, t_8. In series I, however, t_3 was preserved, whereas in series II the whole episode was reduced very quickly to its localistic kennel.

(b) The final outcome (mainly in series II) corresponds to the final unit which links this episode to the following one. The preparing displacement of the protagonists is syntactically contracted to an adverbial ("bei der Überfahrt" = at the crossing).

(c) The analysis of the reorganization in the imagined scenario shows a tendency towards a homogeneous, stable configuration. This result is parallel to results in visual reorganization.

In general we can conclude that the mental reorganization of stories is strongly governed by changes on the level of processual and spatial imaging. Although the transformation of an input-story is not rule-governed, it shows rather simple regularities due to strong (slaving) parameters which govern mental imagery. Future research in synergetic psycholinguistics should extract these slaving parameters and try to build a synergetic analogon which recognizes variants of a story and allows for creative variation of a basic plot.

References

Bartlett, F. Ch. (1967). *Remembering. A Study in Experimental and Social Psychology.* Cambridge: Cambridge U.P. (first published in 1932).

Gilmore, R. (1980). *Catastrophe Theory for Scientists and Engineers.* New York: Wiley.

Haken, H. (1983). *Synergetics. An Introduction* (third revised and augmented edition). Berlin: Springer.

Haken, H. (1990). Synergetics as a Tool for the Conceptualization and Mathematization of Cognition and Behaviour - How Far Can We Go? In: Haken & Stadler (eds.), 1990, 2-31.

Haken, H., & Stadler, M. (eds.) (1990). *Synergetics of Cognition.* Proceedings of the International Symposium at Schloß Elmau. Berlin: Springer.

Köhler, R., & Altmann, G. (1986). Synergetische Aspekte der Linguistik. *Zeit schrift für Sprachwissenschaft 5,* 253-265.

Kosslyn, S. M. (1980). *Image and Mind.* Cambridge (Mass.): Harvard U. P.

Stadler, M., & Kruse, P. (1990). *The Self-Organization Perspective in Cognitive Research: Historical Remarks and New Experimental Approaches.* In: Haken & Stadler (eds.), 1990, 32-52.

Stadler, M., & Wildgen, W. (1987). Ordnungsbildung beim Verstehen und bei der Reproduktion von Texten. *Siegener Periodicum zur internationalen empirischen Literaturwissenschaft (SPIEL) 6,* 101-144.

Thom, R. (1988). *Esquisse d'une sémiophysique. Physique aristotélicienne et théorie des catastrophes.* Paris: Interéditions.

Thom, R. (1990). *Apologie du Logos.* Paris: Hachette.

Toffoli, T. (1984). Cellular Automata as an Alternative to (rather than an approximation of) Differential Equations in Modeling Physics. *Physica 100,* 117-127.

Toffoli, T., & Margolus, N. (1987). *Cellular Automata Machines.* Cambridge (Mass.): MIT-Press.

Wagner, K. H., & Wildgen, W. (1990). *Studien zur Grammatik und Sprachtheorie* (BLICK 2). Bremen: Universitätsverlag.

Weidlich, W., & Haag, G. (1983). *Concepts and models of a quantitative sociology. The dynamics of interacting populations.* Berlin: Springer.

Wildgen, W. (1979). *Verständigungsdynamik. Bausteine für ein dynamisches Sprachmodell.* (Habilitationsschrift). Regensburg: University of Regensburg (Ms.).

Wildgen, W. (1981). Archetypical Dynamics in Word Semantics: An Application of Catastrophe Theory. In: Eikmeyer, Hans-Jürgen, & Rieser, Hannes (eds.),

Words, Worlds, and Contexts. New Approaches to Word Semantics. Berlin: de Gruyter, 234-296.

Wildgen, W. (1982). *Catastrophe Theoretic Semantics. An Elaboration and Application of René Thom's Theory.* Amsterdam: Benjamins.

Wildgen, W. (1985). *Archetypensemantik. Grundlagen einer dynamischen Semantik auf der Basis der Katastrophentheorie.* Tübingen: Narr.

Wildgen, W. (1986). Synergetische Modelle in der Soziolinguistik. Zur Dynamik des Sprachwechsels Niederdeutsch-Hochdeutsch in Bremen um die Jahrhundertwende (1880-1920). *Zeitschrift für Sprachwissenschaft 5,* 105-137.

Wildgen, W., & Mottron, L. (1987). *Dynamische Sprachtheorie. Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese.* Bochum: Studienverlag Brockmeyer.

Wildgen, W. (1987b). Processual Semantics of the Verb. *Journal of Semantics 5,* 321-344.

Wildgen, W. (1988). Konfiguration und Perspektive in der dynamischen Semantik. *Linguistische Berichte 116,* 311-343.

Wildgen, W. (1989). La structure dynamique du récit. *DRLAV. Revue de Linguistique 41 (Écriture et formalismes),* 53-81.

Wildgen, W. (1990a). Basic Principles of Self-Organization in Language. In: Haken & Stadler (eds.), 1990, 415-426.

Wildgen, W. (1990b). Sketch of an Imaginistic Grammar for Oral Narratives. In: Wagner & Wildgen 1990, 85-121.

Wildgen, W. (1994) *Process, Image and Meaning. A Realistic Model of the Meaning of Sentences and Narrative Texts.* Amsterdam: Benjamins (Companion Series: Pragmatics and Beyond).

Glottometrika 15, 1996, 96-101

Word Associations and Text

Luděk Hřebíček, Prague

In *Glottometrika 13* G. Altmann (1992) presented a deep analysis of the distribution of word associations. In this paper the concordance between the observed distributions on the one hand, and different possible theoretical distributions, as proposed in the history of this research, on the other hand, is investigated in detail. It is shown that the agreement of the data with the Zipf-Dolinskij distribution is excellent and that other theoretical distributions do not appear to be satisfactory. However, there still remains the problem of systematization which is a necessary condition of theory building, as Altmann himself characterizes the situation. The present paper tries to fill in this gap in the theory. We also try to stress its semantic consequences.

Word associations represent a psychological test enabling us to examine certain properties of the human mind. The nature of word associations, however, can be better understood when they are also viewed from the linguistic side of the problem. In psychological experiments linguistic entities are used as a means to look into the relevant spaces of the human mind. Logically, we need to obtain non-trivial information concerning the language used in psychological experiments. This non-trivial information can be offered especially by quantitative linguistics which is a branch of the theoretical knowledge of natural languages based on methodologically acceptable approaches to language. This mainly concerns the engagement in formulating general linguistic laws such as the Menzerath-Altmann law which is applied in the present theoretical attempt. Unfortunately, in linguistic circles such laws are not as well-known as they deserve to be, and philologists, also sometimes called linguists, are not always able to understand the far-reaching consequences of these laws.

It is not difficult to suspect word associations of having some links to the systems of meaning usually called 'semantic systems'. Words are vehicles of meanings, and associations exhibit the relations between their meanings. However, what should one understand by the concept of 'semantic system', i.e. a system of meanings? Sometimes semantic systems (and also entire languages) are supposed to be a mythical cistern from which the language users draw units bearing meanings for their communication needs. This conception can hardly be supposed to be resistant

to deeper criticism. It is more adequate to suppose that language is software implemented in individual human heads with their biological hardware. There are smaller or larger deviations among individuals as far as semantic systems in their heads are concerned. The process of implementation of this software goes on in every communication process. The concept of communication is very broad; it does not represent only communication processes using natural languages (cf. Bunge, 1983). And the 'semantic system' of linguists who operate with meanings of language entities is nothing but an informal abstraction from a certain number of texts. We can also say that this abstract semantic system is a non-explicit inductive generalization from the real individual semantic systems existing in human minds. The idea of individual personal semantic systems is evidently more realistic than that mystical cistern of meanings. And there are many agreements and many differences among the individual semantic systems, not only because of the differences in the interconnections of neurons carrying our knowledge (cf. Bindra, 1976) but also because of the different ways of obtaining this knowledge through communication events. These differences are worthy of deeper investigation. The *abstraction,* however, is questionable as there is no information about the individual semantic systems. The main task of semantics is to search for what is common to these individual systems.

We cannot look into the brain and investigate these systems. One possible way to obtain some information is the psychological approach based on word associations. The other way is offered by linguistics, especially by text linguistics.

A natural environment for words is a text. By the term 'lexical units' we mean here the units which are semantically interpreted either in the way which is presented in standard dictionaries of natural languages, or as they are interpreted by normal language-users dealing with texts. We mean users without lexicological training. For the analyses in text linguistics any semantic interpretation of a word in a text is correct. Each text contains a dictionary of (interpreted or not interpreted) words, which we call 'lexical units'. This indicates that words are combined with the entities of semantic systems in relations which are to a certain degree free. Semantic systems appear to be something *behind* languages and not *in* languages. Language units are often supposed to be firmly joined with their meanings; this is useful and advantageous when language structure is described with the help of the intuitively grasped meanings. Our task, however, is to approach the semantic system from the side of language structures. This is one of the aims of testing word associations. Language is a window looking at the landscape of semantics.

Relations among words signal the relations among meanings. These relations are evident from grammatical and reference relations of words in sentences and in text. Consequently, from grammar and text references inferences to semantic

relations can be made. It is rational to begin with the entire text and to take into account all its structural levels. The linguistic levels are explicitly defined by the Menzerath-Altmann law, which is one of the universal linguistic laws. More than one decade after its formulation (see Altmann, 1980; Altmann & Schwibbe, 1989) the law has been corroborated on different levels and in different languages. This law was used for derivation of further fundamental laws of natural languages, especially in the domain of self-regulation (cf. Köhler, 1986). Since we try to use this law for the solution of our problem, let us describe it briefly. The law states that

The longer the language construct the shorter its constituents.

The formula derived by Altmann is as follows:

(1) $y = Ax^b$

where $x =$ the length of the construct,
$y =$ the (mean) length of its constituents,
$A, b =$ constants; with regard to the law b takes negative values.

With the help of this law a special aspect of the semantic structure of a text can be disclosed. Let us consider a word, i.e. a lexical unit which obtained a communicator's semantic interpretation. Each word occurs in a certain sentence or sentences of the text. These sentences - joined semantically by means of this word - form an entity which we call 'text aggregate' or simply 'aggregate'.[1] It was shown that aggregates are constructs in the sense of the Menzerath-Altmann law: the higher the number of sentences in an aggregate the shorter the mean sentence length measured by the number of words. The details are described in Hřebíček (1989, 1992), where the theory was tested on Old Ottoman and Turkish texts, and also in Schwarz (1992), who checked it on German texts and derived a theoretical distribution of aggregates. Let us remark that a text also appears to be a construct in relation to its aggregates representing its constituents (cf. Hřebíček,1993).

It is quite natural to assume that all lexical units occurring in an aggregate point to a subsystem of the respective semantic system (proper to its producer or to its interpreter). Thus aggregates also offer an insight into semantic systems. Consequently, we have two narrow openings to this system and when we are looking

[1] M. Bunge (1977:263) calls a thing an aggregate if its parts are not bonded. Thus text aggregates are, as a matter of fact, systems because their parts are bonded.

through one or the other window, we should see approximately the same picture. Word associations and aggregates reflect the situation existing at a given moment in the mind of the tested person or of the communicator. We may thus treat this situation as follows:

When x is the number of words associated with a given word, it is quite natural to suppose that the function f_x has x as its argument. The observed distributions of word associations indicate that for each distribution the value f_1 corresponding to $x = 1$ represents a parameter to which all other values of f_x are related. This leads to the following hypothesis:

The fraction f_1/f_x is proportional to x.

This relation can be presented in a logarithmic transformation, i.e. $\ln(f_1/f_x) \sim \ln x$, so that the hypothesis can be written in the form:

(2) $\ln f_1 - \ln f_x = \ln c\ (\ln x).$

When this equation is divided by $\ln x$, the following expression is obtained:

(3) $$\left(\frac{f_1}{f_x}\right)^{\frac{1}{\ln x}} = c.$$

Parameters similar to c are usually constants, but this is not the case for c in (3). From the different observed values of the distribution of word associations it is evident that this quantity in fact equals some structure having the character of a variable. And here the second hypothesis derived from the modelling idea explained above is to be introduced:

Word associations represent lexical units occurring in a real or potential text aggregate.

This means that we assume that the tested person is able to produce a text in which the associated words form an aggregate. Aggregates abide by the Menzerath-Altmann law. Therefore we put c equal to the right-hand side of (1). From (3) we then obtain:

$$(4) \qquad \left(\frac{f_1}{f_x}\right)^{\frac{1}{\ln x}} = Ax^b.$$

With $A = e^a$ we obtain from (4):

$$(5) \qquad f_x = f_1 x^{-(a + b \ln x)}.$$

This, however, is the so-called Zipf-Dolinskij function (cf. Hammerl, 1991) which is in perfect concordance with the observed word association data. This result enables us to formulate some consequences which have the status of conjectures.

The second one of the two basic hypotheses presented above indicates that the assumed semantic system represents a linguistic level, i.e. a language construct in the sense of the Menzerath-Altmann law. The size of this construct (i.e. aggregate) can be measured not only in the number of sentences, but also in the number of words, as is the case for word associations; of course, this does not exclude the possibility of also measuring its size in other linguistic units. And thus lexical units, or better still their meanings, appear to be constituents of these constructs.

These conclusions are not in contradiction with the ideas presented above according to which a semantic system appears to be a phenomenon related to different communication means, one of which is natural language. Both these phenomena are simply governed by the same principle, i.e. by the Menzerath-Altmann law. Or it would be more correct to say that semantics constitutes a connection between language and non-linguistic reality. Perhaps language is a phenomenon partly mapping the shape of the semantic system. All these are only conjectures, but there is no doubt that a large new field for theoretical linguistic investigations is opened due to the Menzerath-Altmann law.

References

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

Altmann, G. (1992). Two models for word association data. *Glottometrika 13,* 105-120.

Altmann, G., & Schwibbe, M.H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Bindra, D. (1976). *A theory of intelligent behavior.* New York: Wiley.

Bunge, M. (1977). *Treatise on basic philosophy. Vol. 3. Ontology I: The furniture of the world.* Dordrecht: Reidel.

Bunge, M. (1983). *Treatise on basic philosophy. Vol. 5. Epistemology & Methodology I.* Dordrecht: Reidel.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik. Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.

Hřebíček, L. (1989). Menzerath-Altmann's law on the semantic level. *Glottometrika 11,* 47-56.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures.* Bochum: Brockmeyer.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., & Rieger, B. (eds.), *Contributions to Quantitative Linguistics.* Dordrecht: Kluwer, 33-39.

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

Schwarz, C. (1992). *Zur Verteilung von Aggregaten in Texten.* [Unpublished seminar paper]. Bochum: Sprachwissenschaftliches Institut der Ruhr-Universität.

Glottometrika 15, 1996, 102-111

Diversification Processes of the Word

Gabriel Altmann, Bochum

1. In the empirical sciences one always has striven to reduce disparate phenomena to a common denominator, i.e. one has tried to construct a theory from which one could derive models for a large class of phenomena. As a rule, this was not possible until one recognized behind certain phenomena the same principle or mechanism, and introduced mathematization. The history of sciences amply evidences progress of this kind.

In linguistics there have also been attempts of this kind, e.g. to reconcile phonology, morphology, syntax and semantics, but these attempts have concerned above all descriptive principles; no theory was constructed. G.K. Zipf succeeded in finding a very powerful principle, namely that of least effort (cf. 1949) holding in all domains of language and appropriate to be used in mathematical modelling. This principle leads the speaker to unify or to diversify the properties of language units in order to alleviate his/her physical or mental effort. These two processes are not to be considered simply as fluctuations or random deviations from a norm but as eternal processes operating in the self-regulation of language and resulting in special curves and probability distributions of language properties.

If we succeed in deriving them from a general principle then we are on the threshold of a theory of language. The first step in this direction was taken by Köhler (1986) who introduced a general principle for one control cycle of language. A number of diversification phenomena, theoretical problems and models can be found in Rothe (1991).

In this paper we are not interested in the development of the mathematical apparatus (cf. Altmann, 1991; Köhler & Altmann, 1996) - but rather in the enumeration of diversification phenomena associated with the word.

2. We consider language units as subsystems of language and the idiolect carrier either as a subsystem or the environment of language. The idiolect carrier gets through his communicative needs whereby he/she acts upon the shaping of language units. Thus language units are exposed to an incessant stress that can be adjusted only by means of self-regulation. Very rapid changes would lead to the destruction of communication, slow changes lead to evolution, rise of dialects, sociolects, etc. Self-regulation does not mean a simple return from a fluctuation to

the old state, but also the maintenance of a steady state requiring not constant forms but invariant stochastic relations between variable forms.

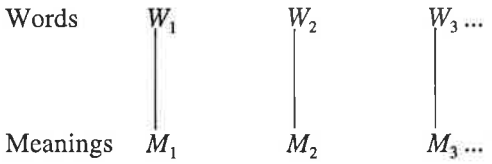
The shaping of the properties of language units caused by the intervention of the idiolect carrier follows certain laws[1]. And it is just the formulation of these laws that can establish a portion of a theory of language. Diversification is one of the lawlike processes operating on all levels of language, but of course it is not the only process in language. At the same time the unification process (and surely other, as yet unknown processes) operates against diversification, they are in competition with one another - they generate steady states. The "Zipfian processes" of diversification and unification, being the consequences of economy and other needs of idiolect carriers, are the central processes that make language a synergetic whole in which cooperation, competition and self-regulation play a fundamental role.

In this contribution we restrict ourselves to the diversification of some basic units such as word, morpheme or stem and show fourteen different aspects.

3. Diversification of meaning

(a) The law of polysemy

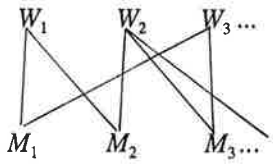
Under extreme circumstances that would fully satisfy the needs of the hearer, every word (morpheme, stem) taken as an acoustic pattern would have exactly one meaning. Schematically



This is perhaps the "birth state" of every word; it can change in the course of time. If the word is more frequently used, it "captures" further meanings since this better meets the needs of the speaker who wants to express more with one word, according to Zipf (1949). From time to time a new junction between the elements

[1] "Everything abides by laws" (Bunge, 1977:17).

of the set W (words) and that of M (meanings) arises. After a time it may look e.g. as follows



In these cases we say that the meaning of the word diversifies, that the word becomes polysemantic or polylexic. If there were no restriction on this process then after a sufficiently long time all words would have all meanings. This is of course not the case, because (i) if the speaker ascribes the meaning M_j to the word W_i then he/she no longer needs W_j. W_j can disappear - so that finally the speaker would possess only one single word - or it can lose the meaning M_j; (ii) if the hearer does not agree with the new assignment $W_i \rightarrow M_j$ then he/she (i.e. the community) need not accept it and it can be eliminated.

It is known that there is no natural language exhibiting these extreme states, i.e. there are neither languages having only one single word nor languages in which every word has exactly one meaning, nor languages where all words have all meanings. Because of the above-mentioned synergetic situation, every language has its own measure of word polysemy (polylexy) that must follow a law providing steady states, i.e. there must be a law in the form of a probability distribution prescribing that there are f_x words having x meanings ($x = 1,2,3,...$).

Krylov (1982a,b) was the first to attempt to derive this law, which we therefore call *Krylov's law* or the law of polysemy.

(b) The rank-frequency law of (denotative) meanings

If a word has several (grammatical or lexical) meanings they are not used with the same frequency; the frequencies are heterogeneous. This heterogeneity is a measure against the diversification in point (a) above taken under the pressure of the hearer: To eliminate one of the meanings of a word means to reduce the frequency of occurrence of *this* meaning of the word to zero. If the process in (a) is lawlike then this process is lawlike, too, and it must be possible to show that the frequencies of occurrence of the meanings of the word form an ordered whole. If we rank the frequencies of the individual meanings of a word according to their magnitude then we can ascertain that they follow the negative binomial

distribution (see Beöthy & Altmann, 1984a,b; Altmann, 1985; Rothe, 1985; Altmann, Best & Kind, 1987), which can be derived from different assumptions.

While Krylov's law represents the interplay of Zipfian forces (unification and diversification) in language as a whole, this law, which can be called *Beöthy's law*, operates on individual words. Since, as a matter of fact, they control the same property on a macro- and a micro-level, they must be derivable from a common basis.

(c) Associative diversification

If the hypothesis in (b) is correct and the rank-frequency distribution of denotative meanings abides by Beöthy's law, then the same must also hold for the rank-frequency distribution of the connotative meanings of the word, which can easily be ascertained through the investigation of word associations (see Cramer, 1968).

Every speaker of a language community connects with every word certain associations, some of which are exclusively subjective, while the others are common property. The common part can be recognized by its frequency of occurrence when interviewing test persons. The more frequent an association, the stronger it will be anchored in the language consciousness of the idiolect carrier, and the easier it is for it to become a denotation. The connotations, the associations, are the source from which polysemy formation springs. Since they constitute merely a parallel to denotations, their rank-frequency distribution must abide by Beöthy's law. Several tests with the material of Palermo & Jenkins (1964) have corroborated this assumption without exception though there are other models too (cf. Altmann, 1992; and Hřebíček, 1996).

(d) The law of synonymy

The scheme presented in (a) can also be imagined in reverse. If the speakers have the need to ascribe a (new) meaning to a state of affairs, then they must assign to it a word (an acoustic pattern). If they take for this purpose an existing word (e.g. "group" in sociology and algebra, "tree" in botany and graph theory) then the word becomes polysemic; if they invent a new word, then at the beginning it will be monosemic, but once the given meaning is embodied, in the course of time it can find expression also in other acoustic patterns (words), which then become synonymic. The formation of synonymy is thus the reverse of the process leading to the formation of polysemy; in the scheme in (a) we simply have to exchange the places of W_i and M_i. Full synonymy is extremely rare because two words with one common meaning can have several different meanings or different connotations.

There is no reason to assume that the formation of synonymy would proceed in another way than polysemy formation, since both of them are merely two sides of the same coin. Therefore it is plausible first to test whether Krylov's law holds here, too. Whatever the law by which these processes abide, it should be the same for both of them.

If this is true, then the rank-frequency distribution of the synonyms also follows the same law as that of the individual meanings of a word, i.e. *Beöthy's law.*

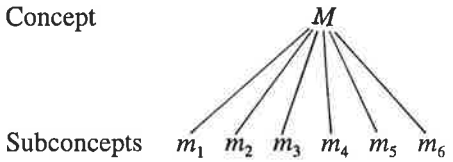
These hypotheses have not been tested as yet.

4. Diversification of subconcepts

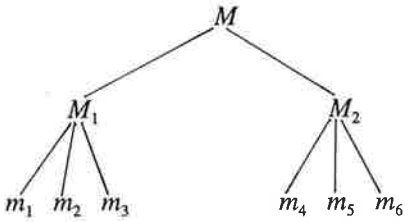
Let us consider a concept like German "Waffe" (weapon). It is the superconcept of a set of subconcepts such as "Feuerwaffe" (fire-arm), "Stichwaffe" (thrust weapon, e.g. dagger), ..., which in turn are superconcepts of other concepts. In the course of language evolution new subconcepts arise incessantly, which causes the enlargement of the intension of the set of subconcepts, i.e. it diversifies. This diversification of the set of subconcepts considered for the language as a whole probably follows a stochastic birth process resulting in a probability distribution of the number of concepts in a set of subconcepts.

Considering one isolated set of subconcepts and the frequency of the use of individual concepts in it, one sees that they are not equal, but if ranked according to their magnitude they exhibit a regular trajectory that can be captured with a rank-frequency hypothesis.

In the course of language evolution not only the sets of subconcepts increase but also those of superconcepts. If a concept includes too many subconcepts then the idiolect carrier wishes to restructure the whole field so as to reduce his memory effort (= minimization of memory effort, see Köhler, 1986). A set originally with the structure



will be restructured in that new intermediate concepts are introduced e.g. in the form



The subconcepts m_i "glide" to a lower level which is thereby enriched, while on their former level the number of concepts is reduced. This process of concept formation is evidently neither chaotic nor deterministic, but follows some law of proportionality. The first investigation in this direction was performed by Martin (1974) so that the corresponding hypothesis (cf. Altmann & Kind, 1983; Schierholz, 1991; Sambor & Hammerl, 1991) is called *Martin's law.*

5. Grammatical diversification

In every natural language the words can be classified according to different criteria, e.g. morphological, syntactic, age, origin, etc. One speaks about paradigmatic, syntagmatic, grammatical categories or classes, etc.

Let us consider a morpheme or a word (combination of morphemes). It is plausible to assume that in the "beginning" (i.e. at the date of birth of such a unit) it belongs only to a single class within a single classification, e.g. to one word class (Latin "facere" is a verb) or to a class of unchangeable stems (e.g. German "sehr") or to a class of the so called basic words as in Indonesian, or only to a single morphological class because of a unique vocalization, as with the roots of Semitic languages. In the course of time an expansion or diversification can begin, the unit loses its unique class membership, it spreads into several additional classes. The cause may be the diversification originated by the speaker, or the structure of language in which such processes go on automatically, by analogy.

Two circumstances are relevant:

(1) What are the possible diversifications?
(2) What laws do they follow?

The possibility that these processes are chaotic, merely exposed to the mercy of the speaker, must be excluded since in the synergetic pattern every change sets off damping, competing or a self-regulating process.

Some frequent diversification processes are as follows:

(a) The word enlarges its class membership without any change, e.g. through conversion ("the hand", "to hand").

(b) The stem enlarges its class membership through derivation, e.g. German "Bild", "bilden", "bildhaft", or vocalization in Semitic languages (cf. Skalmowski, 1964) etc.

(c) The stem enlarges its applicability within one class through derivation, e.g. German "Blut", "Blutung", "Bluter", or through vocalization etc.

(d) The stem enlarges its applicability within one class through compounding, e.g. "Blut", "Blutdruck", "Blutdurst", etc.

(e) If a language abandons the isolating type, then morphemes diversify into several morphs because of agglutination or inflection.

(f) The word enlarges its applicability in the sentence by acquiring several functions, i.e. it enlarges its dispositional properties, which are different from the constant grammatical properties.

(g) The word enlarges its valence, above all the verbs.

(h) The word enlarges its cotextuality (cf. Köhler, 1986), i.e. its ability to occur in several contexts (where "context" can be defined in several ways). The reverse of this kind of diversification process is a part of style formation, where a "position" diversifies, i.e., a position in a given context can be filled with different units (words, sentences, etc.). This is of course not restricted to the word so that this problem does not belong here.

6. Diatopic diversification

This kind of diversification is a special case of the synonymy law from Section 3a. In a language consisting of several dialects a concept can be expressed by different forms. We find f_1 concepts having only one form in the whole area, f_2 concepts having two variants, etc. The same holds for the diastratic diversification of concepts. Measurements for dialects were performed by Goebl (1984), and it turned out that the negative binomial distribution holds here as well (cf. Altmann, 1985b). This hypothesis was called *Goebl's law.*

7. Kinds of hypotheses

In the above cases one can ask three questions, from which hypotheses arise:

(1) What is the probability distribution of the number of classes to which the words of a language belong? This encompasses all cases mentioned above. In order to set up an empirical distribution, e.g. for case (4a) one must ascertain the number f_1 of words belonging to only one word class, the number f_2 of words belonging to two word classes (simultaneously) etc. In case (4b) one must ascertain the number f_0 of words not admitting any derivation, the number f_1 of words admitting exactly one derivation etc.

(2) What is the rank-frequency distribution of the items of a single word? Empirically it means e.g. in case (4b) finding all derivations of a certain word, to ascertain their frequency of occurrence and to rank them according to their magnitude. There are several candidates for this law: Zipf-Mandelbrot's law corroborated for complete texts (cf. Orlov, Boroda & Nadarejšvili, 1982), the negative binomial distribution with rank as independent variable corroborated in semantics (cf. Beöthy & Altmann, 1984a,b; Altmann, 1985a), different distributions shown in Chitashvili & Baayen (1993) or a function shown in Altmann (1994). Case (1) concerns the language as a whole, case (2) concerns individual words.

(3) Since there are no isolated entities in language, one can ask what it depends on that a word diversifies in a given way. Evidently all diversifications are closely interconnected. It may turn out that the polysemy of a word exerts a direct influence on one of the grammatical diversifications and vice versa. The frequency of occurrence of the word is also relevant. Some of these interconnections were examined by Köhler (1986), resulting in the finding that in all cases examined there is the same interconnection, namely *Menzerath's law.* So the hypothesis is plausible that in other cases it is Menzerath's law that controls the processes as well.

If that is not the case, then one can assume that all interconnections are "enslaved" by the same principle, which in special cases will be realized by Menzerath's law, i.e. one can track down this more general law by means of the generalization of the differential equation leading to Menzerath's law.

8. Modelling

When modelling the above diversification processes one can take recourse to different mathematical means.

(a) Since we speak about processes, we can use the theory of stochastic processes (especially the birth-and-death ones) or differential equations.

(b) Since entities enlarge their class membership, urn models may turn out to be useful (cf. Krylov, 1982b).

(c) Since in these processes the needs of idiolect carriers play an important role because they cooperate and compete, one can use a kind of "synergetic" modelling where all relevant factors are set in relation to one another so that the parameters appearing in the formulas obtain an interpretation (cf. Altmann & Köhler, 1996). From this approach all the alternatives of Section 7 can be derived. Here we start from the assumption that the language community must abide by these laws because they guarantee optimal communication. Since the community is the source of these laws, it must "know" their trajectories intuitively. In order to capture them we can set up hypotheses about the relative rate of change of a variable. We equate it with a ratio of two functions, where in the numerator there are the diversifying, creative, changing forces shaping the word-norm, while in the denominator the unifying, conservative, regulating forces can be found, which are in competition with the above ones. According to the character of the variable we obtain curves or probability mass functions or densities that can be tested in the usual way. The results existing hitherto serve as positive corroboration of this kind of modelling.

References

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

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

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe (1991), 33-46.

Altmann, G. (1992). Two models for word association data. *Glottometrika 13,* 105-120.

Altmann, G. (1994). Phoneme counts. *Glottometrika 14,* 54-68.

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

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

Altmann, G., & Köhler, R. (1996). Synergetic modelling of language phenomena. In this volume.

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

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

Bunge, M. (1977). *Treatise on Basic Philosophy, Vol . 3. Ontology I: The Furniture of the World.* Dordrecht: Reidel.

Chitashvili, R.J., & Baayen, R.H. (1993). Word frequency distributions. In: Hřebíček, L., & Altmann, G. (eds.), *Quantitative Text Analysis.* Trier: WVT, 54-135.

Cramer, P. (1968). *Word association.* New York: Academic Press.

Goebl, H. (1984). *Dialektometrische Studien I-III.* Tübingen: Niemeyer.

Hřebíček, L. (1996). *Word associations and text.* In this volume.

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

Krylov, J.K. (1982a). Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: Guiter, H., & Arapov, M.V. (eds.), *Studies on Zipf's Law,* 234-262.

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

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

Palermo, D.S., & Jenkins, J.J. (1964). *Word association norms.* Minneapolis: Unversity of Minnesota Press.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Medienverlag.

Sambor, J., & Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze I.* Lüdenscheid: RAM.

Schierholz, S. (1991). *Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive.* Tübingen: Niemeyer.

Skalmowski, W. (1964). A note on the distribution of Arabic verbal roots. *Folia Orientalia 6,* 97-100.

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

Glottometrika 15, 1996, 112-133

The Theory of Word Length: Some Results and Generalizations[1]

Gejza Wimmer, Bratislava
Gabriel Altmann, Bochum

1. In a previous article (Wimmer et al., 1994) it was assumed that in simple cases word length distribution is generated by the mechanism

$$P_x = g(x)P_{x-1} \tag{1}$$

where P_x is the probability of word length x and $g(x)$ is an organizing (or proportionality) function. The formula represents a difference equation of first order. Adhering to the present trend in synergetic linguistics it has been assumed that originally $g(x) = ax^{-b}$, i.e. $g(x)$ has the form of Menzerath's law which plays an important role both in language self-regulation (cf. Köhler, 1986; Hammerl, 1991) and in hierarchical coordination of language and text levels (cf. Hřebíček, 1994). The resulting distribution has been observed in many cases (see below) but languages seem to abandon this form setting $b = 1$ or $b = 0$ and develop instead new forms modifying the remainder in some way. Though text authors or speakers are free to use words of any length, peculiarly enough, there are special distribution forms to which languages and authors adhere a long time and modify them locally rather than globally, i.e. they shift some frequencies from one class (x) to the neighbouring class (x+1 or x-1), but do not abandon the basic model, as will be shown below.

We consider statement (1) as a law-like hypothesis since it fulfills the requirements put on laws (cf. Bunge, 1967, 1983), above all generality, systematicity and confirmation. Nevertheless it is merely a skeleton that must be filled out with flesh taken from languages, genres and authors, all of them bringing different boundary and subsidiary conditions which can vary in the life of a language or of an author. Thus no *specific* formula following from (1) holds eternally for all languages or even one language. In evolution we shall meet both local and global modifications.

[1] The authors gratefully acknowledge the support of the F. Thyssen Foundation.

A preliminary examination of some hundreds of texts in twenty-one languages[2] enables us to draw some consequences, to report some results and to formulate some generalizations.

2. The evolution of word length distribution which must cope with changing conditions proceeds in the following steps:

(i) Local modification of probability classes which are either restricted or unrestricted.
(ii) Modification of the organizing (proportionality) function $g(x)$.
(iii) Modification of the (control) recursion function.
(iv) Combination of steps (ii) and (iii).
(v) Local modification of probability classes of the models resulting from steps (ii) to (iv).

These changes represent neither a temporal succession in the evolution nor are all of them necessary in a particular language. One or more steps can be skipped and the order of modifications can be changed. It is rather so that a language tends to stick to a specific model and for some time it can display only modifications of type (i). Then step by step the regularity of the model becomes destroyed - which is the usual consequence of the requirement for innovation in language; the mechanism leaves the state of equilibrium and some of the changes (ii) to (iv) follow as a necessary consequence since languages must again attain a steady state. The step towards a new steady state is effected by a new type of control or organization whose later fate cannot be predicted. After each step the process can begin anew at (i) or the process can continue at step (v) or an unpredictable change of type (ii) to (iv) can occur. Local modifications can be of two kinds:

(a) *Unrestricted,* if all frequency classes are modified by scalars whose sum is 1. They arise as a consequence of a shift happening in the given language, for example the rise of zero-syllable words in Slavic languages.

(b) *Restricted,* if only some frequency classes are affected. These modifications are usually personal fluctuations in texts. They can be of course quite characteristic for an author who brings (consciously or unconsciously) new rhythms into his/her texts and lends them a stylistic peculiarity.

Figure 1 shows the modifications observed up to now. Let us consider these steps one after another, comment on them and illustrate them with empirical material.

[2] In a project coordinated by K.-H. Best in Göttingen. All data will be published in different publications by different authors.

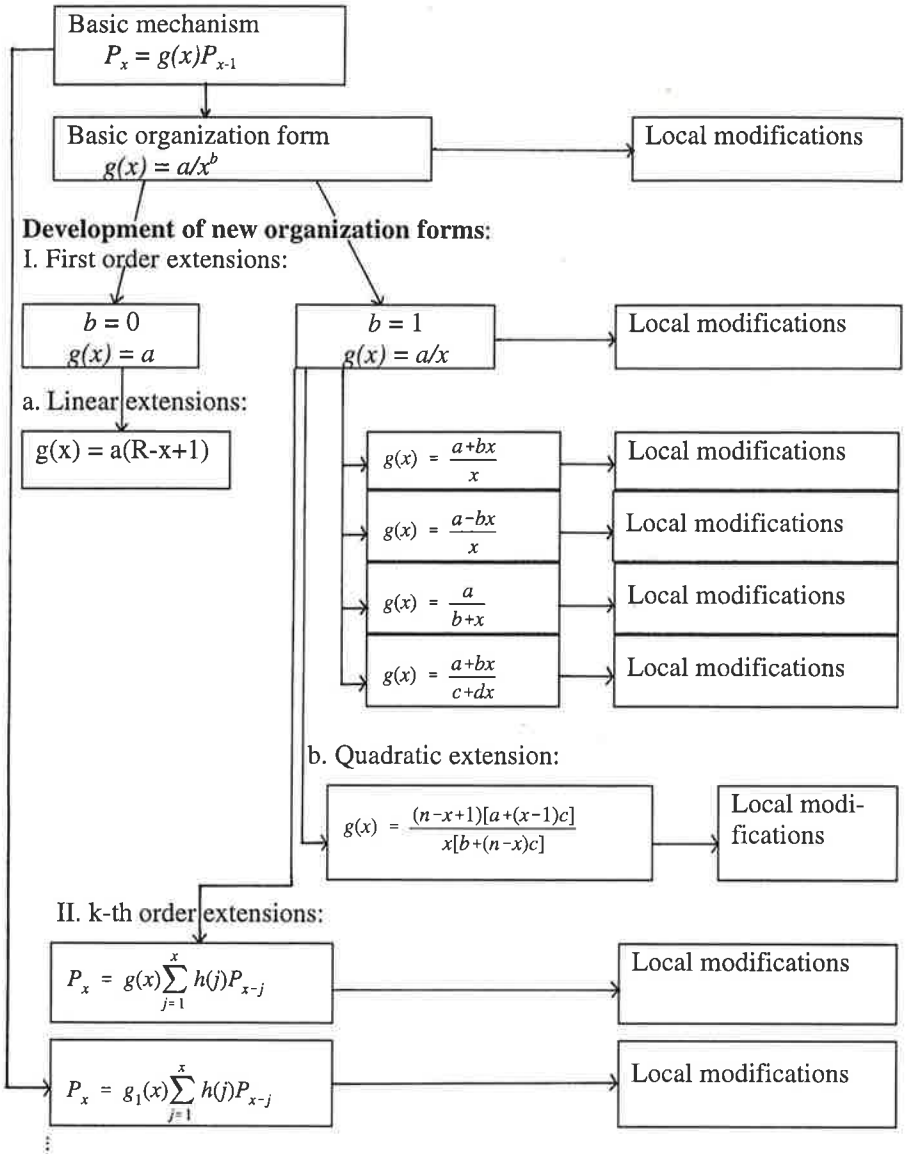


Figure 1 [$g_1(x)$ is an arbitrary proportionality function]

3. The basic form

3.1. The basic organization function is $g(x) = ax^{-b}$ leading to the *Conway-Maxwell-Poisson* distribution (cf. Conway & Maxwell, 1962; Wimmer et al., 1994) and having the form

$$(2a) \qquad P_x = \frac{a^x}{(x!)^b} C_1, \quad x = 0,1,2,...$$

if $P_0 \neq 0$. In languages not having zero-syllabic words in some cases the 1-displaced form

$$(2b) \qquad P_x = \frac{a^{x-1}}{(x-1)!^b} C_2 \quad x = 1,2,3,...,$$

but in some others the positive or zero-truncated form

$$(2c) \qquad P_x = \frac{a^x}{(x!)^b} C_3, \quad x = 1,2,3,...$$

manifests itself. C_i are norming constants. These distributions have been observed in Slovak poetry (cf. Nemcová & Altmann, 1994) and in Korean texts (cf. Kim & Altmann, 1996). In Slovak new kinds of organization have been developed for other kinds of texts (see below), but in Korean several local modifications of (2) took place, i.e. some frequency classes have been inflated at the expense of others, which have been deprived; however, the original model can be adapted to these changes. We observed the following modifications:

$$(2d) \qquad P_x' = \begin{cases} P_x' & x = 1,2,5,6,... \\ P_3 + \alpha P_4, & x = 3 \\ P_4(1-\alpha), & x = 4 \end{cases}$$

that is, a proportion α has been moved from class $x = 4$ to class $x = 3$. This modification is restricted to two classes. The other one is not restricted, and all classes are affected:

$$(2e) \quad P_x' = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1 - \alpha)P_x}{1 - P_1}, & x = 2,3,... \end{cases}$$

that is, class $x = 1$ has been inflated at the expense of all the others. The fitting of the non-modified and the modified distributions to Korean texts is shown in Kim & Altmann (1996).

3.2. A global modification takes place if the parameter b in (2a) to (2c) is set to 0 or to 1.

In the first case ($b = 0$) we obtain (for $0 < a < 1$) the geometric distribution, which has not been observed in this domain as yet. However, the linear extension $g(x) = a(R - x + 1)$ yielding

$$(3a) \quad P_x = \frac{R_{(x)}a^x}{T}, \quad x = 0,1,...,R$$

with $T = \Sigma_{j=0}^{R} R_{(j)}a^j$, $R \in \mathbb{N}$, $a > 0$, $R_{(x)} = R(R-1)...(R-x+1)$, representing the *Palm-Poisson* distribution has been found in Latin and Italian, of course, in 1-displaced form, i.e. as

$$(3b) \quad P_x = \frac{R_{(x-1)}a^{x-1}}{T}, \quad x = 1,2,...,R+1.$$

Some Italian data were published by Gaeta (1994), and others were collected by C. Hollberg; the Latin data were collected by A. Schweers and W. Röttger. In Table 1 the word length in a letter of Pliny[3] is shown.

[3] Letter 5,21 from C. Plinii Caecilii Secundi epistularum libri decem, recognovit brevique adnotatione critica instruxit R.A.B. Mynors, Oxford 1982[6].

Fitting the Palm-Poisson distribution to a letter of Pliny

x	f_x	NP_x
1	44	48.38
2	55	45.48
3	31	34.21
4	16	19.29
5	8	7.25
6	2	1.39
	$a = 0.1880$; $R = 6$, $X_3^2 = 3.62$; $P = 0.31$	

3.3. The other modification, with $b = 1$, is much more frequent in the languages examined. In the simplest case with $g(x) = a/x$ we obtain the usual *Poisson* distribution:

$$(3c) \quad P_x = \frac{e^{-a}a^x}{x!}, \quad x = 0,1,2,...$$

in its standard form, or, solving (1) with $g(x) = a/x$ and $P_0 = 0$, the *positive Poisson* distribution

$$(3d) \quad P_x = \frac{e^{-a}a^x}{x!(1 - e^{-a})} = \frac{a^x}{x!(e^a - 1)}, \quad x = 1,2,3,...$$

The positive Poisson distribution is very widespread especially in Modern German (cf. Altmann & Best, 1996); however the conditions developed so that different local modifications as well as extensions are necessary in order to get acceptable fits.

(i) The first unrestricted local modification is the *positive Pandey-Poisson* distribution (see Pandey, 1965):

$$(3e)\quad P_x = \begin{cases} \dfrac{\alpha a^x}{x!(e^a - 1)}, & x = 1,2,...,c-1,c+2,... \\ 1 - \alpha + \dfrac{\alpha a^c}{c!(e^a - 1)}, & x = c \end{cases}$$

where the affected class $x = c$ can be different in individual texts. This type can be found in German short stories. For example the result for the story "Die Rettung des Kaisers von China" by C. Wilkeshuis (sampled by B. Niehaus) from a 3rd grade school reader can be found in Table 2. In this case neither the negative binomial nor the Poisson distributions, so frequent in German, yielded a good fit.

(ii) A second modification is the *positive Singh-Poisson* distribution (cf. Singh, 1962-63)

$$(3f)\quad P_x = \begin{cases} 1 - \alpha + \dfrac{\alpha a}{e^a - 1}, & x = 1 \\ \dfrac{\alpha a^x}{x!(e^a - 1)}, & x = 2,3,4,... \end{cases}$$

Table 2
Fitting the positive Pandey-Poisson distribution to a German text

x	f_x	NP_x
1	567	567.25
2	383	382.74
3	91	90.64
4	25	22.18
5	2	5.19
$a = 0.9792$; $\alpha = 0.9017$; $c = 2$; $X_1^2 = 2.30$; $P = 0.13$		

which can be found in Italian (cf. Gaeta, 1994) and in Swedish (cf. Best, 1996). Almost all Swedish texts sampled by K.-H. Best displayed this type. An example ("Ungdomar av polisen after att ha sprängt brevlådor". From *Arbetet 1*, 22.4.93, p. 15) is shown in Table 3.

Table 3
Fitting the Singh-Poisson distribution to Swedish data

x	f_x	NP_x
1	98	99.57
2	32	31.22
3	29	22.97
4	9	12.68
5	4	5.59
6	0	2.06
7	1	0.65
8	2	0.26
$a = 2.2076$; $\alpha = 0.5927$; $X_3^2 = 3.15$; $P = 0.37$		

(iii) The third modification of this type is the *positive Cohen-Poisson distribution* (see Cohen, 1960; Haight, 1967) used in rare cases in German, e.g. in the example (Wölfel, U., "Der Nachtvogel". From: *Die grauen und die grünen Felder*. Mülheim 1970) shown in Table 4.

Table 4
Fitting the positive Cohen-Poisson distribution to German data

x	f_x	NP_x
1	272	268.90
2	162	166.65
3	25	26.07
4	6	3.38
	$a = 0.4693$; $\alpha = 0.6214$; $X_1^2 = 2.26$; $P = 0.13$	

3.4. Linear extensions

The four extensions of $g(x) = a/x$ that can be considered linear (see Figure 1) are as follows:

3.4.1. From $g(x) = (a + bx)/x$ we obtain after reparametrization ($a/b = k\text{-}1$, $b = q$) the *negative binomial distribution* in its standard form (for $P_0 \neq 0$) as

$$(4a) \quad P_x = \binom{k + x - 1}{x} p^k q^x, \quad x = 0,1,2,...$$

and for $P_0 = 0$ the *positive negative binomial* distribution

$$(4b) \quad P_x = \frac{\binom{k + x - 1}{x} p^k q^x}{1 - p^k}, \quad x = 1,2,3,...$$

The second model is the main one for German. The majority of texts either adheres to this distribution or to its limiting form, the positive Poisson distribution (see Altmann & Best, 1996; Best, 1996). One finds a great number of corroborations in the above references.

The only local modification of (4) found up to now is the *positive Cohen-negative binomial* distribution (cf. Cohen, 1960):

$$(4c) \quad P_x = \begin{cases} \dfrac{(1 - \alpha)kqp^k}{1 - p^k - \alpha kqp^k}, & x = 1 \\ \dfrac{\binom{k + x - 1}{x} p^k q^x}{1 - p^k - \alpha kqp^k}, & x = 2,3,... \end{cases}$$

found in those cases in German where neither the Poisson nor the negative binomial distributions were adequate, e.g. in the text "Wo ich wohne" from I. Aichinger (Wo ich wohne. Erzählungen, Gedichte, Dialoge. Frankfurt 1963) sampled by F. Laass and shown in Table 5.

Table 5
Fitting the positive Cohen-negative binomial distribution to a German text

x	f_x	NP_x
1	912	901.90
2	497	504.77
3	107	115.10
4	34	26.35
5	5	6.05
6	1	1.83
	$k = 0.9492$; $p = 0.7680$; $\alpha = 0.5961$ $X_2^2 = 3.56$; $P = 0.17$	

3.4.2. The second extension is characterized by $g(x) = (a - bx)/x$ from which we obtain after reparametrization [$a/b = n+1$, $n \in \mathbb{N}$, $b/(b+1) = p$] the *binomial* distribution. For $P_0 \neq 0$ the standard form,

$$(5a) \quad P_x = \binom{n}{x} p^x q^{n - x}, \quad x = 0,1,...,n$$

and for $P_0 = 0$ the *positive binomial* distribution,

$$(5b) \quad P_x = \frac{\binom{n}{x} p^x q^{n - x}}{1 - q^n}, \quad x = 1,2,...,n$$

This seems to be the main model in the Latin letters of Pliny sampled by A. Schweers and W. Röttger, and in the Turkish texts sampled by L. Hřebíček and E. Erat. An example in Table 6 shows an excellent fit of (5b) found 13 times in 18 Turkish cases[4]

[4] From the weekly *Cumhuriyet HAFTA 18.12. - 24.12.92, Ecmel Barutçu, Akilhi Bir Tercih, p. 15*

Table 6
Fitting the positive binomial distribution to a Turkish text

x	f_x	NP_x
1	185	186.26
2	299	290.31
3	272	276.52
4	168	179.58
5	94	83.97
6	28	29.08
7	8	7.55
8	1	1.73
	$n = 13$; $p = 0.2062$; $X_5^2 = 2.65$; $P = 0.75$	

(i) However, in some cases a restricted local shift of frequencies led to a modification in the form

$$(5c)\quad P_x' = \begin{cases} P_1(1 - \alpha), & x = 1 \\ P_2 + \alpha(P_1 + P_3), & x = 2 \\ P_3(1 - \alpha), & x = 3 \\ P_x, & x = 4,5,\ldots,n \end{cases}$$

showing that the class $x = 2$ has been inflated at the expense of both $x = 1$ and $x = 3$. The fitting of this model to the text Halikarnas Balikçisi, Kara Gece (in: H. Balikçisi, Gülen Ada. Istambul, Yeditepe 1957, p. 68-71) sampled by L. Hřebíček is shown in Table 7.

Table 7
Fitting the modified binomial distribution to a Turkish text

x	f_x	NP_x
1	113	109.39
2	281	292.70
3	177	170.65
4	115	117.85
5	45	39.50
6	5	7.36
7	2	0.59
	$n = 7$; $p = 0.3584$; $\alpha = 0.1911$ $X_2^2 = 1.77$; $P = 0.41$	

(ii) In Polish where there are nonsyllabic prepositions another modification seems to be appropriate. In all texts sampled by J. Sambor and W. Lehfeldt the *extended positive binomial* distribution

$$(6)\quad P_x = \begin{cases} 1 - \alpha, & x = 0 \\ \dfrac{\alpha\binom{n}{x} p^x q^{n-x}}{1 - q^n}, & x = 1,2,\ldots,n \end{cases}$$

seems to be adequate. An example[5] is shown in Table 8. The same model has been found by L. Uhlířová (1996) in Czech texts written by B. Hrabal.

3.4.3. The third extension $g(x) = a/(b+x)$ leading to the *hyper-Poisson* distribution

$$(7)\quad P_x = \frac{a^x}{b^{(x)}T}, \qquad x = 0,1,2,\ldots$$

where $b^{(x)} = b(b+1)\ldots(b+x-1)$ and T is the norming constant, has been found in

[5] "O swicie" from J.E. Kucharski, Przyniesione z lasu. Warszawa, Czytelnik 1990, p.11-12.

several languages for whole classes of texts, e.g. Italian (cf. Gaeta, 1994), in Slovak journalistic texts (cf. Nemcová & Altmann, 1994) and peculiarly, all the letters of M. Luther (16th century) displayed the 1-displaced hyper-Poisson distribution (data sampled by B. Müller). An example is shown in Table 9.

Table 8
Fitting the extended positive binomial distribution to Polish data

x	f_x	NP_x
0	6	6.00
1	70	74.99
2	85	78.86
3	51	49.14
4	17	20.10
5	5	5.63
6	2	1.28
$n = 10$; $p = 0.1894$; $\alpha = 0.9746$; $X_3^2 = 1.87$; $P = 0.60$		

Table 9
Fitting the 1-displaced hyper-Poisson distribution to a text of M. Luther[6]

x	f_x	NP_x
1	156	156.56
2	101	101.36
3	24	21.93
4	1	2.85
5	0	0.26
6	1	0.04
	$a = 0.3251$; $b = 0.5021$; $X_1^2 = 0.61$; $P = 0.44$	

Up to now no local modifications of the hyper-Poisson distribution have been found.

[6] Letter to Katharina Luther 5.6.1530 from Luthers Werke VI, Nr. 226, p. 278-279.

3.4.4. The fourth extension is $g(x) = (a + bx)/(c + dx)$ yielding the *hyper-Pascal* distribution $[a/b = k - 1,\ c/d = m - 1,\ b/d = q]$

$$(8)\quad P_x = \frac{\binom{k + x - 1}{x}}{\binom{m + x - 1}{x}} q^x C, \qquad x = 0,1,2,...$$

for which there are extensive corroborations from Slovak; however, we found it sporadically in unique cases in other languages, too.

The only "quadratic" extension found was $g(x) = \frac{(n-x+1)[a+(x-1)c]}{x[b+(n-x)c]}$ yielding the *Pólya* distribution $[a/(a+b) = p,\ b/(a+b) = q,\ c/(a+b) = s]$

$$(9)\quad P_x = \frac{\binom{-\frac{p}{s}}{x}\binom{-\frac{q}{s}}{n-x}}{\binom{-\frac{1}{s}}{n}}, \qquad x = 0,1,...,n$$

which can be observed in Finnish. Unfortunately the number of samples (made by R. Jussila and A. Vettermann) is still too small to venture the hypothesis that this is the general trend for Finnish. An example of fitting is shown in Table 10.

Table 10
Fitting the Pólya distribution to a Finnish text[7]

x	f_x	NP_x
1	85	84.12
2	142	141.38
3	139	143.49
4	114	111.73
5	76	72.45
6	41	40.56
7	17	19.91
8	7	8.60
9	5	3.26
10	0	1.07
11	1	0.43
$n = 14, p = 0.1681, s = 0.0437; X_6^2 = 2.17, P = 0.90$		

For Finnish, too, probably several locally restricted modifications of this distribution will be necessary. The phenomenon of restricted local modifications has been observed in Korean as well as in Turkish and Czech.

4. A "k-th order extension" means that the frequency in class x is controlled not only by the frequency in class x-1 but also by that of k foregoing classes. In order to enlighten the construction of hypotheses, we will show one possibility of derivation in a simple case of second order extension which is typical for French data.

As shown in Wimmer et al. (1994) the recurrence

$$(10) \quad P_x = \frac{a}{x}\sum_{j=1}^{x} h(j)P_{x-j}$$

leads to the class of generalized Poisson distributions if

$$h(j) = j\Pi_j$$

where Π_j is some probability distribution, i.e. we have

[7] Auli Koponen, Kirsu vastaa miljoonaa nenää! *Tiede 2000 8/93*, 39-41

$$(11) \quad P_x = \frac{a}{x}\sum_{j=1}^{x} j\Pi_j P_{x-j}.$$

The solution can be obtained in different ways; here we show the use of the probability generating function

$$(12) \quad G(t) = \sum_x P_x t^x$$

using two of its properties, namely

$$(13) \quad G(1) = 1$$

and

$$(14) \quad G'(t) = \sum_x x P_x t^{x-1}.$$

We can multiply (11) by xt^{x-1} and sum for all x, i.e.

$$(15) \quad \sum_{x=1}^{\infty} x P_x t^{x-1} = a\sum_{x=1}^{\infty}\sum_{j=1}^{x} j\Pi_j P_{x-j} t^{x-1}.$$

Changing the order of summation in the right side expression we obtain

$$(16) \quad G'(t) = a\sum_{j=1}^{\infty}\sum_{x=j}^{\infty} j\,\Pi_j P_{x-j} t^{x-1}$$

$$= a\sum_{j=1}^{\infty} j\,\Pi_j t^{j-1}\sum_{x=j}^{\infty} P_{x-j} t^{x-j}$$

$$= aH'(t)G(t)$$

where $H(t)$ is the probability generating function of the distribution Π_j. This can be solved easily since

$$\frac{G'(t)}{G(t)} = aH'(t)$$

yields after integration

$$\ln G(t) = aH(t) + c$$

and finally

$$(17) \quad G(t) = e^{aH(t) + c}.$$

The constant c can be determined from condition (13), i.e. letting $t = 1$ in (17) we have $1 = e^{a+c}$ which leads to $c = -a$.

Thus we obtain finally

$$(18) \quad G(t) = e^{a[H(t) - 1]}$$

which is the probability generating function of a *generalized Poisson* distribution.

Up to now only two cases of kind (10) and (18) have been observed. The simpler one is a second order extension using the "one-two" distribution (or 1-displaced Bernoulli distribution) of the form

$$(19) \quad \Pi_j = \begin{cases} 1 - \alpha, & j = 1 \\ \alpha, & j = 2 \end{cases}$$

whose probability generating function is

$$(20) \quad H(t) = (1 - \alpha)t + \alpha t^2.$$

Thus, according to (10) we have a second order difference equation:

$$P_x = \frac{a}{x}\left[(1 - \alpha)P_{x-1} + 2\alpha P_{x-2}\right]$$

which can be solved easily if we insert (20) in (18), i.e.

$$(21) \quad G(t) = e^{a[(1 - \alpha)t + \alpha t^2 - 1]}.$$

The probability distribution that can be derived from (21) e.g. through stepwise derivation according to t and setting $t = 0$, i.e.

$$\frac{1}{x!}\left[\frac{d^x G(t)}{dt^x}\right]_{t=0} = P_x$$

yields the *Hirata-Poisson* (or *Hermite)* distribution:

$$(22) \quad P_x = \sum_{i=0}^{[\frac{x}{2}]} \binom{x - i}{i} \frac{e^{-a} a^{x-i}}{(x - i)!} \alpha^i (1 - \alpha)^{x - 2i}, \quad x = 0,1,2,...$$

This distribution yields in 1-displaced form excellent results in many French texts. An example from data collected by S. Dieckmann and B. Judt is shown in Table 11.

Table 11
Fitting the Hirata-Poisson distribution to French data[8]

x	f_x	NP_x
1	426	427.25
2	168	167.83
3	88	85.12
4	28	24.80
5	5	10.00
	$a = 0.5149$; $\alpha = 0.2371$; $X_2^1 = 0.19$; $P = 0.66$	

5. An x-order extension that can be found in many languages is the *Consul-Jain-Poisson* distribution in which Π_j is the Borel distribution, i.e.

$$\Pi_j = \frac{e^{-bj} b^{j-1} j^{j-2}}{(j - 1)!}, \quad j = 1,2,...$$

Though its generating function cannot be presented in a reasonable closed form, its insertion in (11) yields

[8] Stéphane Simon, Projeté à vingt mètres. *France-Soir 7.5.1993*, p.4.

$$(23) \quad P_x = \frac{a(a + bx)^{x-1} e^{-(a+bx)}}{x!}, \quad x = 0,1,2,...$$

An example of fitting this distribution in 1-displaced form to a Maori text sampled by V. Krupa is shown in Table 14.

Table 14
Fitting the Consul-Jain-Poisson distribution to Maori data[9]

x	f_x	NP_x
1	767	770.10
2	479	470.77
3	134	139.93
4	27	26.95
5	5	4.25
	$a = 0.6062$; $b = -0.0083$[10]; $X_2^2 = 0.55$; $P = 0.76$	

6. Conclusions

Research in word length theory is in its initial phase but we hope that the basic mechanism has been found. From the results up to now some conclusions can be drawn:

(i) No language (examined) uses a unique distribution without any modification. This is in line with the fact that language develops and texts come into existence under different conditions. On the other hand, whole specific text classes tend to the same model.

(ii) Though we sometimes find the same model in cognate languages, there is no necessity that cognate languages should tend to the same model.

(iii) The modifications within a specific model are either restricted, i.e. only some classes are affected, e.g. in Korean or Turkish, or unrestricted when all

[9] Katarina Mataira, He Kai Tino Reka. From The Whare Kura. He Pitopito Kórero. Wellington, Owen 1963, p. 26.

[10] A negative b is allowed only because we are working with a slight modification setting $P_5 = 1 - \sum_{x=0}^{4} P_x$

classes are modifed by a scalar or by an appropriate function.

(iv) The modifications surpassing the boundaries of a specific model are extensions of first to x-th order. Possibly in the future one will find distributions with non-Poissonian background as shown in the last line of Figure 1 but up to now none has been observed.

(v) The variation of models can easily be interpreted in terms of attractors from catastrophe theory and the theory of dissipative systems. The stress arising incessantly in systems brings fluctuations which either must be brought under control, or the phenomenon goes away from the state of equilibrium and seeks a new steady state. This can be attained in two ways: Either it takes the nearest attractor which in our case is represented by local modifications of the probability distribution; or it takes the attractor with the highest potential, i.e. a new type of organization must arise as is shown in Figure 1.

(vi) The question arises automatically whether there is a restricted number of possible attractors for this phenomenon or whether the creativity of languages and authors is infinite. We shall, of course, never obtain a definite answer but we can at least embed this phenomenon in a net of language laws.

(vii) The theory developed so far shows the mechanism but it does not explain why a particular language prefers a special model, and even if we knew it, it would be necessary to find other phenomena affecting the values of the parameters of the pertinent models. Thus we have a long way to go, and this journey must be made hand in hand with progress in the mathematical analysis of other parts of language.

(viii) The results, as far as probability distributions are concerned, can be summarized in the following way:

Basic model: *Conway-Maxwell-Poisson d.*
Local modification: *Modified Conway-Maxwell-Poisson d.*

New organization forms:

1. $b = 0$: *Geometric d.* ($0 < a < 1$) or *right truncated geometric d.* ($a > 0$), not yet observed
   Linear extension: *Palm-Poisson d.*

2. $b = 1$: *Poisson d.*
   Local modifications:
   (i) *positive Pandey-Poisson d.*
   (ii) *positive Singh-Poisson d.*
   (iii) *positive Cohen-Poisson d.*
   (iv) *mixed Poisson d.*

First order extensions:

Linear: (a) *negative binomial d.*
Local modification: *positive Cohen-negative binomial d.*

(b) *binomial d.*
Local modifications: (i) *modified positive binomial d.*
(ii) *extended positive binomial d.*
(iii) *mixed binomial d.*

(c) *hyper-Poisson d.*

(d) *hyper-Pascal d.*

Quadratic: *Pólya d.*

Second order extension: *Hirata-Poisson d.*
x-th order extension: *Consul-Jain-Poisson d.*

References

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In this volume.
Best, K.-H. (1996). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. *Glottometrika 16* (in press).
Bunge, M. (1967). *Scientific Research I.* Berlin: Springer.
Bunge, M. (1983). *Exploring the World.* Reidel: Dordrecht.
Cohen, A.C. (1960). Estimation in the truncated Poisson distribution when zeroes and some ones are missing. *J. of the American Statistical Association 55*, 342-348.
Conway, R.W., & Maxwell, W.L. (1962). A queueing model with state dependent service rates. *J. of Industrial Engineering 12,* 132-136.
Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *Z. für empirische Textforschung 1,* 44-48.
Haight, F.A. (1967). *Handbook of the Poisson distribution.* New York: Wiley.
Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: WVT.
Hřebíček, L. (1996). *Text levels.* Trier: WVT.

Kim, I., & Altmann, G. (1996). Zur Wortlänge in koreanischen Texten. In this volume.
Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Nemcová, E., & Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Z. für empirische Textforschung 1,* 40-43.
Pandey, K.N. (1965). On generalized inflated Poisson distribution. *J. of Scientific Research of the Banaras Hindu University 15,* 157-162.
Singh, S.N. (1962-63). Inflated Poisson distribution. *J. of Scientific Research of the Banaras Hindu University 13,* 317-326.
Uhlířová, L. (1996). How long are words in Czech? In this volume.
Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *J. of Quantitative Linguistics 1,* 98-106.

Glottometrika 15, 1996, 134 - 146

How Long Are Words in Czech?

Ludmila Uhlířová, Prague

This is a question commonly asked about Czech as well as about any other language. Words consist of smaller parts, and, obviously, there are many reasons why we should like to know how many smaller parts 'fit' in them. Any question about the length of a linguistic unit at some language level and of some degree of complexity pertains to the basic questions concerning the syntagmatics of language. However, if an answer that will go beyond an elementary, 'impressionistic' level of description is expected (which, anyway, is how the question is put in the title), the answerer must cope with a whole package of items.

First, there is a crucial assumption that the length of words is not arbitrary, but that it is subjected to a more general regularity, and that the regularity - similar as it is to other linguistic units - is of a probabilistic nature. Thus, the first question from the package is: How does the probability structure of word length look like? Here some techniques from probability theory may be helpful; below, we shall try to find out whether the word length distribution in Czech may be modelled by some general type(s) of probability distribution, and which one fits best. By modelling the whole distribution and not relying on a single statistical characteristic (such as, e.g., the average, mode or median) the fact is respected that the frequencies of word classes of different lengths are proportional to each other in one way or another. Generally, the proportionality results from an interplay of a great number of factors of various kinds and force. E.g., beginning from a functional opinion that any language system is a complex paradigm from which a speaker can make choices at all levels of grammar and lexicon, then it is clear that a speaker of a strongly inflectional language has at his disposition quite another set of choices than a speaker of a strongly agglutinating or isolating language.

Another question follows from a necessary prerequisite of any serious empirical research, namely that the data cannot be presented just as 'hanging in the air', but its interpretation should be rooted in a more general framework. That is why the question of the choice of a suitable linguistic framework is pertinent. In the present study, a framework is followed which is described in Wimmer & Altmann (1996), Altmann, Erat & Hřebíček (1996), and Wimmer et al. (1994) and which is grounded in the philosophy of synergism in language: Text is a creative process,

a 'strategy' to be treated 'dynamically', 'as a means leading to a certain arrangement of elements' (Hřebíček, 1993:137; see also Hřebíček & Altmann, 1993; Köhler, 1993). The present study is a contribution to a synergetic project on word length theory coordinated by K.-H. Best (see Altmann & Best, 1996) and supplies it with data on Czech.

In accordance with the project, the word, i.e. the measured unit, is defined at the text level as any word form occurring in a running text (not as the word in its canonical form, as given in lexicons). Thus, the word-length distribution is always studied in a smaller or wider class of texts, and the results are interpreted with regard to that class.

The length of words is measured in number of syllables. In Czech, any syllabic boundary either falls inside the word, or coincides with the word boundary, with the exception of four zero-syllable prepositions *v, k, s, z* ('in', 'to', 'with', 'from'). These prepositions, if pronounced (but not so in spelling!), join the first syllable of the immediately following word, forming a special joint syllable together with it, so that, potentially, one can imagine a text in which the total of words exceeds the total of syllables. If we departed from the phonetic level of syllable and counted the total length of a text in number of syllables, we would have to take the PPs consisting of a zero-syllable preposition + noun as one 'word'. E.g., *s ním* ('with him'), pronounced as ['sňi:m], would be counted as a single one-syllabic 'word', similarly *v Praze* ('in Prague') ['fpra-ze] as a single two-syllabic 'word', and *k nádraží* ('to the station') ['kna:-dra-ži:] as a single three-syllabic 'word'. This is how Ludvíková (1985:22,153) proceeded in her pioneering statistics of word length in terms of syllables. In contrast to Ludvíková, we start from the level of word, and therefore, the zero-syllable prepositions are treated as a separate closed class of zero-syllabic words.

On the other hand, problems of syllable boundary with words containing consonant clusters are left out, because the number of syllables is not affected by placing the syllable boundary at any position within the cluster.

Let us present the data. The corpus of Czech texts consists of twenty-four short stories by a famous contemporary writer, Bohumil Hrabal[1]. The following short stories have been analyzed:

310 Expozé panu ministru informací	334 Romantici
315 Lednová povídka	335 Umělé osudy

[1] B. Hrabal, *Jarmilka*. Sebrané spisy sv. 3. Prague 1992; B. Hrabal, *Rukověť pábitelského učně*. Sebrané spisy sv. 8. Prague 1993, Pražská imaginace. My thanks are due to colleagues from the Computer Corpus of Czech who kindly lent me two diskettes with Hrabal's texts.

316 Únorová povídka
319 Blitzkrieg
323 Protokol
325 Trať číslo 23a
326 Fádní stanice
327 Křtiny
328 Očekávej mě
329 Dětský dům
330 Všední hovor
332 Symposion
338 Setkání
339 Večerní Praha
81 Hostinec u Bernardýna
82 Měsíční noc
83 Pan Metek
84 Zdivočelá kráva
85 Králíčci v křídle
89 Slavnost sněženek
817 Variace na krásnou slečnu
830 Staré noviny.

Further, six journalistic texts from the Weekend Supplement of the daily *Mladá fronta* have been analyzed:

1 Ladislav Verecký: Policie při práci. MF Dnes, Víkend, 16 April 1994
2 Ondřej Neff: Natlučené nosy těch nepravých. MF Dnes, Víkend, 23 April 1994
3 Martin Schmarcz: Protiinflační zákon. MF Dnes, Víkend, 30 April 1994
4 Martin Schmarcz: Pozitivní informace. MF Dnes, Víkend, 14 May 1994
5 Ladislav Verecký: Tajemství bezúhonnosti. MF Dnes, Víkend, 21 May 1994
6 Josef Tuček, Alena Zvěřinová: Ministerský let. MF Dnes, Víkend, 28 May 1994.

Whereas Hrabal's short stories differ in length, ranging from about 500 to 3000 running words, the journalistic texts are of about the same length, approximately 500 running words each. The latter, belonging to a short commentative genre, are regularly printed in the leftmost column of the first page under the same macroheadline *Jenže* ('However') and deal with various current topics of public life; not all texts were written by the same author.

The tests have shown that in all thirty texts (with the exception of one, see below) the word length distribution can be successfully modelled by means of the same type of probability distribution, with two local modifications in some texts.

If X is a variable denoting word length in a text consisting of N running words, and if P_x is the probability that a word will be x syllables long (= theoretical frequency of the class of words of length x) and if

$$P_0 = 1 - \alpha$$

then

$$P_x = \frac{\alpha \binom{n}{x} p^x q^{n-x}}{1 - q^n} \quad for\ x = 1,2,...,n.$$

As can be seen from Table 1 and Table 2, this type of distribution is a suitable model for Hrabal's texts 310, 315, 316, 319, 323, 327, 328, 329, 330, 334, 335, 338, 339, 82, 83, 85, 89 and 817 as well as for the journalistic texts 1, 2, 3 and 6, i.e. for the majority of texts under study, including also those of Hrabal's texts which are relatively long and for which the value of X^2 is expected to be high (that is why also the coefficient of contingency $C = X^2/N$ is given for each text).

Another four of Hrabal's texts, namely 81, 84, 325 and 326, can also be appropriately modelled by the extended positive binomial distribution, on the assumption that the word classes of length $x = 1$ and $x = 2$ are given a special treatment. The word class of the length $x = 1$ seems to be foregrounded in the structures of these texts, whereas the word class of length $x = 2$ seems to be backgrounded. Therefore, if

$$P_x' = \begin{cases} 1 - \alpha, & x = 0 \\ \alpha\,(P_1 + \beta P_2), & x = 1 \\ \alpha\, P_2(1 - \beta), & x = 2 \\ \alpha\, P_x, & x = 3,4,...,n \end{cases}$$

then we get a local modification of the extended binomial distribution; the results of the chi-square test are very good in four cases (see Table 3). (Another text, 839, is on the edge of acceptance.)

There remains one of Hrabal's texts (number 332), which fits neither the basic model, nor any modification of it. Moreover, we have not succeeded in finding *any* model of word-length distribution for this text. This leads to the hypothesis that - to follow the synergetic line of thinking - in the process of writing of the text some special circumstances, though not obvious in its 'surface' form at first glance, were at play which became responsible for the instability of its attractor.

The journalistic texts, too, follow the extended positive binomial distribution, and, also, two of them display a modification of the basic type of distribution. However, this time the modification is different from that in Hrabal's texts. Here, the foregrounded class is $x = 4$ at cost of $x = 3$, so that

$$P_x' = \begin{cases} 1 - \alpha, & x = 0 \\ \alpha\, P_x, & x = 1,2,5,6,...,n \\ \alpha\, P_3(1 - \beta), & x = 3 \\ \alpha\,(P_4 + \beta P_3), & x = 4 \end{cases}$$

Anyway, this is also a modified extended positive binomial distribution (see Table 4). It is worth noting that text 1 (fitting the extended positive binomial distribution without any additional modifications) and text 5 (fitting the local modification) were written by the same author (Verecký), and similarly, text 3 (fitting the extended positive binomial distribution without any modification) and text 4 (fitting the local modification) were written by the same author (Schmarcz).

Thus, both with Hrabal and with Verecký or Schmarcz, it shows that word length distribution in texts written by one and the same author and belonging to the same genre may be described by the same type of probability distribution *and* by a modification of that type. For the time being, not more than one modification has been found for any of the authors.

Hence, on the one hand, Czech texts by different authors - regardless of whether they pertain to the same or different genres - tend to share the same type of word-length distribution, namely the extended positive binomial distribution. On the other hand, even texts of the same genre written by the same author in the same period of time show local distributional modifications. These results are compatible with those presented by Wimmer & Altmann (1996). What seems to be important for the typology of word-length distributions in texts - at least at this tentative step of research in Czech[2] - is not mere evidence of modifications, but of the question of *which* word classes call forth the modifications. The same modification seems to refer to texts of the same genre written by the same or different authors, whereas different modifications with different β parameters seem to go across different genres and thus, perhaps, may indicate or 'measure' generic distances.

[2] Preliminary results from ten of Hrabal's stories are published in Czech (Uhlířová, 1994).

Table 1

x	Text 310		Text 315		Text 316	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	36	36.00	27	27.00	50	50.00
1	352	337.11	391	371.88	1321	1297.64
2	340	360.66	328	342.73	896	919.65
3	216	214.37	172	175.47	409	410.37
4	82	76.45	56	53.90	144	129.26
5	15	16.35	7	9.93	15	30.53
6	2	2.06	1	1.09	7	5.61
7					2	0.94
	$n = 7$; $p = 0.2629$ $\alpha = 0.9654$; $X_3^2 = 2.37$; $P = 0.50$ $C = 0.0023$		$n = 7$; $p = 0.2350$ $\alpha = 0.9725$ $X_3^2 = 2.64$; $P = 0.45$ $C = 0.0027$		$n = 19$; $p = 0.0730$ $\alpha = 0.9793$ $X_3^2 = 5.38$; $P = 0.15$ $C = 0.0019$	

x	Text 319		Text 323		Text 327	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	9	9.00	20	20.00	70	70.00
1	182	185.98	418	402.04	806	786.01
2	176	171.82	292	325.62	796	803.35
3	111	87.77	286	167.45	452	437.90
4	30	39.54	59	61.35	110	134.27
5	8	11.69	19	17.03	19	21.95
6	0	2.64	3	3.72	2	1.52
7	1	0.46	1	0.79		
8	3	0.10				
	$n = 16$; $p = 0.1097$; $\alpha = 0.9827$; $X_3^2 = 5.38$; $P = 0.15$; $C = 0.0103$		$n = 22$; $p = 0.0716$ $\alpha = 0.9800$ $X_3^2 = 6.53$; $P = 0.09$ $C = 0.0059$		$n = 6$; $p = 0.2902$ $\alpha = 0.9690$ $X_3^2 = 5.98$; $P = 0.11$ $C = 0.0026$	

Table 1 (Cont.)

x	Text 328		Text 329		Text 330	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	14	14.00	9	9.00	25	25.00
1	279	252.71	199	192.98	489	504.83
2	235	243.32	150	162.70	387	391.75
3	115	124.95	73	73.16	188	173.71
4	31	36.09	27	18.50	53	48.14
5	3	5.93	1	2.66	9	8.54
6					2	1.03
	$n = 6$; $p = 0.2780$ $\alpha = 0.9793$ $X_2^2 = 5.97$; $P = 0.05$ $C = 0.0088$		$n = 6$; $p = 0.2512$ $\alpha = 0.9803$ $X_1^2 = 3.40$; $P = 0.07$ $C = 0.0074$		$n = 8$; $p = 0.1815$ $\alpha = 0.9783$ $X_3^2 = 3.22$; $P = 0.36$ $C = 0.0028$	

x	Text 334		Text 335		Text 338	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	71	71.00	52	52.00	31	31.00
1	984	1015.28	824	842.86	677	646.08
2	815	818.07	619	634.85	442	470.93
3	452	408.05	273	265.65	203	209.77
4	138	140.91	94	66.69	71	63.71
5	26	35.68	11	10.95	13	13.93
6	7	6.84			1	2.58
7	3	1.01				
8	1	0.16				
	$n = 15$; $p = 0.1032$ $\alpha = 0.9716$ $X_2^2 = 6.79$ $P = 0.03$ $C = 0.0027$		$n = 7$; $p = 0.2007$ $\alpha = 0.9722$ $X_2^2 = 12.20$ $P = 0.002$ $C = 0.0065$		$n = 13$; $p = 0.1083$ $\alpha = 0.9784$ $X_3^2 = 5.32$ $P = 0.15$ $C = 0.0037$	

Table 1 (Cont.)

x	Text 339		Text 82		Text 83	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	70	70.00	72	72.00	97	97.00
1	1255	1203.13	1283	1281.78	1649	1664.89
2	911	973.15	945	857.27	1465	1440.71
3	447	449.79	454	423.65	777	727.25
4	149	129.93	100	123.04	183	235.99
5	20	24.02	28	24.50	42	51.95
6	1	2.98	3	3.38	8	7.36
7			1	0.35	4	0.75
	$n = 8$; $p = 0.1877$ $\alpha = 0.9755$ $X_3^2 = 11.00$ $P = 0.01$ $C = 0.0038$		$n = 10$; $p = 0.1423$ $\alpha = 0.9751$ $X_3^2 = 7.17$ $P = 0.07$ $C = 0.0025$		$n = 9$; $p = 0.1799$ $\alpha = 0.97$ $X_3^2 = 19.4$ $P = 0.0002$ $C = 0.0046$	

x	Text 85		Text 89		Text 817	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	37	37.00	68	68.00	27	27.00
1	739	717.96	1036	1005.20	492	492.70
2	690	687.21	806	852.06	445	452.95
3	363	365.43	418	401.25	256	249.84
4	100	116.59	118	113.37	89	91.87
5	16	22.32	11	19.22	31	23.64
6	4	2.49	4	1.90	2	4.34
7					1	0.66
	$n = 7$; $p = 0.2419$ $\alpha = 0.9810$ $X_3^2 = 5.72$ $P = 0.13$ $C = 0.0029$		$n = 7$; $p = 0.2203$ $\alpha = 0.9723$ $X_2^2 = 6.09$ $P = 0.05$ $C = 0.0025$		$n = 11$; $p = 0.1553$ $\alpha = 0.9799$ $X_3^2 = 3.45$ $P = 0.33$ $C = 0.0026$	

Table 2

x	Text 1		Text 2		Text 3		Text 6	
	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	25	25.00	20	20.00	13	13.00	14	14.00
1	156	156.07	17	182.52	144	138.63	150	141.31
2	165	162.23	9	181.51	147	165.16	158	174.01
3	100	104.39	19	109.39	137	116.59	123	119.04
4	59	46.50	5	44.50	49	54.02	57	48.86
5	5	15.19	95	12.87	14	17.16	8	12.03
6	3	3.76	52	3.21	4	3.78	1	1.73
7	1	0.86	9		1	0.66		
			4					
	$n = 15$ $p = 0.1293$ $\alpha = 0.9513$ $X_1^2 = 0.28$ $P = 0.60$ $C = 0.0005$		$n = 12$ $p = 0.1531$ $\alpha = 0.9639$ $X_3^2 = 5.60$ $P = 0.13$ $C = 0.0101$		$n = 10$ $p = 0.2093$ $\alpha = 0.9745$ $X_3^2 = 6.90$ $P = 0.075 \sim 0.08$ $C = 0.0136$		$n = 7$ $p = 0.2910$ $\alpha = 0.9726$ $X_3^2 = 5.16$ $P = 0.16$ $C = 0.0101$	

Table 3

x	Text 81		Text 84		Text 325	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
0	102	102.00	52	52.00	55	55.00
1	1163	1173.20	759	755.80	633	643.27
2	920	923.94	590	585.62	478	466.76
3	634	605.30	293	322.38	308	308.15
4	205	207.00	146	115.93	113	107.97
5	33	42.47	26	30.32	22	27.80
6	2	5.08	1	5.95		
7			2	1.00		
	$n = 7; p = 0.2548$ $\alpha = 0.9667$ $\beta = 0.13$ $X_2^2 = 6.49$ $P = 0.07$ $C = 0.0018$		$n = 14; p = 0.1156$ $\alpha = 0.972$ $\beta = 0.0498$ $X_2^2 = 13.38$ $P = 0.001$ $C = 0.0072$		$n = 8; p = 0.2189$ $\alpha = 0.966$ $\beta = 0.1509$ $X_1^2 = 1.91$ $P = 0.17$ $C = 0.0012$	

Table 3 (Cont.)

x	Text 326		Text 830	
	f_x	NP_x	f_x	NP_x
0	62	62.00	33	33.00
1	890	897.97	248	256.68
2	545	549.88	285	292.19
3	394	362.21	245	211.18
4	99	112.64	95	95.53
5	17	21.02	14	28.81
6	1	2.28	2	5.79
7			1	0.75
8			0	0.06
9			1	0.00
	$n = 7$; $p = 0.2372$ $\alpha = 0.9691$; $\beta = 0.2132$ $X_2^2 = 6.04$; $P = 0.05$ $C = 0.0030$		$n = 9$; $p = 0.2317$ $\alpha = 0.9643$; $\beta = 0.0264$ $X_2^2 = 14.53$; $P = 0.0007$ $C = 0.0157$	

Table 4

x	Text 4		Text 5	
	f_x	NP_x	f_x	NP_x
0	18	18.00	9	9.00
1	129	126.21	147	154.95
2	143	144.46	156	145.52
3	73	76.18	75	72.75
4	72	70.95	31	35.04
5	16	15.63	6	6.59
6	4	3.79	1	1.14
7	0	0.68		
8	0	0.09		
9	1	0.01		
	$n = 12$; $p = 0.1723$ $\alpha = 0.9605$; $\beta = 0.2398$ $X_2^2 = 0.28$; $P = 0.87$ $C = 0.0014$		$n = 8$; $p = 0.1901$ $\alpha = 0.9159$; $\beta = 0.0874$ $X_2^2 = 1.77$; $P = 0.41$ $C = 0.0165$	

References

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In this volume.

Altmann, G., Erat, E., & Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. In this volume.

Hřebíček, L. (1993). Text as a strategic process. In: Hřebíček, L., & Altmann, G. (eds.), *Quantitative text analysis.* Trier: WVT, 136-150.

Hřebíček, L., & Altmann, G. (1993). Prospects of text linguistics. In: Hřebíček, L., & Altmann, G. (eds.), *Quantitative text analysis.* Trier: WVT, 1-28.

Köhler, R. (1993). Synergetic linguistics. In: Köhler, R., & Rieger, B. (eds.), *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier 1991.* Dordrecht: Kluwer, 41-51.

Ludvíková, M. (1985). Kvantitativní charakteristiky českých fonémů. In: Tešitelová, M. et al. *Kvantitativní charakteristiky současné češtiny.* Praha: Academia, 11-28.

Uhlířová, L. (1994). O jednom modelu rozložení délky slov. *Slovo a slovesnost (in press).*

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In this volume.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1,* 98-106.

Glottometrika 15, 1996, 147-157

Zur Wortlängenhäufigkeit in schwedischen Pressetexten

Karl-Heinz Best, Göttingen

0. Ein naheliegendes und oft auch leicht zugängliches Beobachtungsfeld für den Linguisten ist die Untersuchung der Häufigkeit, mit der bestimmte Einheiten oder Eigenschaften einer Sprache in ihrem System oder in ihren Texten vorkommen. Hierzu bieten sich nun Untersuchungen im Bereich der Laute (Phoneme), Wörter und Sätze als den vertrautesten Sprachphänomenen besonders an. So braucht es nicht zu verwundern, daß Sprachwissenschaftler sich schon seit langem direkt (in Form des Syntheseindex: Greenberg, 1960) oder indirekt (in Form der morphologisch orientierten Sprachtypologie schon des 19. Jhds.: vgl. Horne, 1966:17f.) mit Wortlängen in verschiedenen Sprachen befaßten. Dabei hatte der Syntheseindex die Funktion, Sprachen hinsichtlich der durchschnittlichen Wortlängen in ihren Texten zu charakterisieren und miteinander zu vergleichen. Außerdem ergab sich damit die Möglichkeit, die Interaktion zwischen der Wortlänge und anderen, zumeist morphologischen Eigenschaften von Sprachen zu untersuchen (Krupa & Altmann, 1966). Die Fruchtbarkeit dieses typologischen Ansatzes erweist sich in kontinuierlichen Forschungen bis in die jüngste Zeit (vgl. Silnitsky, 1993). Problematisch ist allerdings die Praxis, die morphologischen Indizes für jede Sprache auf der Basis eines Textausschnitts von nur je ca. 100 Wörtern zu bestimmen; der stilistischen Variabilität der Sprachen wird man damit sicher nicht annähernd gerecht.

Eine ganz zentrale Rolle spielt die Wortlänge auch bei den Versuchen, Sprache als ein selbstregulierendes System zu modellieren. So sieht Köhler (1986:74) die Wortlänge im Zentrum seines Basismodells der Lexik. Sie wird beeinflußt von der Frequenz, der Lexikongröße (d.h. der Zahl der Lexeme im Inventar) und der Phonemzahl (d.h. der Anzahl der Phoneme im Phonemsystem der betreffenden Sprache). Andererseits aber wirkt sich die Wortlänge auf die Polylexie (d.h. die Anzahl der Bedeutungen pro Lexem) aus. Hammerl hat auf der Grundlage seiner Forschungen zur Lexik des Polnischen Modifikationen des Köhlerschen Modells vorgenommen, sieht aber auch die Wortlänge als eine zentrale Größe an (Hammerl, 1991:218f.).

Es gibt eine Vielzahl weiterer Forschungsansätze, die sich mit der Wortlänge

befassen oder bei denen Wortlängen wenigstens einen wichtigen Teilaspekt darstellen; sie reichen von der Analyse des literarischen Stils (Fucks & Lauter, 1965) über den Ausdruck der Höflichkeit (Ogino, 1989) bis zur Interaktion zwischen Konstrukt und Konstituente, formuliert im sog. Menzerathschen Gesetz, wobei die Wortlänge sowohl als Konstruktlänge als auch als Konstituentenlänge betrachtet werden kann (Altmann & Schwibbe, 1989:46ff.). Es wird niemand wundern, daß die Wortlänge auch die Lesegeschwindigkeit beeinflußt (Pierce, 1965:265f.). Weniger naheliegend dürfte der Hinweis S. Freuds (1992:37) sein: "Bekanntlich neigt man gerade bei einsilbigen Familiennamen besonders dazu, den Vornamen mitzunennen."

Das Panorama der Zusammenhänge, bei denen die Wortlänge eine wichtige Rolle spielt, ist damit längst nicht ausgeschöpft; für weitere Aspekte vgl. Best & Zhu (1994:19) und Best (1994).

1. In dieser Untersuchung geht es um einen ganz speziellen Aspekt, um die Frage nämlich, mit welcher Häufigkeit Wörter verschiedener Länge in schwedischen Pressetexten vorkommen. Grundlage der Untersuchung ist die Annahme, daß Wortlängenhäufigkeiten nicht chaotisch auftreten, sondern bestimmten Gesetzmäßigkeiten folgen. Für diese Annahme existieren inzwischen gute Bestätigungen zu einer ganzen Reihe von Sprachen, darunter für Deutsch (u.a. Grotjahn, 1982; Altmann & Best, 1996); Italienisch (Gaeta, 1994) und Slowakisch (Nemcová & Altmann, 1994). Weitere Untersuchungen werden in einem Projekt zur "Wortlängenhäufigkeit" derzeit durchgeführt. Mehrere Resultate findet man in diesem Band. In allen Fällen liegen gute Ergebnisse vor. Es gibt also hinreichend Grund zu der Annahme, daß dies auch im Schwedischen gelten sollte. Theoretische Begründungen findet man in Altmann & Best (1996), Wimmer et al. (1994) und Wimmer & Altmann (1996).

2. Die Untersuchung zum Schwedischen wurde wie folgt durchgeführt: Die Auswahl der Texte ergab sich relativ zufällig und willkürlich. Kriterium war lediglich, daß es sich um Texte aus nur einem Funktionalstil (Fleischer & Michel, 1977:243ff.) handeln sollte; da bei Untersuchungen zu anderen Sprachen Pressetexte mit guten Ergebnissen bearbeitet worden waren, konnte auch für diese Arbeit auf solche Texte zurückgegriffen werden, zumal diese relativ leicht zugänglich sind. Es wurden nur solche Texte ausgewählt, die von dem jeweiligen Presseorgan selbst verantwortet werden, also nicht etwa Texte Außenstehender wie Anzeigenkunden, Behörden oder dgl.

Ein weiteres Kriterium bestand darin, daß es hinreichend viele Texte eines Funktionalstils sein sollten, um entscheiden zu können, welche der vielen denkbaren Modelle für die Verteilung der Wortlängen in schwedischen Texten am ehe-

sten infrage kommen. Erfahrungsgemäß genügen dazu ca. 10-20 Texte einer Textklasse. Es gibt Hinweise darauf, daß Texte verschiedener Funktionalstile/ Textsorten innerhalb einer Sprache unterschiedlichen Modellen folgen können, wie z.B. die Untersuchungen zum Slowakischen gezeigt haben (Nemcová & Altmann, 1994). Deshalb sollten die Texte auch nicht aus unterschiedlichen Stilen stammen. Eine Untersuchung schwedischer Gedichte, Prosa oder anderer Textklassen müßte also ggfs. gesondert erfolgen; dabei kann sich herausstellen, daß in diesen Fällen andere Modelle gelten als hier gefundenen.

3. Bei der Bearbeitung der Texte wurde wie folgt verfahren: Berücksichtigt wurden nur die Überschriften (auch Zwischenüberschriften), der Lead und der laufende Text. Alle weiteren Textbestandteile wie Autor, Ort, Datum etc. blieben unberücksichtigt. Als "Wort" wurde wie in den anderen Arbeiten das "orthographische Wort" (Bünting & Bergenholtz, 1989^2:36f.) gewählt; es läßt sich besonders leicht definieren und hat sich bis auf wenige Ausnahmen bisher gut bewährt. Die Silbenzahl pro Wort wurde nach der Zahl der im Wort vorkommenden Vokale bestimmt. In Zweifelsfällen wurden einschlägige Handbücher (z.B. Prisma. Handwörterbuch Schwedisch - Deutsch. 1992) zu Rate gezogen.

Ein besonderes Problem stellen in Pressetexten die Zahlwörter, speziell die Jahreszahlen dar. In dieser Untersuchung wurden sie folgendermaßen bearbeitet: Jahreszahlen wurden wie alle anderen Zahlen auch entsprechend den Angaben in Ritte (1986:31) behandelt, also "1992" = gesprochen "nittonhundranittiotwå" (1 Wort, 8 Silben). Es gibt inzwischen Erfahrungen damit, daß Zahlwörter in manchen Textsorten (wissenschaftliche Texte, Pressetexte etc.) die Häufigkeitsverhältnisse gelegentlich derart entstellen, daß die Verteilungen sich nur mit komplizierten Modifikationen modellieren lassen. Aus diesem Grund, und weil in vielen Sprachen die als Wörter, nicht als Zahlzeichen realisierten Zahlwörter orthographisch sehr verschieden behandelt werden (z.B. frz. "1993": "mil neuf cent quatre-vingt-treize"), liegt es nahe, Zahlwörter abweichend von den orthographischen Konventionen als mehrwortige Syntagmen aufzufassen. In dieser Untersuchung wurde von einer entsprechenden Datenrevision abgesehen, weil Zahlwörter in den bearbeiteten Texten nur vereinzelt auftraten und daher keine solchen "Störungen" verursachten.

Die Zehnerzahlen des Schwedischen (z.B. "trettio/tretti") wurden in der längeren Version aufgenommen; also: "trettio" (1 Wort, 3 Silben). Der angehängte Artikel (z.B. *huset* = das Haus) wurde als Wortteil, nicht als eigenständiges Wort behandelt. Abkürzungen wurden in ihre Vollform aufgelöst und entsprechend gewertet (z.B. "bl.a." = "bland annat": 2 Wörter mit 1 bzw. 2 Silben).

4. Die 18 schwedischen Pressetexte wurden daraufhin überprüft, mit welchen Wahrscheinlichkeitsfunktionen sie am besten modelliert werden können; es zeigte sich, daß alle der positiven Singh-Poisson-Verteilung folgen, einige aber noch besser mit der 1-verschobenen Consul-Jain-Poisson-Verteilung erfaßt werden können. Die Formeln für die beiden Verteilungen lauten:

Positive Singh-Poisson-Verteilung:

$$P_x = \begin{cases} 1 - \alpha + \dfrac{\alpha a e^{-a}}{1 - e^{-a}}, & x = 1 \\ \dfrac{\alpha a^x e^{-a}}{x!(1 - e^{-a})}, & x = 2,3,4,... \end{cases}$$

$$a > 0$$
$$0 < \alpha < 1$$

Consul-Jain-Poisson-Verteilung (in 1-verschobener Form):

$$P_x = \frac{a[a + b(x - 1)]^{x-2} e^{-[a + b(x-1)]}}{(x - 1)!}, \quad x = 1,2,3,...$$

$$a, b > 0$$

Als Kriterium der Anpassungsgüte benutzten wir den Chiquadrattest. Eine Anpassung betrachten wir als zufriedenstellend, wenn $P \geq 0.05$, wobei P die Überschreitungswahrscheinlichkeit des Chiquadrats darstellt. Bei umfangreichen Daten versagt dieses Kriterium des öfteren, weil das Chiquadrat mit der Stichprobengröße arithmetisch wächst. In solchen Fällen benutzten wir lediglich $C = X^2/N$, was ein Diskrepanzmaß darstellt, bei dem die Freiheitsgrade unberücksichtigt bleiben; als zufriedenstellend wird $C \leq 0.02$ betrachtet. a, b und α sind Parameter. Der Index beim Chiquadrat gibt die Zahl der Freiheitsgrade an. x ist die Zahl der Silben im Wort, n_x die Zahl der Wörter mit Silbenzahl x im Text. NP_{xSP} gibt die theoretische Häufigkeit der Singh-Poisson-Verteilung an, NP_{xCJ} die der Consul-Jain-Poisson-Verteilung.

5. Die Resultate der Untersuchung sind in Tabelle 1 dargestellt. Es wurden folgende Texte benutzt:

1 *Polisens fel att tjuv blev av med löständernä?* (Verlor ein Dieb seine falschen Zähne wegen eines Dienstvergehens der Polizei?). Arbetet 1, 22.4.93, S. 3, Regionales aus Schonen; Bericht
2 *Patienters rättigheter måste stärkas* (Patientenrechte müssen gestärkt werden). Arbetet 1, 22.4.93, S.3, Regionales aus Schonen; Bericht
3 Katarina Ström: *Alnarp lär ut trädgårdstips* (Alnarp gibt Gartentips). Arbetet 1, 22.4.93, S. 11, Regionales aus Malmö; Bericht
4 *Förlorade jobbet trots friande dom* (Stelle verloren trotz Freispruch). Arbetet 1, 22.4.93, S. 15, Regionales aus Lund; Bericht. Der Text enthält Zwischenüberschriften, die mitgezählt wurden.
5 *Ungdomar greps av polisen efter att ha sprängt brevlådor* (Jugendliche von Polizei nach Sprengung von Briefkästen aufgegriffen). Arbetet 1, 22.4.93, S. 15, Regionales aus Lund; Bericht
6 *Rena slavjobbet på krogen?* (Reine Sklavenarbeit im Wirtshaus?). Arbetet 1, 22.4.93, S. 15, Regionales aus Lund; Bericht
7 Rode Möller: *Utvecklingsstörda kan fara illa i vanliga skolan* (Entwicklungsgestörte können in der normalen Schule Schaden nehmen). Arbetet 1, 22.4. 93, S. 16, Regionales aus Schonen; Bericht
8 Ingalill Löfgren: *Studenthemmets huvudnyckel stals* (Hauptschlüssel des Studentenwohnheims gestohlen). Göteborgs-Posten, 2.10.93, S.7, Bericht
9 *Fyra skadades i trafikolycka* (Vier Verletzte bei Verkehrsunfall). Göteborgs-Posten, 2.10.93, S .6, Bericht
10 *Första spadtaget för äldrecentrum* (Erster Spatenstich für Altenzentrum). Göteborgs-Posten, 2.10.93, S. 12, Bericht
11 *Syncentral invigd* (Augenzentrum eröffnet). Göteborgs-Posten, 2.10.93, S. 12, Bericht
12 Thomas Höjeberg: *Globalt miljardstöd till palestinskt självstyre* (Milliardenunterstützung aus aller Welt für palestinensische Selbstverwaltung). Göteborgs-Posten, 2.10.93, S. 14, Bericht
13 Helge Ögrim: *Greenpeace ber Clinton straffa Norge* (Greenpeace fordert Clinton auf, Norwegen zu bestrafen). Göteborgs-Posten, 2.10.93, S. 14, Bericht
14 *Onödig strid om EG-omröstning* (Überflüssiger Streit über EG-Abstimmung). Göteborgs-Posten, 2.10.93, S. 2, Kommentar
15 *Gamla svenskar mår oftast bra* (Alte Schweden fühlen sich meistens wohl). Göteborgs-Posten, 2.10.93, S. 6, Bericht

16 *Eftergifter tas tillbaka* (Zugeständnisse werden zurückgenommen). Göteborgs-Posten, 2.10.93, S. 14, Bericht
17 Karin Zillén: *Orimligt bara pröva halva bron* (Es ist absurd, nur die halbe Brücke zu prüfen). Göteborgs-Posten, 2.10.93, S. 18, Bericht
18 Marit Larsdotter: *Hexkonst uppåt väggarna* (Hexenkunst ganz verrückt). Göteborgs-Posten, 2.10.93, S. 26, Bericht über eine Kunstausstellung.

Tabelle 1

x	Text 1		Text 2		Text 3	
	n_x	NP_{xSP}	n_x	NP_{xSP}	n_x	NP_{xSP}
1	94	95.07	33	31.77	80	79.68
2	32	29.83	12	13.18	42	40.53
3	19	20.03	11	12.64	22	25.77
4	11	10.09	14	9.09	15	12.29
5	2	4.06 |	5	5.23	2	4.69 |
6	1	1.36 |	0	2.51	2	1.49 |
7	0	0.39 |	0	1.03 |	1	0.40 |
8	2	0.17 |	1	0.55 |	0	0.09 |
9					1	0.06 |
	$a = 2.015$; $\alpha = 0.593$ $X_2^2 = 0.46$; $P = 0.80$		$a = 2.878$; $\alpha = 0.702$ $X_4^2 = 5.74$; $P = 0.22$		$a = 1.908$; $\alpha = 0.775$ $X_2^2 = 1.28$; $P = 0.53$	

Tabelle 1(Fortsetzung)

x	Text 4			Text 5		Text 6		
	n_x	NP_{xSP}	NP_{xCJ}	n_x	NP_{xSP}	n_x	NP_{xSP}	NP_{xCJ}
1	129	131.88	119.12	98	99.57	303	301.79	306.06
2	76	77.86	81.93	32	31.22	188	179.11	171.37
3	40	45.93	43.73	29	22.97	61	77.09	71.52
4	16	20.32	21.58	9	12.68	20	24.88 |	26.98
5	19	7.19 |	10.34 |	4	5.59	12	6.42 |	9.75
6	6	2.82 |	9.30 |	0	2.06 |	6	1.38 |	3.45
7				1	0.65 |	1	0.33 |	1.87
8				2	0.26 |			
	$a = 1.770$ $\alpha = 0.847$ $X_1^2 = 4.63$ $P = 0.03$ $C = 0.016$		$a = 0.876$ $b = 0.242$ $X_2^2 = 4.48$ $P = 0.11$	$a = 2.208$ $\alpha = 0.593$ $X_3^2 = 3.15$ $P = 0.37$		$a = 1.291$ $\alpha = 0.959$ $X_1^2 = 4.90$ $P = 0.03$ $C = 0.0082$		$a = 0.658$ $b = 0.161$ $X_4^2 = 7.80$ $P = 0.10$

x	Text 7		Text 8		Text 9	
	n_x	NP_{xSP}	n_x	NP_{xSP}	n_x	NP_{xSP}
1	176	174.60	106	105.59	41	41.83
2	94	90.09	63	63.78	23	19.81
3	64	64.12	36	36.52	7	10.87
4	24	34.22	17	15.68 |	2	4.47 |
5	18	14.61	3	5.38 |	5	1.47 |
6	3	5.20 |	1	1.54 |	0	0.40 |
7	4	1.58 |	3	0.51 |	1	0.15 |
8	2	0.58 |				
	$a = 2.135$; $\alpha = 0.766$ $X_3^2 = 4.40$; $P = 0.22$		$a = 1.718$; $\alpha = 0.863$ $X_2^2 = 0.15$; $P = 0.93$		$a = 1.647$; $\alpha = 0.775$ $X_1^2 = 2.27$; $P = 0.13$	

Tabelle 1 (Fortsetzung)

x	Text 10		Text 11		Text 12	
	n_x	NP_{xSP}	n_x	NP_{xSP}	n_x	NP_{xSP}
1	28	27.90	31	30.75	122	119.14
2	14	14.16	14	12.83	75	72.36
3	14	11.74	14	12.72	53	57.44
4	6	7.29	8	9.45	27	34.20
5	2	3.63	4	5.62	14	16.29
6	2	1.50	2	2.78	16	6.46
7	1	0.78	0	1.18	2	3.11
8			1	0.44		
9			2	0.23		
	$a = 2.487$; $\alpha = 0.754$ $X_1^2 = 1.63$; $P = 0.65$		$a = 2.974$; $\alpha = 0.709$ $X_4^2 = 1.90$; $P = 0.75$		$a = 2.382$; $\alpha = 0.811$ $X_2^2 = 3.49$; $P = 0.17$	

x	Text 13		Text 14		Text 15		
	n_x	NP_{xSP}	n_x	NP_{xSP}	n_x	NP_{xSP}	NP_{xCJ}
1	132	137.13	94	92.28	48	47.30	50.87
2	53	47.55	35	31.65	38	34.41	32.51
3	27	29.08	18	25.04	12	18.53	15.70
4	11	13.33	14	14.85	5	7.48	6.91
5	5	4.89	10	7.05	7	2.42	2.92
6	4	1.49	0	2.79	0	0.65	1.21
7	1	0.39	0	0.94	0	0.15	0.50
8	1	0.14	4	0.40	1	0.06	0.38
	$a = 1.834$; $\alpha = 0.636$ $X_2^2 = 3.82$; $P = 0.15$		$a = 2.373$; $\alpha = 0.625$ $X_3^2 = 3.65$; $P = 0.30$		$a = 1.616$ $a = 0.780$ $\alpha = 0.958$ $b = 0.199$ $X_1^2 = 3.16$ $X_1^2 = 2.06$ $P = 0.08$ $P = 0.15$		

Tabelle 1 (Fortsetzung)

x	Text 16		Text 17			Text 18	
	n_x	NP_{xSP}	n_x	NP_{xSP}	NP_{xCJ}	n_x	NP_{xSP}
1	109	108.65	248	241.84	235.13	228	229.45
2	58	57.82	114	106.91	134.87	122	116.36
3	41	40.94	73	79.22	67.17	47	49.04
4	20	21.74	32	44.03	32.76	19	19.82
5	12	9.23	17	19.57	16.04	5	7.94
6	2	3.27	9	7.25	7.93	4	3.19
7	0	0.99	7	2.30	3.96	2	1.28
8	1	0.36	1	0.64	2.00	1	0.92
9			1	0.24	2.14		
	$a = 2.124$; $\alpha = 0.777$ $X_4^2 = 1.55$; $P = 0.82$		$a = 2.223$ $a = 0.759$ $\alpha = 0.710$ $b = 0.279$ $X_3^2 = 10.30$ $X_5^2 = 5.94$ $P = 0.02$ $P = 0.31$ $C = 0.0201$			$a = 0.623$ $\alpha = 0.206$ $X_4^2 = 2.01$ $P = 0.73$	

6. Als Ergebnis dieser Untersuchung kann festgestellt werden: Das grundlegende Modell für die Verteilung der Wortlänge in schwedischen Pressetexten ist die positive Singh-Poisson-Verteilung. Sie bewährt sich für Texte 1-16 und 18; Text 17 läßt sich mit ihr gerade noch modellieren. Für die Texte 4, 6, 15 und 17 stellt aber die 1-verschobene Form der Consul-Jain-Poisson-Verteilung ein besseres Modell dar. Hier könnte sich eine Änderung in der Strukturierung schwedischer Pressetexte andeuten. Text 17 spielt insofern eine besondere Rolle, als aus inhaltlichen Gründen (Bericht über den Bau der Öresundbrücke) vielsilbige Wörter wie "Öresundskontoriet" gehäuft auftreten und so zu einem höheren Anteil längerer Wörter führen, als man sonst erwarten würde.

Literatur

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In diesem Band.

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten (erscheint).

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, U. (Hrsg.), *Computatio Linguae II.* Stuttgart: Steiner, 19-30.

Bünting, K.-D., & Bergenholtz, H. (1989[2]). *Einführung in die Syntax.* Frankfurt: Athenäum.

Fleischer, W., & Michel, G. (1977[2]). *Stilistik der deutschen Gegenwartssprache.* Leipzig: VEB Bibliographisches Institut.

Freud, S. (1904/1992). *Zur Psychopathologie des Alltagslebens.* Frankfurt: Fischer.

Fucks, W., & Lauter, J. (1965; 1971[4]). Mathematische Analyse des literarischen Stils. In: Gunzenhäuser, R., & Kreuzer, H. (Hgg.), *Mathematik und Dichtung.* München: Nymphenburger, 107-122.

Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *Zeitschrift für empirische Textforschung 1,* 44-48.

Greenberg, J.H. (1960). A quantitative Approach to the Morphological Typology of Languages. *International Journal of American Linguistics 26,* 178-194.

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

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.

Horne, K.M. (1966). *A Critical Evaluation of Morphological Typology. With Particular Emphasis on Greenberg's Quantitative Approach as Applied to Three Historic Stages of German.* Washington D.C., Diss.phil. (University Microfilms International).

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

Krupa, V., & Altmann, G. (1966). Relations between Typological Indices. *Linguistics 24,* 29-37.

Nemcová, E., & Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirische Textforschung 1,* 40-43.

Ogino, T. (1989). Length and Politeness of Honorific Expressions. In: Mizutani, Sh. (ed.), *Japanese Quantitative Linguistics.* Bochum: Brockmeyer, 188-199.

Pierce, J.R. (1965). *Phänomene der Kommunikation.* Düsseldorf-Wien: Econ.

Prisma. Handwörterbuch Schwedisch - Deutsch. Völlige Neubearbeitung. Berlin: Langenscheidt 1992.

Ritte, H. (1986). *Schwedische Grammatik.* München: Hueber.

Silnitsky, G. (1993). Typological Indices and Language Classes: A Quantitative Study. *Glottometrika 14,* 139-160.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In diesem Band.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1,* 98-106.

Glottometrika 15, 1996, 158-165

Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen

Sandra Dieckmann, Göttingen
Birga Judt, Göttingen

0. In dieser Untersuchung geht es um die Frage, ob die Häufigkeit, mit der Wörter unterschiedlicher Länge in Texten verwendet werden, auch bei weiterem französischen Sprachmaterial der Hirata-Poisson-Verteilung folgt. Dieses Modell wurde bereits in Feldt, Janssen & Kuleisa (1994) für französische Briefe und einige Pressetexte vorgeschlagen. Gegenstand der Untersuchung sind weitere Pressetexte und, als kleiner Ausblick, auch zwei literarische Prosatexte.

An dieser Stelle sei auf eine vergleichende Untersuchung zur Wortlänge hingewiesen, in der Fucks (1968) auf einer breiten Datenbasis 11 Sprachen, darunter das Französische, miteinander vergleicht. Französisch erwies sich dabei nach dem Englischen als eine Sprache mit besonders starker Tendenz zu einsilbigen Wörtern. Während W. Fucks die Vergleiche mit Hilfe des Mittelwertes durchführt, werden hier die Daten der analysierten Texte einzeln in Form von Häufigkeitsverteilungen vorgestellt. Das Ziel von Fucks war Charakterisierung und Vergleich, unser Ziel ist die Überprüfung eines Modells.

1. Die Einheit "Wort" wird als "orthographisches Wort" bestimmt: "Là où le linguiste part d'un texte écrit, il peut accepter si cela lui convient, pour définir une unité 'mot', les critères de l'orthographie: *un mot dans l'écriture* est un segment séparé des autres segments par des espaces blancs" (Martinet, 1969:252f.). Damit wird das gleiche Kriterium für "Wort" verwendet, das auch Feldt, Janssen & Kuleisa (1994) angewendet haben.

Zur Bestimmung von "Silbe" bezieht sich Lausberg auf den Begriff des "Schallfüllegrades", wobei jeder Laut eine ihm eigene Ausprägung aufweise, und führt aus: "Die Abfolge der einzelnen Schallfüllegrade gleicht nun einer Berg- und Talfahrt: Es gibt Schallfüllegipfel und Schallfülletäler" (Lausberg, 1956: 62). Und weiter: "Man pflegt nun den Schallfüllegipfel zusammen mit den zu ihm gehörigen Talbereichen 'Silbe' zu nennen. Der Schallfüllegipfel selbst ist als 'Silbengipfel'

Hauptträger der Silbe" (Lausberg, 1956:63). Als solche Silbengipfel werden Vokale und Diphthonge aufgefaßt; in Zweifelsfällen werden Aussprachewörterbücher zu Rate gezogen.

2. Bei der Datenaufnahme wurden folgende Konventionen angenommen:
- Es wurden die orthographischen Wörter in ihrer gesprochenen Form ausgewertet. Die Halbkonsonanten (z.B. in *lui, moi, yeux*) wurden nicht als Silbenträger gewertet.
- Das "e-instable" wird je nach Umgebung gesprochen oder es bleibt stumm (vgl. Klein, 1963:91f.).
- Apostrophierte Personalpronomen, Artikel und apostrophiertes *ne* (= *n'*) werden als nullsilbig gewertet. Aber *aujourd'hui:* 1 Wort.
- *t-* in *a-t-il* zählt nicht als Wort, da es eine rein phonetische Einheit ist.
- *Dessous* wird hier als zweisilbig gewertet.
- 1992: mille neuf cents quatre-vingt-douze: 4 Wörter.
- *58% = cinquante-huit pour cent:* 3 Wörter.
- *100km/h = cent kilomètres par heure*: 4 Wörter.
- $3h^{03}$ = *trois heures trois*: 3 Wörter.
- XIX^{e} = *dix-neuvième:* 1 Wort.
- Abkürzungen, wie z.B. *CGS (contribution sociale généralisée), PS (Parti socialiste)* usw. gelten als 1 Wort.
- 1984-1985: 9 Wörter.
- Ein Problem bilden die apostrophierten, dadurch nullsilbigen Wörter wie *d'*, etwa in der syntaktischen Verbindung *d'un*, die im Französischen relativ häufig sind. Es hat sich bei vorläufigen Modellierungsversuchen gezeigt, daß man die Texte sowohl mit als auch ohne diese nullsilbigen Wörter modellieren kann und in beiden Fällen die Hirata-Poisson-Verteilung akzeptabel ist. Um auch hierbei eine möglichst einheitliche Datenerhebung zu präsentieren, werdem wie in Feldt, Janssen & Kuleisa (1994) keine nullsilbigen Wörter in die Datentabellen aufgenommen; sie werden vielmehr als phonetische Bestandteile ihrer Nachbarwörter aufgefaßt.
- Für Zweifelsfälle wurde Martinet & Walter (1973) zu Rate gezogen.
- In allen Texten wurde der Autorenname nicht mitgezählt.
- Die Überschriften wurden in Texten 1-8 mitgezählt, sonst nicht.
- Ortsangaben wurden nicht mitgezählt.
- Der in der Mitte eingefügte, aber durch eigene Überschrift und Fettdruck abgesetzte Text wurde nicht mitgezählt.

3. Für die Untersuchung wurden folgende Texte gewählt:

T1 - Stéphane Simon: *Projeté a vingt mètres*. France-Soir 7.5.1993, S.4.
T2 - Marie-Amélie Lombard: *Feu vert pour les 'descentes de police'*. Figaro 20.5.1993, S.6.
T3 - Pierre Bocev: *La lutte politique en Russie*. Figaro 22/23.5.1993, S.2.
T4 - François Bonnet: *Un rôle d'opposant à réapprendre*. Libération 26.4. 1993, S. 6/7.
T5 - P. Hu: *La police allemande s'est réjouie trop vite*. Libération 31.5.1993, S.2.
T6 - F.B.: *Simone Veil défend une continuité tranquille*. Libération 26.4. 1993, S. 10.
T7 - Jean-Yves Lhomeau: *Le symbole d'une aventure naufragée*. Libération 3.5.1993, S.3.
T8 - Gérard Nirascou: *Nice: trois hommes pour un fauteuil*. Figaro 22/23.5. 1993, S. 5.
T9 - Valerie Massoneau: *Bosnie-Herzegovine. L'horreur qu'on ne voudrait plus voir*. Paris-Match 20.8.1992, S. 62.
T10 - Céline Buanic: *L'anniversaire de la reine mère*. Paris-Match 20.8.1992, S.23.
T11 - Lou Barthélémy: *Elle n'a pas fini d'étonner!* Marie Claire, Mai 1993, S. 28.
T12 - Jacque de Danne: *Au chevet de la France*. France-Soir 6.5.1993, S. 3
T13 - Marie Sigaud: *Grande stade: alors, on se secoue?* France-Soir 6.5.1993, S.11.
T14 - Alain Le Kim: *Joelle Cramailh*. Marie Claire, Mai 1993, S. 101-102.
T15 - Charles Baudelaire: *La fausse monnaie*. In: Ch. Baudelaire: *Le Spleen de Paris*. Paris, Flammarion 1987: S. 136-137.
T16 - Guy de Maupassant: *Coco*. In: G. de Maupassant: *Contes du jour et de la nuit*. Paris, Gallimard 1984, S. 143-148.

4. An die Daten wurde die Hirata-Poisson-Verteilung, die sich als ein Spezialfall des Ansatzes von Wimmer et al. (1994) ergibt, angepasst (vgl. Wimmer & Altmann, 1996), und zwar in 1-verschobener Form. Die nicht-verschobene Formel lautet:

$$P_x = \begin{cases} e^{-a} & x = 0 \\ \sum_{i=0}^{\left[\frac{x}{2}\right]} \binom{x-i}{i} \frac{e^{-a} a^{x-i}}{(x-i)!} b^i (1-b)^{x-2i}, & x = 1,2,... \end{cases}$$

$a > 0$
$0 \leq b < 1$

$[z]$ über der Summe bedeutet die größte ganze Zahl von z. Die Ergebnisse sind in Tabelle 1 zu sehen. Zu jedem Text werden folgende Werte angegeben:

x - Wortlänge, gemessen in der Zahl der Silben
f_x - Zahl der Wörter des Textes mit Länge x
Np_x - theoretische Werte der Hirata-Poisson-Verteilung
a, b - Parameter der Hirata-Poisson-Verteilung
X_k^2 - der Wert des Chiquadrats mit k Freiheitsgraden
P - Überschreitungswahrscheinlichkeit des Chiquadrats
$C = X^2/N$ - Diskrepanzmaß, das vor allem bei großer Gesamtzahl (N Wörter) benutzt wird.

Das Ergebnis des Anpassungstests wird dann als zufriedenstellend betrachtet, wenn $P \geq 0.05$ oder - besonders bei großem N - $C \leq 0.02$.

Tabelle 1
Anpassung der Hirata-Poisson-Verteilung an französische Daten

x	T 1		T 2		T 3	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	426	427.25	489	493.55	434	428.51
2	168	167.83	171	171.37	158	157.81
3	88	85.12	143	127.83	97	109.38
4	28	24.80	42	37.50	51	33.15
5	5	10.00	7	21.75	7	18.15
	$a = 0.5149$ $b = 0.2371$ $X_1^2 = 0.19$; $P = 0.66$		$a = 0.5460$ $b = 0.3640$ $X_1^2 = 3.61$; $P = 0.06$		$a = 0.5557$ $b = 0.3373$ $X_1^2 = 2.35$; $P = 0.12$	

x	T 4		T 5		T 6	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	372	362.73	440	438.09	373	373.39
2	140	142.48	107	107.93	154	153.98
3	78	93.15	103	105.63	89	88.18
4	49	29.26	33	23.84	35	27.63
5	4	15.38	7	11.60	4	9.51
6			2	3.91	1	3.31
	$a = 0.5725$ $b = 0.3138$ $X_1^2 = 4.32$; $P = 0.04$ $C = 0.0067$		$a = 0.4572$ $b = 0.4611$ $X_1^2 = 0.15$; $P = 0.70$		$a = 0.5635$ $b = 0.2682$ $X_1^2 = 0.01$; $P = 0.91$	

x	T 7		T 8		T 9	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	525	512.55	678	678.79	167	164.79
2	147	151.68	271	270.96	96	96.30
3	104	122.12	172	169.86	40	44.25
4	52	31.71	67	53.41	20	14.90
5	6	14.22	8	19.81	3	5.76
6	1	3.30	4	7.17		
7	2	1.42				
	$a = 0.4904$ $b = 0.3966$ $X_1^2 = 5.26$; $P = 0.02$ $C = 0.0062$		$a = 0.5698$ $b = 0.2994$ $X_1^2 = 0.05$; $P = 0.82$		$a = 0.6821$ $b = 0.1434$ $X_1^2 = 0.71$; $P = 0.40$	

x	T 10		T 11		T 12	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	242	242.20	390	389.96	350	350.48
2	92	92.78	146	145.38	161	163.92
3	34	32.14	40	41.19	92	87.09
4	8	7.77	9	8.62	32	28.78
5	1	2.11	2	1.85	8	12.73
	$a = 0.4424$ $b = 0.1342$ $X_2^2 = 0.67$; $P = 0.71$		$a = 0.4090$ $b = 0.0884$ $X_2^2 = 0.07$; $P = 0.97$		$a = 0.6068$ $b = 0.2293$ $X_2^2 = 2.44$; $P = 0.30$	

x	T 13		T 14		T 15	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	415	412.87	621	620.48	316	314.81
2	161	163.55	226	225.59	118	117.65
3	83	85.03	75	77.09	51	53.72
4	30	25.13	19	18.09	15	14.60
5	8	10.42	5	4.75	6	5.22
	$a = 0.5236$ $b = 0.2435$ $X_2^2 = 1.60$; $P = 0.45$		$a = 0.4217$ $b = 0.1379$ $X_2^2 = 0.12$; $P = 0.94$		$a = 0.4745$ $b = 0.2124$ $X_2^2 = 0.28$; $P = 0.87$	

x	T 16	
	f_x	NP_x
1	863	857.66
2	376	390.42
3	126	107.60
4	16	22.01
5	1	4.31
	$a = 0.4771$ $b = 0.0458$ $X_2^2 = 7.87$; $P = 0.02$ $C = 0.0057$	

5. Als Ergebnis der Untersuchung kann festgestellt werden, daß alle Texte der 1-verschobenen Hirata-Poisson-Verteilung folgen. Das gilt auch für die beiden literarischen Texte (15 und 16), die hier ergänzend zu den Pressetexten bearbeitet wurden. Es deutet sich damit an, daß diese Verteilung zumindest für ein breiteres Spektrum des gegenwärtigen Französischen ein geeignetes Modell darstellt. Seine tatsächliche Reichweite muß allerdings noch erprobt werden, da sich bisher in jeder Sprache zu einem Grundmodell Modifikationen ergaben (vgl. die Artikel in diesem Band).

Vergleicht man das Ergebnis dieser Untersuchung mit entprechenden Arbeiten zum relativ nahe verwandten Italienischen (vgl. Gaeta, 1994; Holberg, 1994), so kann man feststellen, daß aufgrund der bisherigen Resultate das Französische zunächst einmal bei etwa vergleichbarer Basis homogener wirkt, da für die italienischen Texte mehrere Modelle angenommen werden müssen, während die französischen Texte alle ein und demselben Modell folgen.[1] Dabei muß in Betracht gezogen werden, daß die Datenbasis für beide Sprachen noch relativ schmal ist, und zwar sowohl hinsichtlich der Anzahl als auch der Textsortenzugehörigkeit der bearbeiteten Texte. Es ist außerdem damit zu rechnen, daß für verschiedene Entwicklungsstufen einer Sprache unterschiedliche Modelle gelten.

[1] Für einige französische Texte konnte als Modell auch die 1-verschobene gemischte Poisson-Verteilung verwendet werden, jedoch meistens mit deutlich schlechteren Ergebnissen.

Literatur

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten *(erscheint).*

Feldt, S., Janssen, M., & Kuleisa, S. (1994). Untersuchung zur Gesetzmäßigkeit von Wortlängenverteilungen in französischen Briefen und Pressetexten *(erscheint).*

Fucks, W. (1968). *Nach allen Regeln der Kunst.* Stuttgart: Deutsche Verlags-Anstalt.

Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *Zeitschrift für empirische Textforschung 1,* 44-48.

Hollberg, C. (1994). Wortlänge in italienischen Pressetexten. Msk.

Klein, H.W. (1963). *Phonetik und Phonologie des heutigen Französisch.* München: Hueber.

Lausberg, H. (1956). *Romanische Sprachwissenschaft I.* Berlin: de Gruyter.

Martinet, A. (éd.) (1969). *La linguistique. Guide alphabétique.* Paris: Presse Universitaire de la France.

Martinet, A., & Walter, H. (1973). *Dictionnaire de la prononciation française dans son usage réel.* Paris: France Expansion.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Glottometrika 15, 1996, 166-180

Zur Länge der Wörter in deutschen Texten

Gabriel Altmann, Bochum
Karl-Heinz Best, Göttingen

1. In der klassischen/Standard-Linguistik spielt die Wortlänge keine Rolle; in der theoretischen Linguistik muß sie wie jede andere sprachliche Größe in das nomologische Netz der Sprache eingebettet werden. Ein erster Schritt dahin erfolgte durch G.K. Zipf (besonders 1949). In zahlreichen Studien untersuchte man anschließend einzelne Zusammenhänge der Wortlänge mit anderen Spracheigenschaften (vgl. Miller, Newman & Friedman, 1958; Bürmann, Frank & Lorenz, 1963; Guiter, 1974; Skalička, 1979; Rothe, 1983; Fickermann, Markner-Jäger & Rothe, 1984; Köhler, 1986; Ogino, 1989; Hammerl, 1991; Tuldava, 1991; Miyajima, 1992, um nur einige wenige zu nennen).

Bei der Gestaltung von Texten ist die unterschiedliche Länge der Wörter und ihre Häufigkeit eine zentrale Größe. Die mit ihrer Erforschung verbundene Problematik wird in Grotjahn & Altmann (1993) behandelt. Modellansätze findet man in Wimmer et al. (1994).

W. Fucks (1955a,b) schlug als erstes Wortlängenverteilungsmodell die Poisson-Verteilung vor. Grotjahn (1982) stellte fest, daß diese Verteilung in Texten selten adäquat ist und erweiterte sie so, daß er den Parameter der Poisson-Verteilung als eine gamma-verteilte Variable betrachtete und dadurch die negative Binomialverteilung erhielt. In gestutzter Form scheint diese Verteilung als Modell für das Deutsche sehr gut geeignet zu sein (s. unten); außerdem kann man sie auch aus einem "synergetischen" Ansatz mit linguistischer Begründung ableiten. Das Resultat kann man als *Grotjahn-Gesetz* bezeichnen.

Wir gehen davon aus, daß jede sprachliche Variable in Texten eine "anständige" Häufigkeitsverteilung aufweist, die weder der diskreten Rechteckverteilung noch einem chaotischen Gebilde ähnelt. Dies folgt aus der Annahme, daß sie einem versteckten, organisierenden Mechanismus folgt, der entschlüsselt werden muß. Eine Möglichkeit dazu besteht darin, anzunehmen, daß die Häufigkeitsklassen eine Art Selbstregulation entwickeln, indem sich die benachbarten Klassen (x und x-1) in ein proportionales Verhältnis ordnen: $P_x \sim P_{x-1}$. Die Proportionalität kann entweder konstant sein, z.B. $P_x = aP_{x-1}$ (mit $0 < a < 1$), oder sie ist als eine Funktion darstellbar, z.B. $P_x = g(x)P_{x-1}$ Die Funktion $g(x)$ übernimmt dabei die Rolle der Selbstorganisation, wenn die betroffene Spracheinheit in einen "turbulenten" Zustand gerät, beispielsweise dann, wenn Lemmata des Lexikons im Text verwendet werden, wo sie den Regeln der Grammatik, der Willkür oder Bindung des Stils, des Genres usw. ausgesetzt werden. Die Funktion $g(x)$ ist ein "Ordner",

der dafür sorgt, daß ein Text seinem Zweck dient, daß die kommunikative Funktionalität des sprachlichen Systems erhalten bleibt, daß der Sprecher nicht übermäßig fluktuiert usw.

Für die Wortlängenverteilung in deutschen Texten gehen wir davon aus, daß die Funktion $g(x)$ aus drei Teilen besteht und als

$$g(x) = \frac{a + bx}{cx}$$

darstellbar ist.

Hier symbolisiert a den konstanten Anteil der gegebenen Sprache, der die Längen im Lexikon darstellt. So findet man im deutschen oder englischen Lexikon große Mengen einsilbiger Wörter, die auch im Text nicht verändert werden, während im Indonesischen z.B. nur wenige Lemmata einsilbig sind.

Der Sprecher oder Autor modifiziert diese Längen (X), indem er seinem Stil oder Genre entsprechend kürzere oder längere Wortformen benutzt. Die einfachste Form, diese autorenspezifische Veränderung zu modellieren, ist die Annahme einer linearen Modifikation, d.h. bx.

Der Hörer (die Sprachgemeinschaft) reagiert darauf in der gleichen Weise und setzt einen analogen "Bremsmechanismus" cx ein, der die Gestaltungsmöglichkeiten des Texterzeugers in einem kommunikationstechnisch zulässigen Rahmen hält (technisch: die Divergenz der Funktion P_x verhindert). Wegen der Konvergenz muß natürlich $c > b$ sein. Man darf dies aber nicht so auffassen, als ob der gegebene oder potentielle Hörer selbst direkt in die Kommunikation eingreife; es ist vielmehr davon auszugehen, daß der Sprecher sich unbewußt in die Rolle des Hörers versetzt, wodurch sich eine bestimmte Proportion zwischen seinen eigenen kommunikativen Interessen und denen des Hörers einstellt. Die obige Funktion unterscheidet sich von dem Fucksschen Gesetz nur um die Größe bx.

Wir erhalten also

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

und lösen die Gleichung mit der Bedingung $P_0 = 0$, weil es im Deutschen keine Wörter mit silbischer Länge $x = 0$ gibt (die Wortlänge wird in Anzahl der Silben

gemessen).[1] Klammern wir b im Zähler aus und bezeichnen $a/b = r\text{-}1$, $b/c = q$, so erhalten wir

$$(2) \qquad P_x = \frac{b}{c} \cdot \frac{\frac{a}{b} + x}{x} P_{x-1} = q \frac{r + x - 1}{x} P_{x-1}$$

Die schrittweise Lösung dieser Gleichung ergibt

$$P_2 = \frac{r + 1}{2} q P_1$$

$$P_3 = \frac{(r + 1)(r + 2)}{2 \cdot 3} q^2 P_1$$

$$\vdots$$

$$P_x = \frac{(r + 1)(r + 2)...(r + x - 1)}{2 \cdot 3 \cdot ... \cdot x} q^{x-1} P_1$$

$$= \frac{(r + x - 1)!}{x!\,r!} q^{x-1} P_1$$

$$= \frac{1}{qr} \binom{r + x - 1}{x} q^x P_1. \qquad (3)$$

Um P_1 zu bestimmen, verlangen wir, daß $\Sigma P_x = 1$ und erhalten

$$\frac{P_1}{qr} \sum_{x=1}^{\infty} \binom{r + x - 1}{x} q^x = \frac{P_1}{qr} \sum_{x=1}^{\infty} \binom{-r}{x} (-q)^x = \frac{P_1}{qr} \left[(1 - q)^{-r} - 1 \right] = 1,$$

woraus sich

$$(4) \qquad P_1 = \frac{qr}{(1 - q)^{-1} - 1}$$

[1] Nullsilbige Wörter können im Deutschen allerdings dann vorkommen, wenn man "'s" für "es" als nullsilbiges Wort betrachtet. Es handelt sich dabei aber vor allem um ein Phänomen der gesprochenen Sprache, das schriftsprachlich nur relativ selten in Erscheinung tritt. Man kann es auch als einen emergenten Fall der *objektiven Konjugation*, analog dem Ungarischen, betrachten.

ergibt. Bezeichnen wir $1 - q = p$ und setzen (4) in (3) ein, so erhalten wir schließlich

$$(5) \qquad P_x = \binom{r + x - 1}{x} \frac{p^r q^x}{1 - p^r}, \quad x = 1,2,3,...$$

und damit die 0-gestutzte oder "positive" negative Binomialverteilung.

2. Diese Verteilung wurde auf 26 vollständige deutsche Texte angewendet, wie in Tabelle 1 aufgeführt (vgl. auch Best & Zhu, 1993 und Laass, 1996). Auch wenn der Chiquadrat-Test nicht in allen Fällen ein befriedigendes Resulat liefert, da einige P-Werte (= Wahrscheinlichkeit des gegebenen oder eines extremeren Chiquadrat-Wertes) recht klein sind (z.B. 0.02), bleiben wir aus mehreren Gründen bei diesem Verfahren:

(i) In einigen Fällen zeigt schon der Koeffizient $C = (X^2/N)^{1/2}$, der nur von N, nicht aber von den Freiheitsgraden abhängt, daß die Anpassungen akzeptabel sind. Die Kontrolle weiterer Koeffizienten, wie z.B. λ^2 von Pederson & Johnson (1990), bei dem auch die Freiheitsgrade und die linearen Abweichungen in Betracht gezogen werden, verbessert den Eindruck beträchtlich.

(ii) Die Anwendung anderer Anpassungstests, wie z.B. der Systeme von Cressie & Read (1984) oder Moore & Spruill (1975) u.a. zeigen, daß die Apassungen höhere P's aufweisen als der klassische Chiquadrat-Test.

(iii) In der Hoffnung, daß andere Textologen und Linguisten sich dieser Forschung widmen werden, ist es aber ratsam, bei allgemein bekannten Techniken zu bleiben, damit die Einheitlichkeit gewährleistet ist.

(iv) Bei Texten muß man immer in Betracht ziehen, daß nachträgliche Korrekturen des Autors (die man nicht kennt) den Wortlängenrhythmus des Textes, der ja völlig unbewußt ist, beeinträchtigen können. Daher muß man auch mit gelegentlichen Falsifikationen des Modells rechnen und in solchen Fällen entweder andere Techniken verwenden oder "Manuskriptforschung" betreiben oder schließlich zu dem in Abschnitt 3 angesprochenen Kriterium (Parametervergleich) greifen, das in jedem Fall in Betracht gezogen werden sollte. In der Praxis bedeutet das lediglich, daß man feststellen soll, ob die ermittelten q- und r-Werte mit der Kurve in Tabelle 2 übereinstimmen.

Die Rechnungen wurden mit dem Programm FITTER iterativ durchgeführt und stellen eine optimale Anpassung dar, die sich durch weitere Iterationen nur unbeträchtlich verbessern läßt, wenn überhaupt. Es handelt sich um folgende Texte:

T 1. Krieg statt Hilfe (*Der Spiegel, Nr. 47, Jg. 46, 1992: 98-101*)
T 2. Notfalls neue Wege (*Der Spiegel,* Nr. 47, Jg. 46, 1992, 41-44)
T 3. Tja, neue Ziele setzen (*Der Spiegel,* Nr 47, Jg. 46, 1992, 44-47)

T 4. Doofe aus dem Westen (*Der Spiegel,* Nr. 47, Jg. 46, 1992, 110-112)
T 5. Tödlicher Anschlag (*Der Spiegel,* Nr. 47, Jg. 46, 1992, 121-123)
T 6. An die Wand (*Der Spiegel,* Nr. 48, Jg. 46, 1992, 21-22)
T 7. Wie beim Jo-Jo (*Der Spiegel,* Nr. 48, Jg. 46, 1992, 29-30)
T 8. Nur lachen (*Der Spiegel,* Nr. 48, Jg. 46, 1992, 36-37)
T 9. Historische Achse (*Der Spiegel,* Nr. 48, Jg. 46, 1992, 120-122).
T 10. Ganz Wertfrei (*Der Spiegel,* Nr. 48, Jg. 46, 1992, 126-128).
T 11. Thaddaeus Troll, Der himmlische Computer
T 12. K. Kusenberg, Die ruhelose Kugel
T 13. R. Katz, Mondnacht bei den Pyramiden
T 14. Goethe: Brief 641
T 15. Goethe: Brief 605
T 16. Goethe: Brief 644
T 17. Goethe: Brief 659
T 18. Goethe: Brief 612
T 19. Goethe: Brief 589
T 20. Goethe: Brief 596
T 21. Goethe: Brief 591
T 22. Goethe: Brief 667
T 23. Goethe: Der Totentanz
T 24. Schiller Die Kraniche des Ibycus
T 25. Goethe: Erlkönig.
T 26. Goethe : Brief 647

Die Texte 1-10 standen in der Sparte *Deutschland* des Magazins *Der Spiegel.* Die Texte 11-13 stammen aus *Kleine Bettlektüre für den vielseitigen Zwilling* (Scherz Verlag s.d.), die Daten der Texte 14 - 26 aus Grotjahn (1979).

Aus typographischen Gründen wurde die Anordnung der Texte in Tabelle 1 etwas umgestellt. Den Koeffizienten C haben wir nur in einigen problematischen Fällen angegeben.

Mehrere Umstände sprechen für diese Art der Modellierung:

(1) Sie interpretiert die Parameter des Modells linguistisch, auch wenn es schwer sein wird zu erforschen, welche Umstände jeweils zu den gegebenen konkreten Werten führten (s.u).

(2) Sie stimmt mit der synergetischen Ausrichtung der theoretischen Linguistik überein, in der Selbstregulation und -organisation eine eminente Rolle spielen. Denn einerseits ist (wegen $P_x = f(P_{x-1}, P_{x-2}, ..., P_1)$ und $P_1 = f(P_2, P_3, ...)$) evident, daß eine Rückkopplung stattfindet, und andererseits erweist sich die ordnende Funktion

$$g(x) = \frac{a + bx}{cx} = \frac{a}{c}x^{-1} + \frac{b}{c} = Ax^{-1} + B$$

als ein Spezialfall des Menzerathschen Gesetzes (vgl. Altmann, 1980, Altmann &

Schwibbe, 1989), dessen eminente Rolle in der sprachlichen Selbstregulation bekannt ist (vgl. Köhler, 1986; Hammerl, 1991). Seine Rolle als Proportionalitätsfunktion wurde von Hřebíček (1993) erkannt.

(3) Betrachten wir r als eine Beziehung des Sprachmaterials zu der Modifikation durch den Sprecher, wie oben angesetzt, d.h.

$$r - 1 = \frac{a}{b} = \frac{Sprache}{Sprecher}$$

und q als eine Beziehung zwischen Sprecher und Hörer, d.h.

$$q = \frac{b}{c} = \frac{Sprecher}{Hörer},$$

dann sehen wir, daß der Sprecher in dem Maße, wie er von der Norm abweicht bzw. das sprachliche Material modifiziert, von dem Hörer in seine Schranken verwiesen wird; oder umgekehrt, er darf der Sprache nur so viel Veränderung zumuten, wie der Hörer erlaubt, wenn die Kommunikation nicht zusammenbrechen soll. Dies läßt vermuten, daß r eine Funktion von q sein könnte oder umgekehrt. Da aber die Gemeinschaft der Hörer die dem Einzelsprecher übergeordnete Ebene/Instanz ist, läßt sich analog zu der hierarchischen Ordnungsfunktion des Menzerathschen Gesetzes auch hier vermuten, daß $r = f(q)$ eine Potenzfunktion sein sollte.

Tabelle 1.
Anpassung des Grotjahn-Gesetzes an deutsche Daten

x	T 21		T 23		T 25	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	59	59.00	218	223.73	152	156.31
2	27	27.61	99	88.96	65	53.81
3	10	8.76	21	23.67	6	12.38
4	2	2.63	4	5.64	2	2.50
	$r = 56.7739$ $p = 0.9838$ $X_1^2 = 0.33$ $P = 0.56$		$r = 269.6746$ $p = 0.9971$ $X_1^2 = 2.05$ $P = 0.15$		$r = 321.94$ $p = 0.9979$ $X_1^2 = 5.83$ $P = 0.016$	

In der Tat erhalten wir dieses Resultat, wenn wir aus Tabelle 1 alle q's ($q = 1 - p$) den entsprechenden r gegenüberstellen, wie in Tabelle 2 dargestellt. Die iterativ berechnete Kurve

$$r = cq^d = 0.7871q^{-1.0364}$$

liefert den Determinationsquotienten $D = 0.8974$ und unterstützt dadurch diese Annahme.

Tabelle 1 (Fortsetzung)

x	T 14		T 17		T 18		T 19	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	266	267.91	151	151.26	164	166.27	219	220.75
2	133	131.04	68	64.90	105	98.90	125	116.01
3	44	42.93	16	20.27	35	39.42	32	42.03
4	11	10.60	7	5.15	15	11.84	15	11.80
5	1	2.52	1	1.42	1	3.57	3	3.41
	$r = 209.5730$ $p = 0.9954$ $X_2^2 = 0.99$ $P = 0.61$		$r = 9.8165$ $p = 0.9207$ $X_2^2 = 1.83$ $P = 0.40$		$r = 192.2165$ $p = 0.9938$ $X_2^2 = 3.58$ $P = 0.17$		$r = 28.2053$ $p = 0.9640$ $X_2^2 = 4.02$ $P = 0.13$	

x	T 20		T 22		T 24	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	134	141.41	77	76.35	580	586.47
2	102	90.89	51	52.32	296	288.02
3	36	39.01	26	25.31	97	94.59
4	13	12.57	10	9.70	24	23.37
5	2	4.12	4	4.32	1	5.55
	$r = 593.6368$ $p = 0.9978$ $X_2^2 = 3.33$ $P = 0.19$		$r = 15.8956$ $p = 0.9189$ $X_2^2 = 0.09$ $P = 0.96$		$r = 351.6932$ $p = 0.9969$ $X_1^2 = 0.88$ $P = 0.35$	

Tabelle 1 (Fortsetzung)

x	T 12		T 13		T 15		T 26	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	486	484.24	739	745.07	428	431.98	259	260.51
2	317	315.30	452	422.90	215	201.49	132	123.37
3	131	140.20	148	181.54	51	63.16	37	47.23
4	57	47.86	74	65.37	19	14.96	19	15.94
5	13	13.37	24	20.83	1	2.86	6	4.94
6	1	4.03	7	8.29	1	0.55	1	2.01
	$r = 5.2313$ $p = 0.9683$ $X_2^2 = 3.02$ $P = 0.22$		$r = 6.4377$ $p = 0.8474$ $X_3^2 = 10.06$ $P = 0.02$ $C = 0.08$		$r = 123.029$ $p = 0.9925$ $X_2^2 = 5.94$ $P = 0.08$ $C = 0.09$		$r = 3.7042$ $p = 0.7987$ $X_2^2 = 3.42$ $P = 0.18$	

x	T 2		T 5		T 11		T 16	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	302	301.39	439	447.10	588	575.45	338	344.30
2	221	205.45	337	302.96	380	394.57	242	219.54
3	117	119.97	132	164.85	196	208.90	76	96.76
4	41	64.20	79	78.70	116	94.28	36	33.11
5	43	32.47	40	34.42	34	38.13	11	9.37
6	14	15.79	16	14.13	14	14.23	1	2.28
7	7	7.45	6	5.53	4	4.99	0	0.49
8	8	6.28	2	3.31	1	2.45	2	0.15
	$r = 2.5090$ $p = 0.6115$ $X_5^2 = 13.77$ $P = 0.02$ $C = 0.13$		$r = 3.8885$ $p = 0.7228$ $X_5^2 = 12.20$ $P = 0.03$ $C = 0.11$		$r = 5.2313$ $p = 0.7931$ $X_5^2 = 8.11$ $P = 0.15$		$r = 26.2050$ $p = 0.9531$ $X_3^2 = 7.40$ $P = 0.06$ $C = 0.10$	

Tabelle 1 (Fortsetzung)

x	T 3		T 4		T 6		T 8	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	459	456.16	536	535.80	524	516.31	352	347.45
2	211	208.60	366	358.03	288	288.17	225	225.99
3	95	104.65	173	186.61	136	150.08	127	126.18
4	51	54.82	97	83.54	80	75.35	62	64.64
5	44	29.44	29	33.71	42	36.99	28	31.33
6	16	16.07	7	12.61	13	17.88	18	14.61
7	7	8.88	7	4.45	12	8.55	3	6.62
8	5	4.94	1	|1.50	4	4.05	6	|2.93
9	2	6.44	1	|0.75	2	3.62	1	|2.25
	$r = 0.5492$ $p = 0.4096$ $X_6^2 = 11.82$ $P = 0.07$ $C = 0.12$		$r = 4.8834$ $p = 0.7728$ $X_5^2 = 7.97$ $P = 0.16$		$r = 1.5024$ $p = 0.5539$ $X_6^2 = 5.82$ $P = 0.44$		$r = 2.4561$ $p = 0.6258$ $X_5^2 = 3.95$ $P = 0.56$	

x	T 10		T 7		T 9		T 1	
	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*	n_x	*PoNB*
1	309	295.01	456	450.95	316	318.52	383	375.63
2	181	198.92	277	275.92	192	174.64	249	260.04
3	121	117.16	132	151.11	75	91.59	165	153.73
4	56	64.00	100	77.97	52	46.95	76	83.11
5	42	33.33	31	38.74	21	23.73	38	42.41
6	21	16.78	17	18.75	9	11.88	26	20.78
7	7	6.25	10	8.91	10	5.90	10	9.88
8	3	3.98	4	4.17	3	2.92	6	4.59
9	1	3.57	2	1.93	-	1.44	-	2.09
10	-	-	-	|0.88	-	|0.70	-	|0.94
11	-	-	1	|0.77	1	|0.72	1	|0.80
	$r = 2.2236$ $p = 0.5817$ $X_6^2 = 8.97$ $P = 0.18$		$r = 1.9100$ $p = 0.5796$ $X_7^2 = 10.79$ $P = 0.15$		$r = 1.2992$ $p = 0.5231$ $X_7^2 = 10.69$ $P = 0.15$		$r = 2.5593$ $p = 0.6110$ $X_7^2 = 6.62$ $P = 0.47$	

Tabelle 2
Beziehung zwischen den Parametern q und r

Text	q	r	$r = cq^d$
Goethe: Erlkönig	0.0021	321.94	468.8863
-"- : Brief 596	0.0022	593.64	446.8592
-"- : Totentanz	0.0029	269.68	335.7974
Schiller: Die Kraniche...	0.0031	351.69	313.4148
Goethe: Brief 612	0.0062	192.22	153.0231
-"- : Brief 641	0.0072	209.57	131.0953
-"- : Brief 605	0.0075	123.10	125.6753
-"- : Brief 591	0.0162	56.77	56.6652
Kusenberg: Die ruhelose Kugel	0.0317	40.09	28.2986
Goethe: Brief 589	0.0360	28.21	24.8100
-"- : Brief 644	0.0469	26.20	18.8718
-"- : Brief 659	0.0793	9.82	10.9618
-"- : Brief 667	0.0810	15.90	10.7240
Katz: Mondnacht bei den Pyr.	0.1526	6.44	5.5699
Goethe : Brief 647	0.2013	3.70	4.1451
Troll: Der himmlische Computer	0.2169	5.32	3.8717
T 4	0.2272	4.88	3.6903
T 5	0.2772	3.89	3.0041
T 8	0.3742	2.48	2.2025
T 2	0.3885	2.51	2.1187
T 1	0.3890	2.56	2.1159
T 10	0.4183	2.22	1.9628
T 7	0.4204	1.91	1.9527
T 6	0.4461	1.50	1.8364
T 9	0.4769	1.30	1.7139
T 3	0.5904	0.55	1.3743
$c = 0.7871$, $d = -1.0364$, *Determinationskoeffizient* = 0.8974			

Vergleicht man nun die Daten in Tabelle 2, so sieht man, daß Gedichte (*Erlkönig, Totentanz, Die Kraniche des Ibycus*) ein niedriges q (und großes r) haben, Briefe von Goethe ein etwas größeres q (bei kleinerem r), andere Prosa noch höhere q (und niedrigere r)-Werte und Zeitungsartikel ihre Wortlängenverteilung mit dem relativ größten q (bzw. kleinsten r) gestalten. Diese Beobachtung ist zwar einstweilen noch sehr vorläufig und vage, aber man sieht, daß bei einem dem Stil des

Autors entsprechenden Wert von b (mit dem er den gegebenen Text gestaltet) der Parameter c sich automatisch so einstellt, daß $r = f(q)$ einer "Menzerathschen" Abhängigkeit entspricht. Diese Selbstorganisationstechnik ist sehr ähnlich den numerischen Optimierungsalgorithmen, bei denen man einen Parameter festsetzt und iterativ ein Minimum für den zweiten Parameter sucht. Dieser erstaunliche Mechanismus erfordert weitere Untersuchungen, besonders von Literaturwissenschaftlern und evtl. Psychologen, die die Werte von q oder r mit textologischem Inhalt füllen müßten.

3. Angesichts dieser Resultate sollte folgendes berücksichtigt werden:

(i) Weitere deutsche Daten könnten dieses Modell und seine Interpretation falsifizieren, aber auch bestätigen.

(ii) Es ist denkbar, daß für andere Sprachen ganz andere Modelle entwickelt werden müssen. Für das Deutsche haben wir im Grunde genommen lediglich das Fucksschе Gesetz um den Faktor "Sprechereinfluß" erweitert (s. Grotjahn & Altmann, 1993), um durch linguistisch interpretierbare zusätzliche Bedingungen zum Grotjahnschen Gesetz zu kommen; das Fucksschе Gesetz muß aber womöglich nicht in allen Sprachen durch Zusatzbedingungen auf die gleiche Weise wie hier modifiziert werden. Es sind jedenfalls noch viele theoretische und empirische Untersuchungen nötig, um das gesamte Umfeld der Wortlänge zu erforschen (vgl. Wimmer & Altmann, 1996).

(iii) Die Tatsache, daß (5) gegen die positive Poisson-Verteilung konvergiert, wenn $r \to \infty$, $q \to 0$ und $rq \to a$, sollte dazu führen, daß mindestens einige der ersten Texte in Tabelle 2 sich auch mit der positiven Poisson-Verteilung erfassen lassen. In der Tat kann man zeigen, daß, solange in Tabelle 2 $q < 0.1$, die Texte auch dem einfacheren Fucksschen Gesetz folgen. Ab etwa $q > 0.1$ ist dies nicht mehr der Fall.

(iv) Das Resultat dieser Untersuchung ist nicht nur eine Bestätigung des Fucksschen und des Grotjahnschen Gesetzes, sondern eine wichtige zusätzliche Bestätigung des Menzerathschen Gesetzes, das sowohl im selbstregulatorischen als auch im selbstorganisatorischen Bereich eine wichtige Rolle spielt. Das Fucksschе Gesetz kann man in Begriffen unseres Ansatzes als $P_x = ax^{-1}P_{x-1}$ darstellen, wobei $g(x) = ax^{-1}$ ein "störungsfreier" Spezialfall des Menzerathschen Gesetzes ist. Es ist damit eine weitere Bestätigung der Entdeckung Hřebíčeks (1993), daß dieses Gesetz oft dort zur Wirkung kommt, wo eine niedrigere Ebene durch eine höhere "geordnet" wird - hier die Wortlängenverteilung, die im Text durch (stilistische u.a.) Bedürfnisse des Texterzeugers in Turbulenz gebracht und durch das Bedürfnis nach Minimierung des Dekodierungsaufwands der Sprachgemeinschaft geordnet wird. Wird der "Ordner" selbst modifiziert, dann geschieht dies auch auf eine reguläre Weise. In unserem Fall wurde hier nach dem Prinzip der Selbstähnlichkeit verfahren: Die Modifikation des Ordners $g(x) = ax^{-1}$ im Fucksschen Gesetz befolgt

das gleiche Prinzip $r = cq^{-d}$, und wie man sieht, ist auch $d \approx 1$, oder etwas freier ausgedrückt, eine höhere Ebene gestaltet die unmittelbar niedrigere nach ihrem eigenen Bild. Wie weit diese Selbstähnlichkeit geht, wissen wir noch nicht; man kann aber vermuten, daß die Verfolgung dieser Spur uns neue sprachliche (Zwischen)Ebenen zu entdecken ermöglicht (vgl. Hřebíček, 1992, 1993).

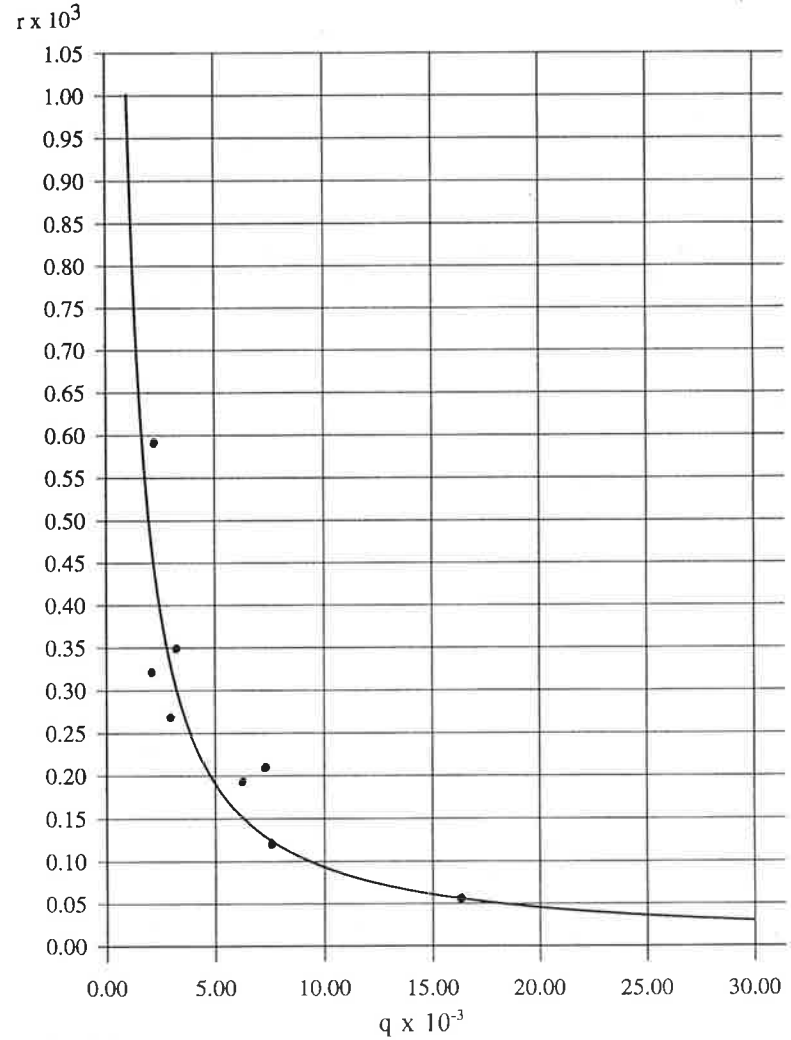


Abbildung 1a. Beziehung zwischen r und q, $q \in \langle 0.001, 0.03 \rangle$

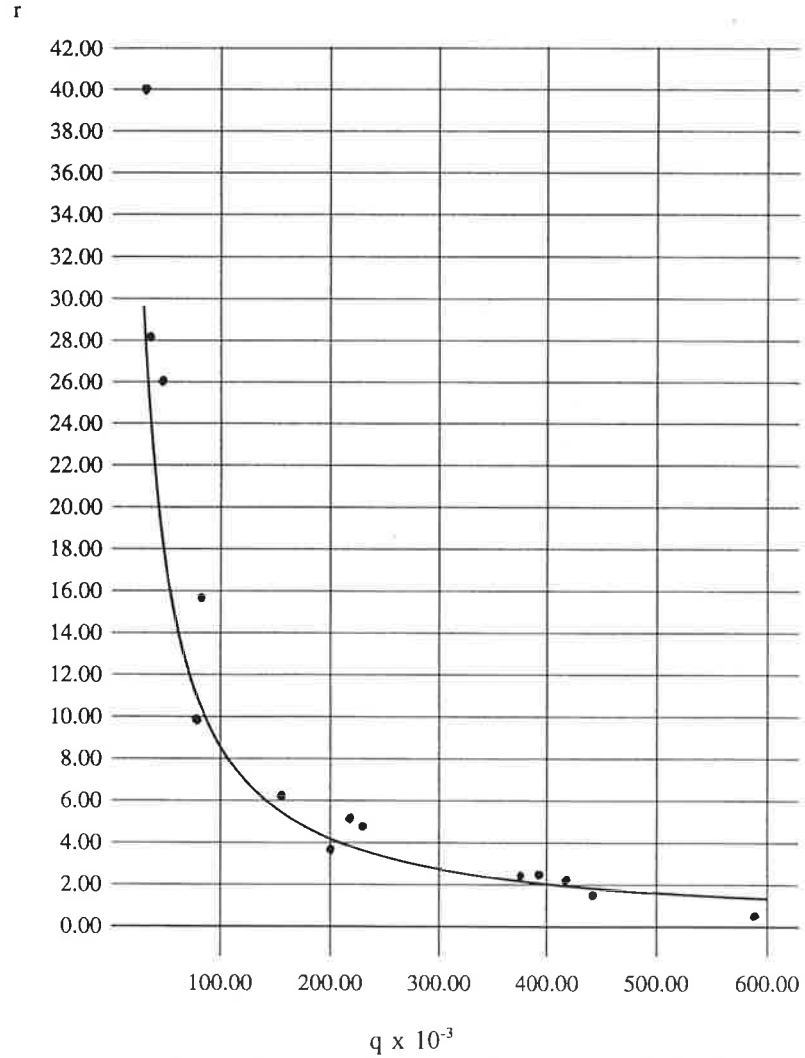


Abbildung 1b. Beziehung zwischen r und q, $q \in \langle 0.03, 0.60 \rangle$

Literatur

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

Altmann, G., & Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Best, K.-H., & Zhu, J. (1993). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, U. (Hrsg.), *Computatio Linguae II.* Stuttgart: Steiner.

Bürmann, G., Frank, H., & Lorenz, L. (1963). Informationstheoretische Untersuchungen über Rang und Länge deutscher Wörter. *Grundlagenstudien aus Kybernetik und Geisteswissenschaft 4,* 73-90.

Cressie, N., & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B 46,* 440-464.

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

Fucks, W. (1955a). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln-Opladen: Westdeutscher Verlag.

Fucks, W. (1955b). Theorie der Wortbildung. *Mathematisch-physikalische Semesterberichte 4,* 195-212.

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum: Brockmeyer.

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

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length: some methodological problems. In: Köhler, R., & Rieger, B. (Hgg.) *Contributions to Quantitative Linguistics.* Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991. Dordrecht: Kluwer, 141-153.

Guiter, H. (1974). Les relations (fréquence-longueur-sens) des mots (langues Romanes et Anglais). *Atti di Congresso Internazionale di Linguistica 14/4,* 373-381.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.

Hřebíček, L. (1992). *Text in Communication: Supra-sentence Structures.* Bochum: Brockmeyer.

Hřebíček, L. (1996). Word associations and text. In diesem Band.

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

Laass, F. (1996). Zur Verteilung der Wortlänge in deutschen Lesebuchtexten. In diesem Band.

Miller, G.A., Newman, E.B., & Friedman, E.A. (1958). Length-frequency statistics for written English. *Information and Control 1,* 370-389.
Miyajima, T. (1992). Relationships in the length, age and frequency of Classical Japanese words. *Glottometrika 13,* 219-229.
Moore, D., & Spruill, M.C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics 3,* 599-616.
Ogino, T. (1989). Length and politeness of honorific expressions. In: Mizutani, Sh. (ed.), *Japanese Quantitative Linguistics.* Bochum: Brockmeyer, 188-199.
Pederson, S.P., & Johnson, M.E. (1990). Estimating model discrepancy. *Technometrics 32,* 305-314.
Rothe, U. (1983). Wortlänge und Bedeutungsmenge. Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5,* 101-112.
Skalička, V. (1979). Ein 'typologisches Konstrukt'. In: Skalička, V., *Typologische Studien.* Braunschweig: Vieweg, 335-341.
Tuldava, J. (1991). O verojatnostno-statističeskom modelirovanii pričinnosledstvennych zavisimostej v jazyke. In: *Evrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka.* Smolensk, 9-11.
Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In diesem Band.
Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *J. of Quantitative Linguistics 1,* 98-106.
Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge: Addison-Wesley.

Glottometrika 15, 1996, 181-194

Zur Häufigkeit der Wortlängen in deutschen Lesebuchtexten

Françoise Laass, Göttingen

1. Die vorliegende Untersuchung entstand im Rahmen des Wortlängenprojekts, das in Göttingen von K.-H. Best organisiert und koordiniert wird. Die bisher durchgeführten Untersuchungen zum Deutsch der Gegenwart (Altmann & Best, 1996; Best, 1994; Best & Zhu, 1994) haben gezeigt, daß ein weites Textspektrum unserer Sprache mit einem einzigen Modell, der positiven negativen Binomialverteilung (= 0-gestutzten negativen Binomialverteilung), erfaßt werden kann. Die Frage, ob es sich hierbei um ein Modell für alle deutschen Texte oder nur für einen Teil davon handelt - und wenn letzteres: für welchen? - soll mit der Bearbeitung von Lesebuchtexten weiter verfolgt werden. Lesebuchtexte sind, so darf man erwarten, ein vielseitiges Erprobungsfeld für eine Theorie der Wortlängenverteilung: Auch wenn die Texte nicht alle von vornherein für Kinder und Jugendliche als Publikum konzipiert wurden, so sind sie doch sicher so ausgewählt, daß sie den Herausgebern auch sprachlich als für die entsprechenden Altersklassen geeignet erschienen.

Die Untersuchung wurde nach den gleichen Prinzipien durchgeführt, wie sie in Best & Zhu (1994) und Best (1996) beschrieben sind. Hier in Kürze die Operationalisierung von "Wort" und "Silbe".

Als "Wort" wird das graphemische Wort gewertet[1], d.h. das Wort grenzt sich durch Leerzeichen und Interpunktion von den anderen Wörtern ab. Ausgenommen von dieser Regel sind Trennungs- und Bindungsstrich, da diese Interpunktionszeichen die Einheit eines Wortes hervorheben. Die Wortlänge wird nach der Zahl seiner Silben bestimmt.

Die Anzahl der Silben im Wort richtet sich nach der Anzahl der Vokale und der Diphthonge pro Wort. In Zweifelsfällen ist es hilfreich, die Bühnenaussprache zur Orientierung heranzuziehen. So muß z.B. beim Auftreten zweier

[1] Vgl. Bünting & Bergenholtz (1989:36-39). Bünting & Bergenholtz beleuchten die Definition des Wortes unter verschiedenen Aspekten: Neben dem phonologisch/graphematischen Aspekt gehen sie auch auf den semantischen und den distributionalen Aspekt ein. Denn zur Lösung von Problemfällen bedarf es nicht nur eines Aspekts, sondern aller drei.

aufeinanderfolgender Vokale, sofern sie nicht zu den Diphthongen gehören, unterschieden werden: "Saal" und "Kiel" sind einsilbige Wörter, aber "Studie" und "Studium" sind dreisilbige Wörter.

Im weiteren gilt, daß Zahlen, Abkürzungen o.ä. gemäß ihrer Aussprache gewertet werden:

"1093": = eintausenddreiundneunzig : 1 Wort - 7 Silben;
"bzw.": = beziehungsweise : 1 Wort - 5 Silben;
"NATO": = [na:to:] : 1 Wort - 2 Silben;
"Sauberkeits- und Ordnungsprinzip": 3 Wörter - 3/1/4 Silben.[2]

Einige Probleme ergaben sich lt. ersten Berechnungen bei der Zählung von nullsilbigen Wörtern (z.B. "es" in "gibt's"). Aus diesem Grund werden diese Wörter als phonetischer Bestandteil der vorangehenden Silbe gewertet und somit nicht eigens in der Datentabelle aufgeführt.

Die in den untersuchten Texten vorkommenden Fremdwörter und ausländischen Namen werden entsprechend ihrer jeweiligen Ausspracheregelung notiert.

Bei der Auszählung blieben die Überschriften aller Texte unberücksichtigt (Fußnoten, Bildunterschriften etc. waren nicht vorhanden). Nur die Anreden der Texte 10 ("Meine Lieben!") und 11 ("Masuren!") gingen mit in die Datentabelle ein.

2. Die Ergebnisse der Untersuchung finden sich in den folgenden Tabellen. Damit man sich ein besseres Bild von den Texten machen kann, wird nicht nur die Lesebuchquelle angegeben, sondern auch noch eine Kurzcharakteristik nach Funktionalstil, Textsorte und einigen weiteren Aspekten durchgeführt.

In den Tabellen bedeuten:

x = die Wortlänge in Silben,
n_x = die absolute Häufigkeit der Wortlänge x im Text.
NP_{x1} = die theoretischen Werte nach der Positiven Poissonverteilung,
NP_{x2} = die nach der Positiven negativen Binomialverteilung,
NP_{x3} = die nach der Positiven Cohen - Poissonverteilung,
NP_{x4} = die nach der Positiven Cohen - Negativen Binomialverteilung.

[2] Vgl. Best & Zhu (1994:20).

Die jeweiligen Formeln dieser Verteilungen lauten:

(1) $$P_{x1} = \frac{a^x}{x!(e^a - 1)}, \quad x = 1,2,3,...$$

(2) $$P_{x2} = \frac{\binom{k+x-1}{x} p^k q^x}{1 - p^k}, \quad x = 1,2,3,...P_{x2}$$

(3) $$P_{x3} = \begin{cases} \frac{(1-\alpha)\,a}{e^a - 1 - \alpha a}, & x = 1 \\ \frac{a^x}{x!(e^a - 1 - \alpha a)}, & x = 2,3,4,... \end{cases}$$

(4) $$P_{x4} = \begin{cases} \frac{(1-\alpha)kqp^k}{1 - p^k - \alpha kqp^k}, & x = 1 \\ \frac{\binom{k+x-1}{x} p^{kq^x}}{1 - p^k - \alpha kqp^k}, & x = 2,3,4,... \end{cases}$$

X^2 ist der Wert des Chiquadrats, P die Überschreitungswahrscheinlichkeit des jeweiligen Chiquadrats und C das daraus berechnete Diskrepanzmaß, $C = X^2/N$. Die übrigen Größen sind Parameter. Die Modellierung gilt als zufriedenstellend, wenn $P \geq 0.05$ oder, bei größerem N, $C \leq 0.02$. Da für die Anpassung manche theoretischen Werte zu niedrig waren, mußten diese zusammengefaßt werden ("|" hinter der Zahl).

Zur Datenaufnahme:

Zu Text 3: Bei dem 8-silbigen Wort handelt es sich um "Autoreparaturstelle".
Zu Text 10: "Quijote": 1 Wort - 2 Silben. "Versailles": 1 Wort - 2 Silben. "Miguel": 1 Wort - 2 Silben.
Zu Text 11: "Menage": 1 Wort - 3 Silben. (Fremdwort aus dem Französischen). Die Überschrift wurde bei der Auszählung nicht mitberücksichtigt, da sie von den Lesebuchtextmitarbeitern hinzugefügt wurde.

Zu Text 13: Bei den zwei 10-silbigen Wörtern handelt es sich um die Zahl "dreitausendzweihundertzweiundzwanzig". "Toulon": 1 Wort - 2 Silben. "Ingenieur": 1 Wort - 4 Silben.
Zu Text 15: "Sergeant": 1 Wort - 2 Silben. "Colville": 1 Wort - 2 Silben. "Bore": 1 Wort - 1 Silbe. "Saloon": 1 Wort - 2 Silben. "Helene": 1 Wort - 2 Silben. "Hope": 1 Wort - 1 Silbe.

Die Ergebnisse sind in Tabelle 1-17 dargestellt.

Text 1: Ott, Inge: Die Geschichte vom rosaroten Regenschirm

x	n_x	NP_{x1}	NP_{x2}	
1	316	320.78	321.72	Text aus: Ott, Inge: Das Mäxchen und die Karolin. Stuttgart 1964. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 2. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1972. S. 84-85. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend kurze Sätze (1 - 5 Satzglieder), geringer Wortschatz
2	173	158.28	157.50	
3	40	52.06	51.80	
4	18	15.88	15.98	
		$a = 0.99$ $X_2^2 = 4.52$ $P = 0.10$	$k = 126.07$ $p = 0.99$ $X_1^2 = 4.58$ $P = 0.03$	

Text 2: Wölfel, Ursula: Der Nachtvogel

x	n_x	NP_{x3}	
1	272	268.90	Text aus: Wölfel, Ursula: Die grauen und die grünen Felder. Mülheim 1970. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 2. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1972. S. 112-113. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend mittellange Sätze (6 - 10 Satzglieder)
2	162	166.65	
3	25	26.07	
4	6	3.38	
	$a = 0.47$; $\alpha = .62$ $X_1^2 = 2.26$; $P = 0.13$		

Text 3: Höfle, Helga: Peter sammelt die Zeit

x	n_x	NP_{x4}	
1	397	396.58	Text aus: Höfle, Helga: Das gestreifte Krokodil. Recklinghausen 1971. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 3. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1973. S. 92-93. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, etwa zur Hälfte aus wörtlicher Rede bestehend, überwiegend mittellange Sätze, häufiges Vorkommen der Wörter 'Zeit' (37x), 'Stunden' und 'Minuten'
2	228	228.15	
3	47	47.30	
4	10	10.12	
5	2	2.20	
6	0	0.48	
7	0	0.10	
8	1	0.07	
	$k = 0.66$; $p = 0.77$ $\alpha = 0.66$; $X_1^2 = 0.01$; $P = 0.91$		

Text 4: Nöstlinger, Christine: Der kleine Jo

x	n_x	NP_{x1}
1	393	391.91
2	179	180.83
3	40	55.62 |
4	30	12.83 |
5	0	2.36 |
6	2	0.45 |
	$a = 0.92$; $X_1^2 = 0.03$; $P = 0.86$	

Text aus: Bödecker, Hans (Hrsg.): Die Kinderfähre. Stuttgart 1972. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 3. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1973. S. 117-118. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend mittellange Sätze

Text 5: Richter, Hans Peter: Der Ziegenbart

x	n_x	NP_{x1}
1	660	662.42
2	336	334.08
3	115	112.32
4	27	28.32
5	5	5.71
6	1	1.15
	$a = 1.01$; $X_4^2 = 0.25$ $P = 0.99$	

Text aus: Richter, Hans Peter: Das war eine Reise. Nürnberg 1962. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 4. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1973. S. 128-130. Funktionalstil: Stil der schönen Literatur Textsorte/Gattung: Kurzgeschichte für Kinder Textcharakteristika: Ich-Erzähler, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend mittellange Sätze

Text 6: Bichsel, Peter: Der Erfinder

x	n_x	Np_{x1}
1	686	694.54
2	335	324.01
3	105	100.77
4	13	23.50 |
5	5	4.38 |
6	1	0.68 |
7	1	0.09 |
8	2	0.03 |
	$a = 0.93$; $X_2^2 = 2.21$ $P = 0.33$	

Text aus: Bichsel, Peter: Kindergeschichten. Neuwied 1969. In: Pregel, Dietrich u.a. (Hrsg.): Texte für die Primarstufe. 4. Schuljahr. Hermann Schroedel Verlag KG, Hannover 1973. S. 142-144. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend mittellange Sätze, häufige Aneinandereihung der Haupt- und Nebensätze durch Komma oder 'und'

Text 7: Schubiger, Jürg: Der Meteorit

x	n_x	NP_{x4}
1	474	472.32
2	243	236.16
3	47	64.62
4	34	19.89 |
5	4	6.53 |
6	1	3.48 |
	$k = 0.000004$ $p = 0.59$; $\alpha = 0.59$ $X_0^2 = 7.78$ $C = 0.0096$	

Text aus: Schütz, Christel (Hrsg.): Das neue Narrenschiff. Frankfurt/Main 1980. S. 7ff. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 5. Ernst Klett Verlag, Stuttgart 1981. S. 97-98. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, etwa zur Hälfte aus wörtlicher Rede bestehend, überwiegend mittellange Sätze

Text 8: Wölfel, Ursula: Das blaue Wagilö

x	n_x	NP_{x4}	Text aus: Wölfel, Ursula: Das blaue Wagilö. Düsseldorf 1969. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 6. Ernst Klett Verlag, Stuttgart 1981. S. 6-10. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte für Kinder; Textcharakteristika: auktorialer Erzähler, Tempus - Präsens, z.T. wörtliche Rede, überwiegend mittellange Sätze.
1	1196	1195.12	
2	544	533.69	
3	82	99.25	
4	28	21.94	
	$k = 1.32$; $p = 0.83$ $\alpha = 0.56$; $X_0^2 = 4.88$ $C = 0.0026$		

Text 9: Senger, Valentin: Max Himmelreich

x	n_x	NP_{x2}	Text aus: Senger, Valentin: Kaiserhofstraße 12. Darmstadt 1978. S. 105-109. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 6. Ernst Klett Verlag, Stuttgart 1981. S. 62-65. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: autobiographische Erzählung (Romanauszug); Textcharakteristika: Ich-Erzähler, Tempus - Präteritum, z.T. wörtliche und indirekte Rede, überwiegend mittellange Sätze
1	668	670.15	
2	288	281.85	
3	95	102.69	
4	40	34.60	
5	13	11.08	
6	1	4.90	
	$k = 2.31$; $p = 0.75$ $X_2^2 = 1.81$; $P = 0.40$		

Text 10: Kästner, Erich: Ein Vorwort zu Don Quijote

x	n_x	NP_{x2}	Text aus: Kästner, Erich: Leben und Taten des scharfsinnigen Ritters Don Quichotte. Wien 1963. S. 3-6. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 7. Ernst Klett Verlag, Stuttgart 1982. S. 151-152. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Vorwort zu einer Nacherzählung; Textcharakteristika: Ich-Erzähler, z.T. direkte Anrede der Leser, Tempus - Präteritum, z.T. wörtliche Rede, überwiegend lange Sätze (mehr als 10 Satzglieder)
1	296	299.06	
2	176	162.80	
3	55	67.10	
4	23	23.22	
5	10	7.11	
6	0	1.99	
7	1	0.52	
8	1	0.20	
	$k = 6.37$; $p = 0.85$ $X_3^2 = 4.63$; $P = 0.20$		

Text 11: Aufruf zur Anwerbung von Bergarbeitern im Jahr 1908

x	n_x	NP_{x2}	Text aus: "Bergarbeiter Zeitung" vom 8. Aug. 1908. Zitiert nach: "Der Anschnitt". Zeitschrift für Kunst und Kultur im Bergbau/32/1980: 282-284. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 7. Ernst Klett Verlag, Stuttgart 1982. S. 40-41. Funktionalstil: Stil der Publizistik und der Presse; Textsorte/Gattung: Sachtext mit Appellcharakter (Plakatwerbung); Textcharakteristika: z.T. direkte Anrede der Leser in der veralteten Form '3. Person Singular', Tempus - Präsens, überwiegend mittellange Sätze, z.T. veralteter Wortschatz.
1	445	451.65	
2	305	289.82	
3	117	124.31	
4	43	40.09	
5	5	10.37	
6	4	2.76	
	$k = 375.34$; $p = 0.997$ $X_3^2 = 4.90$; $P = 0.18$		

Text 12: Drewitz, Ingeborg: Du und ich

x	n_x	NP_{x2}	Text aus: Ingeborg Drewitz (1983). In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 8. Ernst Klett Verlag, Stuttgart 1983. S. 31-32. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: (autobiographische?)Erzählung; Textcharakteristika: auktorialer/personaler Erzähler, Tempus - Präteritum, wenig wörtliche Rede, kurze und mittellange Sätze, z.T. Ellipsen.
1	497	500.92	
2	219	203.91	
3	62	74.65	
4	25	25.80	
5	5	8.60 |	
6	5	2.79 |	
7	4	0.89 |	
8	1	0.44 |	
	$k = 1.87$; $p = 0.72$ $X_2^2 = 3.73$; $P = 0.15$		

Text 13: Andres, Stefan: Das Trockendock

x	n_x	NP_{x2}	Text aus: Andres, Stefan: Die Verteidigung der Xanthippe. München 1961. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 8. Ernst Klett Verlag, Stuttgart 1983. S. 90-92. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte; Textcharakteristika: auktorialer Erzähler, Tempus - Präsens, wenig wörtliche Rede, überwiegend lange Sätze
1	523	531.84	
2	356	330.07	
3	120	139.93	
4	46	45.56	
5	14	12.15	
6	1	2.76 |	
7	1	0.55 |	
8	0	0.09 |	
9	0	0.01 |	
10	2	0.04 |	
	$k = 39.51$; $p = 0.97$ $X_3^2 = 5.41$; $P = 0.14$		

Text 14: Schönfeldt, Sybil Gräfin: Demo in den eigenen vier Wänden

x	n_x	NP_{x2}	Text aus: Schönfeldt, Sybil Gräfin: Hängt doch die Kinder in den Kamin. Neununddreißig unzeitgemäße Geschichten. München 1983. S. 91-95. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 9. Ernst Klett Verlag, Stuttgart 1984. S. 55-57.) Funktionalstil: Stil der Publizistik und der Presse; Textsorte/Gattung: Essay; Textcharakteristika: Auseinandersetzung mit einem Thema aus der Ich-Perspektive, Tempus - Präsens, überwiegend lange Sätze, z.T. Ellipsen, umgangssprachlicher Wortschatz
1	358	362.30	
2	213	201.08	
3	72	90.56	
4	52	36.05 |	
5	14	13.21 |	
6	1	6.80 |	
	$k = 3.60$; $p = 0.76$ $X_1^2 = 6.70$; $P = 0.01$ $C = 0.0094$		

Text 15: Jens, Walter: Bericht über Hattington

x	n_x	NP_{x2}	Text aus: Jens, Walter: Herr Meister. Dialog über einen Roman. München 1963. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 9. Ernst Klett Verlag, Stuttgart 1984. S. 162-164. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Kurzgeschichte; Textcharakteristika: Ich-Erzähler, Tempus - Präteritum, überwiegend mittellange und lange Sätze, häufige Aneinanderreihung von Hauptsätzen durch Komma oder Semikolon
1	515	529.18	
2	361	333.79	
3	137	140.88	
4	35	44.76	
5	10	11.41	
6	4	2.43 |	
7	0	0.44 |	
8	1	0.11 |	
	$k = 270.30$; $p = 0.995$ $X_3^2 = 6.40$; $P = 0.09$		

Text 16: Aichinger, Ilse: Wo ich wohne (1958)

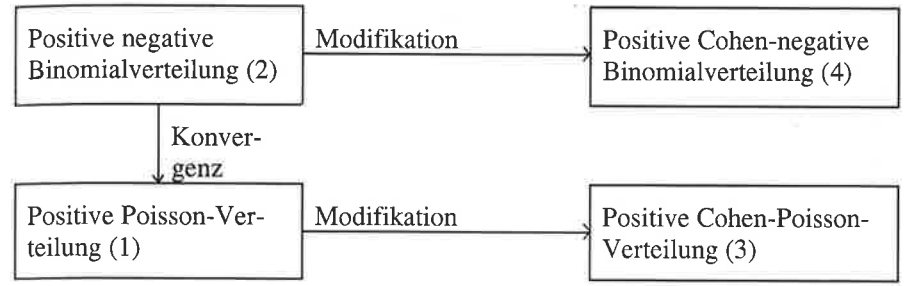
x	n_x	NP_{x4}	Text aus: Aichinger, Ilse: Wo ich wohne. Erzählungen, Gedichte, Dialoge. Frankfurt/Main 1963. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 10. Ernst Klett Verlag, Stuttgart 1985. S. 80-83. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Erzählung; Textcharakteristika: Ich-Erzähler, Tempus - Präsens/Präteritum, z.T. wörtliche Rede, überwiegend mittellange Sätze.
1	912	901.90	
2	497	504.77	
3	107	115.10	
4	34	26.35	
5	5	6.05	
6	1	1.83	
	$k = 0.95$; $p = 0.77$ $\alpha = 0.60$; $X_2^2 = 3.56$ $P = 0.17$		

Text 17: Frisch, Max: Der andorranische Jude

x	n_x	NP_{x2}	Text aus: Frisch, Max: Gesammelte Werke in zeitlicher Folge. Bd. 4. Frankfurt/Main 1976. S. 372. In: Hein, Siegfried u.a.: Lesezeichen. Lesebuch. Ausgabe A/B 10. Ernst Klett Verlag, Stuttgart 1985. S. 182-184. Funktionalstil: Stil der schönen Literatur; Textsorte/Gattung: Erzählung (Dramaentwurf); Textcharakteristika: auktorialer Erzähler, Tempus - Präteritum, überwiegend mittellange und lange Sätze.
1	402	400.73	
2	199	205.04	
3	82	79.40	
4	34	25.81	
5	4	10.02	
	$k = 6.39$; $p = 0.86$ $X_1^2 = 0.40$; $P = 0.53$		

3. Zusammenfassung und Interpretation der Ergebnisse

Es kann festgestellt werden, daß nicht alle Texte des vorliegenden Korpus grundsätzlich der 0-gestutzten (positiven) negativen Binomialverteilung hinsichtlich der Wortlängenhäufigkeiten folgen. Es ergeben sich hier einige Modifikationen, die eher genre- und stilbedingt sind. Eine der Modifikationen ist "lokal" und führt zu der positiven Cohen-negativen Binomialverteilung (4), die andere ist eine Konvergenz gegen die positive Poisson-Verteilung (1), die man immer dann anwenden kann, wenn $k \to \infty$, $q \to 0$ und $kq \to a$. Diese wiederum wird auf die gleiche Weise modifiziert, woraus sich die positive Cohen-Poisson-Verteilung (3) ergibt. Für das Deutsche ergibt sich daher als Modell:



Die Texte können in zwei Gruppen unterteilt werden: die einen (Text 1-8) richten sich an Kinder (bis ca. 12 Jahre), die anderen (Text 9-17) eher an Jugendliche (ab 13 Jahren). Während sich aus der ersten Gruppe kein Text nach der negativen Binomialverteilung modellieren läßt, so trifft dies in der zweiten Gruppe nur für Text 16 zu. Für alle Texte konnten aber geeignete Modelle gefunden werden.

Text 1 scheitert an einem etwas zu niedrigen Wahrscheinlichkeitswert (P = 0.03) und paßt sich besser der Poissonverteilung an; der Koeffizient C ist bei beiden Anpassungen gleich gut. Eine gute Poissonverteilung erreichen ebenfalls die Texte 4, 5 und 6. In Text 2 ist eine Cohen-Poissonverteilung zu erkennen. Text 3 weist die Cohen-negative Binomialverteilung auf. Für die Texte 7 und 8 lassen sich lediglich Cohen-negative Binomialverteilungen ausmachen, jedoch bleiben hier keine Freiheitsgrade übrig, sodaß wir hier die Anpassungsgüte nur mit C messen können.

Mit wenigen Ausnahmen zeigen sich in der zweiten Textgruppe weitaus einheitlichere Ergebnisse. Bis auf Text 16 weisen alle anderen eine akzeptable negative Binomialverteilung auf. Darüberhinaus lassen sich allerdings noch bessere Werte mit einer Poissonverteilung erzielen; dies gilt für die Texte 11, 13 und 15 mit P = 0.30, 0.22 und 0.18. Zwar können auch für die Texte 10 und 17 Poissonverteilungen berechnet werden, doch mit schlechteren Werten. Der niedrige

Wahrscheinlichkeitswert ($P = 0.01$) bei der negativen Binomialverteilung in Text 14 deutet auf einen weniger homogenen Text hin, der sich wahrscheinlich durch den eher essayistischen Stil mit journalistischer Färbung erklären läßt. Aus der Reihe fällt Text 16, dem ähnlich wie den Texten 3, 7 und 8 die Cohen-negative Binomialverteilung zuzuordnen ist.

In Anbetracht dieser Darstellung ist zu vermuten, daß die negative Binomialverteilung nicht gut auf Texte, die sich primär an Kinder richten, anwendbar ist. Lediglich Text 1 (Ott) und die Kindergeschichte von Bichsel (Text 1 in Best & Zhu, 1994) lassen sich bei schwachem P mit dieser Verteilung modellieren. Durch die Verstellung der autoreigenen Sprache, wobei der Autor versucht, eine kindgerechte Sprache zu wählen, entstehen wahrscheinlich Texte, die sich nicht den gleichen Modellen unterwerfen wie die Texte, die für Jugendliche oder Erwachsene geschrieben sind.

Im weiteren zeigt sich in bezug auf die erste Textgruppe eine relativ breite Streuung bei der Wahl einer passenden Verteilung hinsichtlich der Wortlängenhäufigkeit: Poisson-, Cohen-Poisson- und Cohen-negative Binomialverteilung. In der zweiten Gruppe beschränkt sich die Auswahl weitestgehend auf die Poisson- und negative Binomialverteilung (Ausnahme Text 16: Cohen-negative Binomialverteilung). Hierbei drängt sich die Hypothese auf, daß unter den Texten, deren Sprache sich an Erwachsene wendet, eine größere Homogenität zumindest bzgl. der Verteilung der Wortlängenhäufigkeit herrscht als bei jenen, deren Sprache kindgerecht erscheinen soll. Denn diese Sprache ist eine dem Autor nicht eigene, sondern vielmehr verfremdete Sprache, die somit nicht in sich stimmig sein kann. Weitere Untersuchungen müßten sich hieran anschließen, z.B. Bestimmung der Wortlängenhäufigkeit eines Autors mit Texten, die sich an unterschiedliche Altersgruppen richten.

Literatur

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In diesem Band.

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. (Erscheint.)

Best, K.-H., & Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, U. (Hrsg.), *Computatio Linguae II.* Stuttgart, 19-30.

Bünting, K.-D., & Bergenholtz, H. (1989). *Einführung in die Syntax.* Frankfurt/Main.

Glottometrika 15, 1996, 195-204

Word Length Distribution in Turkish Texts

Gabriel Altmann, Bochum
Erkan Erat, Göttingen
Luděk Hřebíček, Prague

This is an attempt to apply to Turkish the word-length theory which was formulated by Grotjahn & Altmann (1993), Wimmer et al. (1994) and Altmann & Best (1996). According to this theory an author, a genre or a language, when forming the actual word length distribution, follow some preferred probability model. This preferred model represents a kind of attractor (as used in catastrophe theory) or in other words, a kind of minimum where the word length finds an equilibrium or steady state which is a characteristic state of texts or languages. There may be several basic models even for one and the same author depending on whether he writes in the same genre or whether the boundary conditions were the same in every case. If the *ceteris paribus* condition is fulfilled, then we expect the adequacy of the given model in all texts under consideration. If it is not fulfilled, then there are two ways to find a new attractor:

(i) The so called perfect-delay dynamics by which the author takes the next possible minimum in order to find an equilibrium. This can be performed by local modification of the individual classes of the basic distribution. These cases seem to be refutations of the model but we consider them as ever-present fluctuations, innovations and first signals of the fact that the governing system begins to be unstable, possibly passes to chaos and must be organized in a different way.

(ii) The so-called Maxwell dynamics by which the author or language seeks the deepest minimum. In this case the text wanders to another probability distribution and we say that it has found a new attractor.

The above theory says that the probability of length x is controlled by that of $x - 1$ in such a way that P_x is proportional to P_{x-1} and the proportionality which organizes this relation is a function $g(x)$, thus

$$(1) \quad P_x = g(x)P_{x-1}.$$

The more complicated case reported in Wimmer et al. (1994) was applicable but not very pertinent in Turkish. In case (i), some (or all) P_x are modified by a scalar so that ΣP_x remains 1; in case (ii) either a change in $g(x)$ takes place or the order

of the difference equation (1) changes, or both.

On the basis of a series of empirical explorations we specified the basic proportionality function holding in Turkish as

$$(2) \qquad g(x) = \frac{a - bx}{cx}$$

($x = 1,2,...,a/b$). Inserting (2) in (1) we obtain

$$(3) \qquad P_x = \frac{a - bx}{cx} P_{x-1} = \frac{b}{c} \cdot \frac{\frac{a}{b} - x}{x} P_{x-1}.$$

Setting $a/b = n+1$, $b/c = \theta$ the expression

$$(5) \qquad P_x = \frac{n - x + 1}{x} \theta P_{x-1}$$

is obtained ($P_0 = 0$, $x = 2,3,...,n$). Solving this simple difference equation we obtain

$$P_x = \frac{1}{n\theta}\binom{n}{x}\theta^x P_1.$$

Under the condition that

$$1 = \sum_{x=1}^{n} P_x = \frac{1}{n\theta} P_1 \sum_{x=1}^{n} \binom{n}{x} \theta^x = \frac{P_1}{n\theta}[(1 + \theta)^n - 1],$$

so that

$$P_1 = \frac{n\theta}{(1 + \theta)^n - 1},$$

we can write

$$P_x = \binom{n}{x} \frac{\theta^x}{(1 + \theta)^n - 1}.$$

If we write $\theta/(1+\theta) = p$, $1/(1 + \theta) = q$, and divide the fraction by $(1+\theta)^n$, the final formula that we are seeking is then

$$(6) \qquad P_x = \binom{n}{x} \frac{p^x q^{n-x}}{1 - q^n}, \qquad x = 1,2,...,n.$$

This is the positive binomial distribution (or zero-truncated binomial distribution) fitted to the following Turkish texts (s. Table 1):

TEXT 1: F.N. Çamlibel: Han Duvarlari. In: Ilhami Soysal (ed.) *20. Yüzyil Türk Şiiri Antolojisi.* Bilgi, 1973, 404-408.
TEXT 2: E. Özkök: Vatandaşin Gözü, Kaya Erdem'im Odasinda. *Hürriyet* 12.3.1990, 17.
TEXT 3: A. Nesin: Elbet bir Bildiğimiz Var. In: Aziz Nesin: *Gidigidi. Istanbul: Cem,* 1973, 49-53.
TEXT 4: O. Akbal: Önsöz. In Oktay Akbal: Şair Dostlarim. Istanbul: Elif, 1964, 7-9.
TEXT 5: N. Uğurlu: Türk Dili mi? Türk Dilleri mi?. *Türk Dili*, Sayi 498/ Haziran, 1993, 407-472.
TEXT 6: E. B. Lâv: Gidişat. In: Ilhami Soysal (ed.) *20. Yüzyil Türk Şiiri Antolojisi.* Bilgi, 1973, 364-367.
TEXT 7: H. Balikçisi. Kara Gece. In: Halikarnas Balikçisi: *Gülen Ada.* Istanbul: Yeditepe, 1957, 69-71.
TEXT 8: A. Nesin: Kelepir bir Işçi. In: Aziz Nesin: *Geriye Kalan.* Istanbul: Cem-May, 1982, 26-29.
TEXT 9: N. Cumali: Sincap. In: Necati Cumali: *Revizyonist.* Ankara: Tekin, 1979, 130-131.
TEXT 10: Maliye, Istanbul'un 50 milyarini vermiyor. *Hürriyet,* 13.3.1990.
TEXT 11: Tarik Dursun K.: Zühre. In: Tarik Dursun K.: *Yabanin Adamlari.* Istanbul: Kurul, 1966, 80-83.
TEXT 12: E. Barutçu: Akilli bir tercih. *Cumhuriyet HAFTA,* 18.12.-24.12.1992, 15.
TEXT 13: Çok satan mi yoksa, çok okunan mi olmali. *Dünya,* 23.12-29.12.1992, 8.
TEXT 14: S. Ketenci: Almanya'da Yaşamak. *Cumhuriyet HAFTA 2, 10.12.*199, 14.
TEXT 15: Gazete okuruyla satin alanlar farkli olabilir. *Dünya,* 27.1.-2.2.1993, 8.
TEXT 16: T. Erdem: Türkiye'de Sosyal Demokrasinin Geleceği. *Cumhuriyet HAFTA,* 14.5. - 20.5.1993, 2
TEXT 17. T. Ates: DYP'nin Sonu mu? *Cumhuriyet HAFTA,* 14.5. - 20.5.1993, 4.
TEXT 18. E. Çölasan: Baba'nin Hayirduasi. *Hürriyet,* 11.6.1993, 11.

Table 1
Fitting the positive binomial distribution to some Turkish texts

x	Text 1		Text 2		Text 3		Text 4	
	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	134	142.24	94	92.25	168	172.10	106	100.08
2	266	261.52	164	169.28	262	248.27	134	150.15
3	252	240.41	176	172.57	182	191.02	149	136.52
4	109	110.50	107	105.56	77	82.67	83	83.78
5	14	20.33	38	38.74	24	19.08	35	36.56
6			4	7.89	2	1.86	12	11.63
7			3	0.69			3	3.28
			1	0.02				
	$n = 5$ $p = 0.4790$ $X_2^2 = 3.10$ $P = 0.21$		$n = 7$ $p = 0.3795$ $X_3^2 = 0.34$ $P = 0.95$		$n = 6$ $p = 0.3659$ $X_3^2 = 2.95$ $P = 0.40$		$n = 12$ $p = 0.2143$ $X_4^2 = 3.33$ $P = 0.50$	

x	TEXT 5		TEXT 6		TEXT 12		TEXT 13	
	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	160	155.76	65	63.83	122	126.04	185	186.26
2	227	245.09	131	127.51	200	198.48	299	290.31
3	241	224.95	122	127.35	198	187.53	272	276.52
4	145	132.73	62	63.59	111	118.11	168	179.58
5	39	52.21	13	12.70	50	52.07	94	83.97
6	11	13.69	3	0.02	20	16.40	28	29.08
7	4	2.57			1	3.69	8	7.55
8					1	0.68	1	1.73
	$n = 9$ $p = 0.2823$ $X_4^2 = 8.43$ $P = 0.08$		$n = 5$ $p = 0.4997$ $X_2^2 = 1.30$ $P = 0.52$		$n = 11$ $p = 0.2395$ $X_3^2 = 1.32$ $P = 0.73$		$n = 13$ $p = 0.2062$ $X_5^2 = 2.65$ $P = 0.75$	

x	TEXT 14		TEXT 15		TEXT 16	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	76	72.03	166	182.93	80	70.03
2	132	134.99	357	324.73	110	128.39
3	141	149.90	331	341.59	139	139.49
4	120	109.24	226	235.81	113	99.45
5	52	54.59	118	111.62	47	48.62
6	19	18.94	35	36.69	13	16.50
7	3	4.50	8	8.27	4	3.84
8	2	0.81	2	1.36	0	0.58
9					0	0.05
10					1	0.05
	$n = 10, \quad p = 0.2940$ $X_4^2 = 2.01,$ $P = 0.73$		$n = 10, \quad p = 0.2829$ $X_5^2 = 6.29, \quad P = 0.28$		$n = 10, \quad p = 0.2895$ $X_4^2 = 6.76, P = 0.15$	

x	TEXT 17		TEXT 18	
	f_x	NP_x	f_x	NP_x
1	116	116.92	101	108.83
2	184	174.66	185	180.75
3	143	159.45	194	171.54
4	106	99.25	80	101.75
5	48	44.48	43	38.62
6	13	14.76	7	9.16
7	3	3.68	2	1.35
8	1	0.81		
	$n = 13, \ p = 0.1993$ $X_4^2 = 3.20, \ P = 0.52$		$n = 8, \ p = 0.3218$ $X_4^2 = 9.61, \ P = 0.05$	

From Table 1 it is evident that this model, regardless of its high correspondence with the observed data in 13 texts out of 18, does not yield satisfactory results for the other texts. Under the assumption that the model is correct, we can try to cope with the anomalies in three different ways:

(i) A more adventurous way is to search for the reasons in the agglutinative character of Turkish. This involves the danger that the same grounds must hold true in all strongly agglutinative languages. A preliminary investigation has shown, however, that Finnish and Korean do not fit the binomial model at all. On the other hand, there are "weakly agglutinative" languages, e.g. Czech and Polish, following the binomial model.

(ii) We could ask the individual authors - who could not answer any question concerning their word length distribution - or one could analyze all their writings - but there is no answer to our question in them.

(iii) We can consider the probability distribution as a kind of attractor, a form existing in any language - i.e. existing unconsciously in the text users - to which the empirical distributions of the given variable tend. Of course, there can be a number of different attractors exerting their impact on individual writers or on individual genres, and each of them can evolve in the course of time. As a matter of fact, formula (1) merely represents a mechanism which can take into account a number of boundary or subsidiary conditions, as is the case in natural laws, too. In practice, if a text deviates from the supposed attractor, we say that it wanders to another attractor, which is quite a normal circumstance in the life of a text producer. This wandering can be expressed in different ways (cf. Wimmer et al., 1994), e.g. in the modification of $g(x)$, in the increase of the order of the difference equation (1), in the modification of the individual frequency classes, in the mixing, compounding or convolution of probability distributions, etc. We shall try to take the third way, on many grounds.

Considering the data diverging significantly from the positive binomial model it was observed that there exists a systematic local shift destabilizing the original attractor: classes $x = 1$ and $x = 3$ have an excess of frequencies while $x = 2$ has a shortage of them. In cases of this kind it can be assumed that the authors are "halfway" to another attractor, leaving the original one but still retaining and modifying it, or that they seek the easiest way to retain the original model in spite of the fact that the boundary conditions changed. In agreement with the observation we take an α-part of P_1 and P_3 and attach them to P_2. Thus we obtain a modified distribution which can be written as

$$P_1' = P_1(1 - \alpha)$$

$$P_2' = P_2 + \alpha(P_1 + P_3)$$

$$P_3' = P_3(1 - \alpha)$$

$$P_x' = P_x, \qquad x = 4,5,\ldots,n$$

In our case we then obtain:

$$(7) \qquad P_x' = \begin{cases} \dfrac{npq^{n-1}(1-\alpha)}{1-q^n}, & x = 1 \\ \dfrac{npq^{n-3}}{1-q^n}\left[\dfrac{(n-1)pq}{2!} + \dfrac{\alpha(n-1)(n-2)p^2}{3!} + \alpha q^2\right], & x = 2 \\ \dfrac{\binom{n}{3}p^3q^{n-3}}{1-q^n}(1-\alpha) & x = 3 \\ \dfrac{\binom{n}{x}p^xq^{n-x}}{1-q^n}, & x = 4,5,.. \end{cases}$$

Modifications of this kind are usual in other sciences too. They give new impulses to both the given empirical science and to probability theory.

The results of fitting model (7) are presented in Table 2. The systematic caracter of this shift is emphasized by the quite constant value of the modification paramerter $\alpha \approx 0.19$.As can be seen in Table 2, four texts (No 7 to 10) show very good agreement with this model. The text by Tarik (No 11) however, did not follow either of these models. Again, we can use the same argument as in (iii) and assume that Tarik already left the binomial model and wandered definitely to another one.

Table 2
Fitting the modified positive binomial distribution to Turkish texts

x	Text 7		Text 8		Text 9	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	113	109.36	47	47.05	39	38.91
2	281	292.70	139	136.65	126	122.27
3	177	170.65	83	83.45	84	82.80
4	115	117.85	52	56.39	73	70.75
5	45	39.50	15	18.46 |	31	31.21
6	5	7.36 |	6	2.00 |	2	8.60
7	2	0.59 |			0	1.36 |
8					1	0.09 |
	$n = 7$, $p = 0.3584$ $\alpha = 0.1911$ $X_2^2 = 1.77$, $P = 0.41$		$n = 6$, $p = 0.4218$ $\alpha = 0.1903$ $X_1^2 = 0.74$, $P = 0.39$		$n = 8$, $p = 0.3554$ $\alpha = 0.1934$ $X_3^2 = 5.41$, $P = 0.14$	

x	Text 10		Text 11	
	f_x	NP_x	f_x	NP_x
1	77	70.67	93	104.97
2	178	195.32	245	252.86
3	138	121.86	157	127.99
4	95	97.04	59	72.01
5	34	41.71	22	17.42
6	16	11.95	1	1.76
7	2	2.20 |		
8	0	0.24 |		
9	1	0.01 |		
	$n = 9$, p = 0.3006, $\alpha = 0.1903$ $X_3^2 = 7.20$, $P = 0.07$		$n = 6$, p = 0.3769, $\alpha = 0.1938$ $X_2^2 = 11.54$, $P = 0.003$	

A number of models can be found; however, for the time being we prefer to stick to a model that has already been found in other languages (e.g. Italian, Slovak and older German texts[1]), namely the hyper-Poisson distribution which can be obtained from the difference equation

$$(8) \qquad P_x = \frac{a}{c + x} P_{x-1}$$

and after reparametrization has the form

$$(9) \qquad P_x = \frac{a^x}{b^{(x)} T}$$

($x = 0,1,2,...$) where T is the norming constant and $b^{(x)} = b(b+1)...(b+x-1)$. At the same time text No. 10 also shows a much better agreement with this model than with the other ones. The fitting of this model to texts No. 10 and 11 gave the result presented in Table 3.

Table 3
Fitting the 1-displaced hyper-Poisson distribution to two Turkish texts

x	Text 10		Text 11	
	f_x	NP_x	f_x	NP_x
1	77	81.65	93	91.82
2	178	166.10	246	241.91
3	138	180.45	157	161.25
4	95	87.65	59	61.52
5	34	37.64	22	16.44
6	16	12.79	1	4.06
7	2	3.60		
8	0	0.86 |		
9	1	0.26 |		
	$a = 1.6329$, $b = 0.8027$ $X_5^2 = 4.64$, $P = 0.46$		$a = 0.8923$ $b = 0.3387$ $X_2^2 = 0.58$ $P = 0.75$	

[1] Personal communications from K.H. Best and L. Gaeta.

In this way we have demonstrated on Turkish data (i) the adequacy of the proposed word-length theory, showing the simple kind of control, and in the form of $g(x)$, the element of self-organization which (ii) may differ even within one author's work and has different forms that do not necessarily depend on the "type" of language but are phenomena of innovation. Extensive research would be necessary in order to determine the extent of this language property in Turkish. However, it is to be expected that the majority of texts analyzed will display an agreement with one of the above three models.

References

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length. Some methodological problems. In: Köhler, R., & Rieger, B. (eds.), *Contribu tions to quantitative linguistics.* Dordrecht: Kluwer, 141-153.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distributions. *J. of Quantitative Linguistics 1,* 98-106.

Altmann, G., & Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In this volume.

Nemcová, E., & Altmann,G. (1994). Zur Wortlänge in slowakischen Texten. *Z. für empirische Textforschung 1,* 40-43.

Glottometrika 15, 1996, 205-213

Zur Wortlänge in koreanischen Texten

Icheon Kim, Düsseldorf
Gabriel Altmann, Bochum

1. Bei Wortlängenzählungen geht man üblicherweise von der orthographischen Form der Wörter aus, die man dann phonologisch zu deuten versucht. Im Koreanischen hat sich durch die Rechtschreibungsreform 1988 einiges verändert; eine einheitliche Orthographie wurde jedoch nicht vollständig erreicht. Dies ist nichts Besonderes und spiegelt lediglich einen Trend wieder, der sich im Bereich der Wortlänge immer wieder zeigt. Da angewandte Linguisten die Wortlänge völlig außer Acht lassen und die Rechtschreibung morphologisch oder phonologisch begründen, besteht keine Garantie, daß sie im Sinne der sprachlichen Selbstregulation handeln und genau das nachvollziehen, was die Sprache "verlangt". Zumindest wird aber eine Annäherung gesucht.

Im Koreanischen ergaben sich einige Probleme mit der Schreibweise der Familien- und Vornamen (zusammen oder getrennt, z.B. *Kimicheon* oder *Kim Icheon),* der Eigennamen, z.B. *Han'guktaehakkyo sabŏmdaehak* bzw. *Han'guk taehakkyo sabŏm taehak* [Han'guk Universität Pädagogische Fakultät], der terminologischen Ausdrücke, die früher getrennt geschrieben wurden, heute aber auch als ein Wort geschrieben werden können, z.B. *junggŏri t'andot'an* bzw. *junggŏrit'andot'an* [Mittelstreckenrakete], der Zahlwörter, die man früher getrennt schrieb, z.B. 1994: *ilchŏn kubaek kusip sa,* heute aber unter "zehntausend" zusammenschreibt, z.B. 107898: *Simman ch'ilch'ŏnp'albaekkusipp'al,* sonst jedoch trennt. Im allgemeinen hielten wir uns an die im Text vorgefundene orthographische Version, wodurch möglicherweise einige Anomalien entstanden sind.

Es wurden relativ willkürlich 24 Texte ausgewählt. Die Texte 1 bis 9 stammen aus den Schulbüchern *Kugŏ* [Landessprache o. Koreanisch]; die Texte 10-11 aus der Sonntagsbeilage in *Han'guk Ilbo; die* Texte 12-20 sind moderne Erzählungen aus *Sin Han'guk munhak chŏnjip Bd. 12: Ch'oe Chŏnghŭi sŏnjip* [Ausgewählte Werke von Ch'oejonghui], Seoul, Omun'gak Verlag 1972, und Bd. 14: *Hwang Sunwŏn sŏnjip* [Ausgewählte Werke von Hwang Sunwon], Seoul, Omun'gak Verlag 1972. Die Texte 21-24 stammen aus der Monatszeitschrift *Sindonga.*

Es handelt sich im einzelnen um folgende Texte:

T 1 *Choŏn.* In: Kugŏ für das 9. Schuljahr, 1983, 22-29.
T 2 *Yun Ponggil.* Ebenda, 32-39.
T 3 *Kojŏg-ŭl ch'ajasŏ* [Auf der Suche nach den altehrwürdigen Stätten]. Eben da, 46-55.
T 4 *Saero naon talnim* [Der Mond geht wieder auf]. Ebenda, 92-105.
T 5 *Abŏji-ŭi sŏnsaengnim* [Vaters Lehrer]. Ebenda, 134-145.
T 6 *Koji-ŭi t'aegŭkki* [Flagge auf der Anhöhe]. Ebenda, 148-159.
T 7 Samil chŏngsin [Der Geist der 1.März Bewegung]. In: Kugŏ für das 11. Schuljahr, 8-17.
T 8 *Sejong taewang* [König Sejong]. Ebenda, 34-43.
T 9 *Simch'ŏng iyagi.* [Die Geschichte von Simch'ong]. Ebenda, 96-109.
T 10 *Pak Kwŏnsam siron* [Kommentar von Park Kwonsam]. Sonntagsbeilage in Han'gukilbo, Han'gukilbosa 6.2.,1994.
T 11 *Yuryŏng chaep'an* [Gespenstergericht]. Ebenda.
T 12 *Hwanyŏng*[Illusion]. In: *Sin Han'guk munhak chŏnjip Bd. 12: Ch'oe Chŏng ŭi sŏnjip* [Neue koreanische Literaturwerke Bd. 12: Ausgewählte Werke von Ch'oe Chŏnghŭi]. Omun'gak Verlag: Seoul 1972, 185-190.
T 13 *Ch'ongnyangniyŏk kŭnch'ŏ* [Nähe der Ch'ongnyangni Station]. Ebenda, 375-377.
T 14 *Pegaenmo* [Die Verzierungen an koreanischen Kopfkissen]. Ebenda, 358-362.
T 15 *Ch'urak toen pihaenggi* [Das abgestürzte Flugzeug]. Ebenda, 190-191.
T 16 *Usan-ŭl chŏbŭmyŏ* [Während man den Regenschirm zumacht]. Ebenda Bd. 14:*Hwang Sunwŏn sonjip* [Ausgewählte Werke von Hwang Sunwon], 521-524.
T 17 *Talg-wa palg-wa* [Mond und Fuß]. Ebenda, 471-475.
T 18 *P'i* [Blut]. Ebenda, 525-530.
T 19 *Ŏdum-sok-e tchik'in p'anhwa* [Der Holzschnitt, der in der Dunkelheit gedruckt wurde]. Ebenda, 483-489.
T 20 *P'ilmuk changsu* [Der Pinselhändler]. Ebenda, 508-513.
T 21 *P'ŭransŭ yŏnghwa-ga millyŏ-onda* [Französiche Filme dringen in Korea ein]. *Sindonga 393,* 1992, 126-127.
T 22 *Hyŏndae Chadongch'a Mabungni yŏn'guso* [Das Forschungsinstitut von Hyondae in Mabungni]. *Sindonga 392,* 527-529.
T 23 *Yi Insik chŏ "saram-gwa k'ŏmp'yut'ŏ"* ["Mensch und Computer" von Yi Insik]. *Sindonga 391,* 609-611.

T 24 *Kim Taejung, "hwahae-ŭi sidae"ikkŭl sinŭi-ŭi chŏngch'i-in* [Kim Taejung, der vertrauenswürdige Politiker, der das "Zeitalter der Versöhnung" anführt]. Ebenda 396, 157-162.

In Sprachen, die eine starke Agglutination haben (Koreanisch, Japanisch, Finnisch, Ungarisch, Türkisch), ist die Morphologie recht strikt geregelt und läßt relativ wenige Ausnahmen zu. Die Wortlängenverteilung ist ziemlich steif und richtet sich nach einem Attraktor. Ein Autor hat natürlich immer die Möglichkeit, die Ordnung zu stören, aber die Anziehungskraft des Attraktors ist in diesen Sprachen sehr stark. Abweichungen erfolgen nicht so sehr durch einen Wechsel von einem Attraktor zum anderen, sondern durch Modifikation des gegebenen, meistens beschränkt auf einige Längenklassen. In der Realität verläuft es so, daß eine Klasse im Vergleich zum Modell zu wenig Mitglieder hat, die Nachbarklasse zu viel. Diese Verschiebungen von Häufigkeiten sind nicht sprachabhängig, sondern eher stilistischer Natur, d.h. einzeltextabhängig. Die Ursachen der Abweichung werden wir möglicherweise nie erfahren, da uns die Randbedingungen der Texterzeugung niemals in ihrer Vollständigkeit bekannt sein werden.

2. Bei der Suche nach einem Modell hat sich herausgestellt, daß die "beste" Funktion von x in $P_x = g(x)P_{x-1}$ für koreanische Wortlänge die Menzerathsche Funktion $g(x) = ax^{-b}$ ist, die sich aber in mehreren Varianten und Modifikationen niederschlägt, wobei noch drei ungelöste Fälle übrigblieben. Das Grundmodell (vgl. Wimmer et al., 1994)

$$P_x = \frac{a^x}{(x!)^b T}, \quad x = 0,1,2,...$$

$$a, b > 0$$

$$T = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b},$$

wobei T immer die Normierungskonstante ist, stellt die Conway-Maxwell-Poisson-Verteilung dar. Da die kleinste Wortlänge im Koreanischen $x = 1$ ist, muß zum Modellieren entweder die 1-verschobene oder die 0-gestutzte Form benutzt werden. Im ersten Fall bekommen wir

$$(1) \qquad P_x = \frac{a^{x-1}}{(x-1)!T}, \qquad x = 1,2,...$$

$$T = \sum_{j=1}^{\infty} \frac{a^{j-1}}{(j-1)!T}.$$

Diesem Modell folgten 10 Texte, wie in Tabelle 1 zu sehen ist. Text 19 liefert noch eine akzeptable Anpassung, da $C = 0.0047$ ist.

Zwei Texte ziehen jedoch die 0-gestutzte Form vor, nämlich

$$(2) \qquad P_x = \frac{a^x}{(x!)^b T}, \qquad x = 1,2,3,...,$$

mit $T = \sum_{j=1}^{\infty} \frac{a^j}{(j!)^b}$, wie in Tabelle 2 zu sehen ist. Bei Text 18 ergibt der Koeffizient $C = 0.0049$, was ein gute Anpassung signalisiert.

Tabelle 1
Anpassung des Modells (1) an 10 Texte

x	Text 1		Text 2		Text 5	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	51	55.46	63	61.04	87	94.88
2	179	178.87	239	237.73	256	251.83
3	186	188.51	242	238.59	245	242.69
4	111	103.27	99	108.33	127	129.30
5	42	35.56	30	28.01	53	45.24
6	2	8.54	3	4.68	6	11.42
7	1	1.79	1	0.54	4	2.64
8			2	0.08		
	$a = 3.23$; $b = 1.61$ $X_2^2 = 0.99$; $P = 0.61$		$a = 3.89$; $b = 1.96$ $X_3^2 = 1.16$; $P = 0.76$		$a = 2.65$; $b = 1.46$ $X_4^2 = 5.45$; $P = 0.24$	

Tabelle 1 (Fortsetzung)

x	Text 7		Text 14		Text 15	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	53	45.45	168	164.73	38	39.27
2	268	280.75	611	619.01	148	147.22
3	356	338.78	620	622.06	153	153.96
4	141	157.27	300	289.00	78	76.30
5	43	37.07	80	77.66	22	22.26
6	3	5.16	7	13.65	5	4.99
7	1	0.52	2	1.89		
8						
	$a = 6.18$; $b = 2.36$ $X_3^2 = 5.83$; $P = 0.12$		$a = 3.76$; $b = 1.90$ $X_4^2 = 3.91$; $P = 0.42$		$a = 3.75$; $b = 1.84$ $X_3^2 = 0.09$; $P = 0.99$	

x	Text 13		Text 19		Text 20		Text 21	
	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	98	97.02	251	281.96	164	148.83	17	21.99
2	293	290.59	792	759.34	515	512.24	129	108.22
3	292	289.55	735	719.55	579	571.92	167	169.16
4	140	151.55	341	370.12	316	330.51	114	135.20
5	60	50.23	126	123.41	102	119.71	71	67.14
6	8	11.68	35	29.40	36	30.17	28	23.05
7	2	2.38	9	5.32	8	6.62	4	5.85
8			1	0.90			2	1.39
	$a = 3.00$; $b = 1.59$ $X_4^2 = 4.04$ $P = 0.40$		$a = 2.69$; $b = 1.51$ $X_4^2 = 10.90$ $P = 0.03$		$a = 3.44$; $b = 1.62$ $X_4^2 = 6.33$ $P = 0.18$		$a = 4.92$; $b = 1.65$ $X_5^2 = 10.65$ $P = 0.06$	

Tabelle 2

x	Text 10		Text 18	
	f_x	NP_x	f_x	NP_x
1	32	32.90	202	203.05
2	94	94.20	634	628.76
3	142	127.27	716	687.95
4	81	100.90	320	359.80
5	61	52.91	104	106.15
6	16	19.79	25	19.62
7	8	5.56	6	2.44
8	1	1.47	1	0.23
	$a = 10.34$; $b = 1.85$ $X_5^2 = 8.84$; $P = 0.12$		$a = 18.33$; $b = 2.57$ $X_3^2 = 9.87$; $P = 0.02$	

In allen weiteren Texten ergab sich die Notwendigkeit, einige Klassen lokal zu modifizieren. In 5 Texten erwies sich die Klasse $x = 1$ als am stärksten abweichend und mußte separat behandelt werden. In solchen Fällen ist es üblich, die Modifikation mit einem separaten Parameter darzustellen, z.B.

$$(3) \qquad P_x = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1-\alpha)a^x}{(x!)^b T}, & x = 2,3,... \end{cases}$$

mit $T = \sum_{j=2}^{\infty} \frac{a^j}{(j!)^b}$. Die Resultate sind in Tabelle 3 zu sehen. Hier ist überall $\alpha = f_1/N$. In Text 17 ist $P = 0.047 \sim 0.05$, aber $C = 0.0049$.

In weiteren 4 Texten wurde beobachtet, daß Klasse $x = 4$ überbelegt ist, wenn man sie mit Modell (2) vergleicht, während Klasse $x = 3$ unterbelegt ist. Hier ist es empfehlenswert eine Modifikation der Form

$$(4) \qquad P_x' = \begin{cases} P_x' & x = 1,2,5,6,... \\ P_3 + \beta P_4, & x = 3 \\ P_4(1-\beta), & x = 4 \end{cases}$$

durchzuführen, die mit $\beta \sim 0.21$ eine recht gute Anpassung liefert, auch wenn der Schweif der Verteilung in Text 11 etwas unregelmäßig ist und zusammengefaßt werden muß. Die Resultate sind in Tabelle 4 zu sehen.

Die restlichen drei Texte konnten mit diesen Modellen nicht erfaßt werden. Zum Zweck der weiteren Forschung bringen wir die Daten in Tabelle 5.

Tabelle 3

x	Text 3		Text 4		Text 6	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	85	85.00	106	106.00	97	97.00
2	187	190.41	357	344.98	208	213.55
3	246	235.43	286	288.51	322	305.10
4	120	125.26	125	138.73	157	172.67
5	30	34.65	43	43.43	44	47.65
6	6	5.62	15	9.57	14	7.31
7	3	0.62	1	1.79	1	0.68
8					1	0.04
	$a = 30.95$; $b = 2.93$ $X_2^2 = 2.60$; $P = 0.27$		$a = 6.92$; $b = 1.92$ $X_3^2 = 5.23$; $P = 0.16$		$a = 49.06$; $b = 3.22$ $X_1^2 = 2.84$; $P = 0.09$	

Tabelle 3 (Fortsetzung)

x	Text 9		Text 17	
	f_x	NP_x	f_x	NP_x
1	166	166.00	95	95.00
2	388	396.03	452	471.81
3	445	422.58	323	324.61
4	182	196.14	185	185.31
5	44	47.73	95	91.53
6	8	6.85 |	64	40.16 |
7	3	0.67 |	13	15.95 |
8			6	8.63 |
	a = 25.63; b = 2.89; X_2^2 = 4.27; P = 0.12		a = 1.40; b = 0.65; X_2^2 = 6.12; P = 0.05	

Tabelle 4

x	Text 11		Text 12		Text 16		Text 23	
	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	27	24.60	141	146.12	96	102.04	49	55.80
2	129	118.33	512	508.81	304	304.23	152	176.70
3	199	211.19	657	659.88	438	421.60	322	295.09
4	96	104.78	269	250.99	176	180.72	171	151.36
5	41	49.76 |	78	93.30	83	90.94	93	100.31
6	22	11.43 |	20	16.74	24	23.51	29	35.40
7	3	1.71 |	1	2.16	5	4.31 |	9	9.18
8	4	0.18 |			2	0.65 |	1	2.15
9	1	0.01 |						
	a = 32.77 b = 2.77 β = 0.20 X_1^2 = 3.62 P = 0.06		a = 22.61 b = 2.70 β = 0.21 X_3^2 = 5.27 P = 0.15		a = 13.95 b = 2.23 β = 0.23 X_3^2 = 2.66 P = 0.45		a = 12.64 b = 2.00 β = 0.23 X_4^2 = 11.60 P = 0.02	

Tabelle 5

x	Text 8	Text 22	Text 24
	f_x	f_x	f_x
1	79	47	85
2	241	210	482
3	368	268	618
4	109	156	305
5	38	99	141
6	10	37	63
7	1	23	11
8	1	7	3
9		1	

Die Resultate sollen nur ein erstes Bild über das Koreanische vermitteln und die Richtung für die weitere Forschung weisen, denn für alle Modelle ist die Zahl der Bestätigungen recht gering. Es ist selbstverständlich auch möglich, daß alle diese Texte mit einem einzigen Modell erfaßt werden können, auch wenn die Wahrscheinlichkeit dafür recht gering ist, da wir beinahe 200 unterschiedliche Verteilungen getestet haben.

Literatur

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1,* 98-106.

Glottometrika 15, 1996, 214-217

Bibliography of Quantitative Linguistics. Bibliographie zur quantitativen Linguistik. Bibliografija po kvantitativnoj lingvistike.

by Reinhard Köhler, Trier. With the assistance of Christiane Hoffmann
(Amsterdam Studies in the Theory and History of Linguistic Science, Series V, Library & Information Sources in Linguistics, Vol. 25.)
Amsterdam/Philadelphia: John Benjamins 1995. LII+780 pp.

Reviewed by L. Hřebíček, Prague

Each scientific branch rests on several pillars enabling it to continue in its further growth. One of these pillars is bibliography. As concerns quantitative linguistics, bibliographical volumes, one published by P. Guiraud (Utrecht 1954), another one by W. Girke, H. Jachnow, & J. Schrenk (München 1974), the handbooks of new editions by L. Doležel and his group (Prague 1964, 1965), by G. Billmeier, & D. Krallmann (Bonn University), and several others were more or less known but not always accessible to researchers in this field. The Current Bibliography appearing in the individual volumes of *Glottometrika,* formerly compiled by Werner Lehfeldt, and recently by Christiane Hoffmann, should also be mentioned. But now we have a bulky volume at hand and we hope that it will be attainable at least in each university library in all countries of the world.

This volume contains records taken from all the above-mentioned bibliographies and from several other ones, mainly from those published in the former Soviet Union. According to his own words, originally the author's aim was to cover as far as possible the entire production of quantitative studies without time and space limitations. This could be reached only partially, says the author. However, the extent reached is admirable, the thesaurus contains a total of 6341 bibliographical entries concerning works published up to the beginning of 1993.

It is completely understandable that the author together with his collaborators could not read all these works and classify them according to their content. The thematic classification of the recorded items into 28 chapters seems

to be done with reference to some external properties of the individual works, for example, to the headings or to the titles of sections. In such a large collection it was not possible to obtain and read all the registered works. The general chapters into which this bibliography is divided, such as Language Theory, Methodology and Information Theory, are followed by that one called Graphematics, then by chapters concerning individual grammatical subdisciplines and also Semantics; the classical linguistic branches for the quantitative approach are Stylistics and Metrics; then comes Textology (not Text Linguistics!), Sociolinguistics, and a chapter called Variation which includes different studies concerning the variation of phonemes, patterns of speaking, diversity of dialects of languages etc.; Psycholinguistics, Language Acquisition, Diachrony, Computational Linguistics, Linguistic Geography, Typology and Genetic Relationship are the other chapters of the volume. Among them Musicology is not missing, and this fact reminds a researcher mostly adapted to analyses of written texts that their sound form is worthy of investigation.

This basic division into 28 chapters is supplemented by a deeper classification; the so-called descriptors and indices characterize the content of individual works. More than one fourth of the volume is dedicated to the lists of indices, viz. of authors, key words of the works, subject headings, subheadings, uncontrolled vocabulary, further to the list of investigated languages and the reviewed publications. The entire apparatus enables the reader to find bibliographical references to a given issue in a large set of items; thanks to the indices a researcher can refer to works which at first sight may not seem to be relevant. For this reviewer it is hardly imaginable how this stupendous amount of work could be managed. (Let us mention only one minor recommendation for the future editions of this bibliography: in the list of key words, especially those taken from the Slavonic languages, the words should be presented in their canonical forms to exclude the rows of word forms like anglijskij, anglijskogo, anglijskoj, anglijskom, anglijskomu, p. 664, which could lead the reader to omission; the same principle applied to the Semitic languages, when they once will be drawn to quantitative investigation, is unthinkable!

As was stressed above, such a bibliographical compendium represents a great help in a firmer constitution of the science of language. It makes sense to raise the question here what kind of linguistic knowledge is offered by quantitative linguistics. The term is rather misleading. The representatives of classical linguistics often support the opinion that quantitative linguistics presents percentages and other quantitative expressions because quantity is an important and in a way philosophically justified concept. This can hardly be accepted as a true position. We can suppose that quantity is not a *sine qua non* of quantitative linguistics. This discipline turns to quantitative expressions with the same purpose

as is the case of an arbitrary science: this approach aims at the formulation of testable theories.

While the other branches of linguistics discuss different possibilities of the description of language phenomena on diverse levels of abstraction - and this is a legitimate aim for the application of linguistic knowledge -, besides this description, there remains a vacancy to be filled in by the knowledge based on *explanatory, empirical and testable* theories. And testable theories are constituents of that branch of the discipline which can be called a real (i.e., scientific, non-discursive) language theory. For this reason we understand the term of "quantitative linguistics" as a synonym of "theoretical linguistics".

In fact, a lot of quantitative works have a purely descriptive character; they can easily be found in Köhler's Bibliography. One cannot say that they are less interesting or less necessary for the progress of knowledge in the field of linguistics. We only want to stress that there is one fundamental classification of the works in quantitative linguistics, namely, the classification differentiating between descriptive and explanatory approaches (for example, between frequency dictionaries and the works by Gustav Herdan). According to our opinion, the described position was for the first time clearly and intentionally formulated in the German school of quantitative linguistics developing its studies mainly at the universities of Bochum and Trier. The latter one is the centre in which the work under review originated.

Besides bibliography, there are other pillars forming the basis of the future development of a scientific discipline, and we can ask which other kinds of editions are necessary for a firm foundation of quantitative linguistics. Köhler's bibliographical compendium clearly testifies that the great amount of theoretical knowledge is distributed among too many works. This situation requires systematization of this knowledge. We feel that this discipline requires a consistent presentation in textbooks. The requisite books should differ substantially from those which appeared decades ago and which are nothing but statistical manuals supplemented by examples taken from languages. They formulate not linguistic, but purely statistical hypotheses, which means that they are also of a descriptive character.

The manuals we have in mind should gather the testable linguistic hypotheses, which were tested in different languages; the future authors of these manuals should seek consequences of the coordination of different linguistic laws (= for the time being not rejected testable hypotheses) and their mutual compatibility. They should discuss and interpret all conceivable consequences of these laws which may become an inspiration for the construction of new theories.

Perhaps the way leading to such textbooks of quantitative linguistics goes through another type of publications; they can be something under the title "A

Reader in Quantitative Linguistics", a collection of reprinted important articles of chapters of books containing ideas which can become the foundation of the sought systematization. Such types of publications can represent a beautiful enlargement of those series the beginning of which is represented by Köhler's Bibliography. The grandiose work done by Reinhard Köhler and his collaborators is a decisive achievement aiming at such a systematization.

Glottometrika 15, 1996, 218-220

Methods in Quantitative Linguistics

by Juhan Tuldava, Tartu. Preface by Gabriel Altmann
(Quantitative Linguistics, Vol. 54)
Wissenschaftlicher Verlag Trier, 1995.

Reviewed by L. Uhlířová, Prague

Juhan Tuldava is a professor emeritus at the University of Tartu (Estonia), the founder of the Tartu school of quantitative linguistics and for many years the head of the text-analytical group of quantitative linguists. The volume under review offers selected papers, ten in number, and gives a good insight into his extensive scientific life-work. The papers were originally published in Russian or Estonian in various rare periodicals and collective volumes (issued by Tartu, Kiev and Vladivostok publishing houses) during the seventies and eighties. Now they have been translated into English, and supplied with recent bibliographical references; brief commentaries and explanations have been inserted at some places of the texts, where the author considered it necessary to make a point more up-to-date. As a separate volume of the well-established and prestigious series "Quantitative Linguistics" (as its 54th (!) volume) Tuldava's papers have become, at last, available to the wide community of readers who are used to communicate in academic English. Hopefully, this is the best way how to make Tuldava's works as popular as they deserve. The initiator, organizer, even sponsor - the true spiritus agens of Tuldava's volume, without whose untiring effort the book would hardly have come into the world, is Gabriel Altmann. Also, he is the author of a nice preface to the volume, which is both atribute to Tuldava (his work is presented by Altmann as "the bestpart of Estonian linguistics", p.VI), and a brief, sophisticated reflection on the international nature of science. It is Altmann to whom thanks are due for "discovering" Tuldava and for making his work easily accessible to anybody who may be interested. Naturally, no less merit is due to both editors, Reinhard Köhler and Burghard Rieger.

The volume is a success. The reader is most impressed by a perfect well-balance which shows in all relevant aspects, as required and expected from a

really mature piece of work by an experienced scientist. Without going into detail, let us point out what we consider to be the dominant aspects of Tuldava's work.

On the one hand, the author presents results of quantitative investigations made on CONCRETE linguistic data. On the other hand, he highlights GENERAL problems of quantitative linguistics and its methodological principles. He is very convincing in demonstrating that quantitative linguistics does not mean just a huge heap of EMPIRICAL data, usually of the frequencies of occurrences, counted for some applied purpose, but that it is a very promising field of THEORETICAL, multidisciplinary studies which is full of new perspectives. The well-balance between empiricism and theoretism, crucial for quantitative linguistics, is quite well manifested by Tuldava's multi-level approach: The lowest-level techniques and procedures are fruitful only if they do not merely describe simple facts and correlations, but if they serve "to seek for higher and higher generalizations, calling for some more and more abstract and complex concepts, setting up hypotheses and theories" (p.1). Therefore, any investigation should be performed in cycles which make possible to proceed from the measurement of random events (on lower cyclic levels) to regularities in the mass (on higher cyclic levels). The final aim of quantitative studies, as Tuldava puts it, is "building an adequate linguistic theory within the framework of the theory of self-organizing systems" (p. 2). With this opinion, Tuldava gets very near to the fundamental axioms of synergetic (systems-theoretical) linguistics, to the principles which have been already proved to be crucial for the further development of quantitative linguistics.

Before applying any quantitative technique, already elaborated in another science or discipline, to a corpus of linguistic data, Tuldava gives a brief, but full and instructive outline of the technique. He is extremely careful to his readers: He does everything to make sure that his readers can follow him. He must be an excellent teacher. His book is not only a monograph in which the author's INDIVIDUAL share in the field under study is presented (and, Tuldava IS, beyond any doubt, a very strong personality). Moreover, his book may serve as a HANDBOOK of quantitative linguistics, as a set of instructions for the use of quantitative techniques, as a valuable source of ideas and inspiration. That is why the term "methods" in the title of the volume suits so well. E. g., the reader learns about the prime notion of quantitative linguistics, namely measurement or "a procedure of ascribing numerical values to the observed objects..."(p. 3), about statistical distributions and functions, about correlation and contingency analysis, regression, factor analysis, parametric and non-parametric tests and about other effective tools of data analysis.

The solution of some linguistic problems often has a long history. Tuldava tries to push the already existing solution(s) a significant step forward, to add something new, to propose a SOLUTION OF HIS OWN, and is very successful in

doing so. At the same time he does not forget to acquaint the reader with the HISTORICAL BACKGROUND of the problem, with the current state of the art, and to remember the most important studies written so far on the issue (his knowledge of literature is wide and profound). E.g., this is true of the history of the well-known Zipf's law, once a simple, but empirically not very well fitting formula describing the relationship between frequency and rank of words which was repeatedly submitted to various modifications and "improvements".Tuldava deals with Zipf's law not only in connection with words and word forms; he also asks whether the phonemic level with its limited inventory (unlike the lexical level with a very large number of units) is also subject to Zipf's law and proposes a solution (pp. 161-184).

Tuldava does not present only already EXISTING techniques. He proposes some NEW ways how to compare vocabularies of two texts, how to calculate vocabulary size as a function of text length, etc. His proposals are motivated both by his delicate LINGUISTIC feeling and by his talent to grasp the exactness of MATHEMATICAL reasoning. Both aspects are well-balanced in his articles, and demonstrate that the author is well-experienced at both disciplines. Thus, the volume "Methods in Quantitative Linguistics" perfectly fits into the series in which it has been published. To read it will be rewarding both for those who wish to become acquainted with the work of a distinguished expert in the field and for those who have decided to be introduced into fundamentals of quantitative linguistics in an easy and skilful way.

Glottometrika 15, 1996, 221-232

CURRENT BIBLIOGRAPHY[1]

Compiled by Dieter Aichele

Sigla

BRJL	Bulletin ruského jazyka a literatury, Praha
CH	Computers and humanities
GLOTT	Glottometrika
JQL	Journal of quantitative linguistics
LLC	Literary and linguistic computing
VT	Virittäjä
ZBAL	Zeitschrift für Balkanologie
ZPHSK	Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung

ECKERT 1991 = Eckert, Penelope (Ed.): New ways of analyzing sound change. San Diego (u.a.): Academic Press, 1991 (Quantitative analyses of linguistic structure ; 5) ISBN: 0-12-229790-3

HŘEBÍČEK 1993 = Hřebíček, Luděk (Ed.); Altmann, Gabriel (Ed.): Quantitative text analysis. Trier: Wissenschaftlicher Verl. Trier, 1993 (Quantitative linguistics; 52) ISBN: 3-88476-080-7

KLENK 1992 = Klenk, Ursula (Ed.): Computatio linguae. Aufsätze zur algorithmischen und quantitativen Analyse der Sprache. Stuttgart: Steiner, 1992 (Zeitschrift für Dialektologie und Linguistik : Beiheft ; 73) ISBN: 3-515-06049-9

KLENK 1994 = Klenk, Ursula (Ed.): Computatio Linguae II. Stuttgart: Steiner, 1994 (Zeitschrift für Dialektologie und Linguistik : Beiheft ; 83) ISBN: 3-515-06420-6

KOIVUSALO 1988 = Koivusalo, Esko (Ed.): Mikael Agricolan kieli. Helsinki: SKS, 1988. (Tietolipas ; 112).

[1]Extract from the *Bibliography of Quantitative Linguistics* (BQL) which is being compiled at the University of Trier.

GENERAL

Altmann, Gabriel (Ed.): Glottometrika 14. Trier: Wissenschaftlicher Verl., 1993 (Quantitative linguistics ; 53) ISBN: 3-88476-081-5

Czyżakowski, W.; Piotrovskij, Raimund G.: Über den gegenwärtigen Stand der automatischen Textverarbeitung in der Forschungsgruppe "Sprachstatistik". (Zum Problem des linguistischen Automaten). IN: GLOTT 14(1993), S.161-189

Faulk, Ramon; Goertzel Gustavson, Frances: Toward a predictive theory of natural language. IN: KLENK 1992, S.16-31

Hřebíček, Luděk: Quantitative linguistics, by Marie Těšitelová. Praha, Academia 1992. 253 pp. IN: GLOTT 14(1993), S.213 [Review]

Hřebíček, Luděk: Fractals in language. IN: JQL 1(1994), S.82-86

Hurford, James R.: The study of language systems. IN: JQL 1(1994), S.43-55

Klein, Harald: INTEXT. A program system for the analysis of texts. IN: HŘEBÍČEK 1993, S.297-307

Kristophson, Jürgen: Einsatz numerischer Verfahren für Textüberlieferungsprobleme. IN: Goebl, Hans (Ed.); Schader, Martin (Ed.): Datenanalyse, Klassifikation und Informationsverarbeitung. Heidelberg: Physica-Verl. Rudolf Liebing, 1992. S.47-53

Kristophson, Jürgen: Ein neuer Beitrag zur Sprachbunddiskussion. IN: ZBAL 29(1993), S.1-11

Muller, Charles: Langue française, débats et bilans. Recueil d'articles, 1986-1993. Paris: Champion, 1993 (Travaux de linguistique quantitative ; 51) ISBN: 2-85203-299-6

DIALECTOLOGY

Adel, Kurt; Dutter, Rudolf; Filzmoser, Heidrun; Filzmoser, Peter: Tiefenstrukturen der Sprache. Untersuchung regionaler Unterschiede mit statistischen Methoden. Wien: Eigenverlag, 1994

Eckert, Penelope: Social polarization and the choice of linguistic variants. IN: ECKERT 1991, S.213-232

Goebl, Hans: Dendrogramme im Dienst der Dialektometrie. Zwei hierarchisch-agglomerative Klassifikationen von Daten des Sprachatlasses AIS. IN: KLENK 1992, S.54-73

Goebl, Hans: Spannungsverhältnisse in dialektalen Netzen. Ein Hinweis zu disziplinübergreifender Diskussion. IN: KLENK 1994, S.63-83

Kemp, William; Yaeger-Dror, Malcah: Changing realizations of a in "(a)tion" in relation to the front a - back a opposition in Quebec French. IN: ECKERT 1991, S.127-184

Knack, Rebecca: Ethnic boundaries in linguistic variation. IN: ECKERT 1991, S.251-272

Labov, William: The three dialects of English. IN: ECKERT 1991, S.1-44

Ogura, Mieko; Wang, William S.-Y.; Cavalli-Sforza, L. L.: The development of Middle English ī in England. A study in dynamic dialectology. IN: ECKERT 1991, S.63-106

Toon, Thomas E.: The sociopolitics of literacy. New methods in Old English dialectology. IN: ECKERT 1991, S.107-125

Tuomi, Tuomo: Suomen murteiden sanakirja. Johdanto. [Dictionary of Finnish dialects. Introduction]. Helsinki: VAPK, 1989 (Kotimaisten kielten tutkimuskeskuksen julkaisuja ; 36)

FINNISH AND FINNO-UGRIC RESEARCH

Häkkinen, Kaisa: Suomen perussanaston etymologiset kerrostumat. [The etymological strata of the Finnish lexicon]. IN: VT 96(1992), S.47-59

Ikola, Osmo; Palomäki, Ulla; Koitto, Anna-Kaisa: Suomen murteiden lauseoppia ja tekstikielioppia. [Syntax and text grammar in Finnish dialects]. Helsinki: 1989 (Suomalaisen Kirjallisuuden Seuran toimituksia; 511)

Jussila, Raimo: Agricolan sanasto ja nykysuomi. [Agricola's vocabulary and modern Finnish]. IN: KOIVUSALO 1988, S.203-227

Jussila, Raimo: Suomen yleiskielen ja murteiden sanastojen suhteista. [Finnish standard and dialectical lexicons compared]. IN: VT 93(1989), S.309-321

Jussila, Raimo: Kiihtelysvaaran murteen sanasto ja yleiskieli. [Vocabulary of Kiihtelysvaara dialect and standard language]. IN: Suni, H. (Ed.): Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi, 10.10.1990. Helsinki: VAPK, 1990. (Kotimaisten kielten tutkimuskeskuksen julkaisuja ; 60). S.40-60

Jussila, Raimo; Nikunen, Erja; Rautoja, Sirkka: Suomen murteiden taajuussanasto. A frequency dictionary of Finnish dialects. Helsinki: VAPK, 1992. (Kotimaisten kielten tutkimuskeskuksen julkaisuja ; 66)

Jussila, Raimo; Nikunen, Erja; Rautoja, Sirkka: Suomen murteiden taajuussanasto. Yksiesiintymäisten sanojen luettelo. A frequency dictionary of Finnish dialects. The listing of Hapax legomena. Helsinki: Kotimaisten kielten tutkimuskeskus, 1992

Jussila, Raimo; Kristiansson-Seppälä, Anna-Liisa: Bibliography of quantitative research into Finnish and other Finno-Ugric languages in Finland. IN: GLOTT 14(1993), S.197-212

Niemikorpi, Antero: Suomen kielen sanaston frekvenssianalyysia. [Frequency analysis of Finnish vocabulary]. Vaasa: Vaasan korkeakoulu, 1990. (Vaasan korkeakoulun julkaisuja : Tutkimuksia ; 150)

Niemikorpi, Antero: Suomen kielen sanaston dynamiikaa. [Dynamics of the Finnish vocabulary]. Vaasa: Vaasan yliopisto, 1991. (Acta Wasaensia / Universitas Wasaensis ; 26)

Pääkkönen, Matti: Grafeemit ja konteksti. tilastotietoja suomen yleiskielen kirjaimistosta. [Graphemes and context: statistical data on the graphology of standard Finnish]. Helsinki: SKS, 1990. (Suomi ; 150)

Pääkkönen, Matti: Graphemes and context: statistical data on the graphology of standard Finnish. IN: GLOTT 14(1993), S.1-53

Saukkonen, Pauli: Main trends and results of Quantitative Linguistics in Finland. IN: JQL 1(1994), S.2-15

GRAMMAR

Greidanus, Tine: Les constructions verbales en français parlé. Etude quantitative et descriptive de la syntaxe des 250 verbes les plus fréquents. Tübingen: Niemeyer, 1990. (Linguistische Arbeiten ; 243) ISBN: 3-483-30243-7

Kiuru, Silva: Agricola tulepi, ios henen tule. Ind. preesensin yks. 3. persoonan muodot Mikael Agricolan kielessä. [The 3rd person singular present indicative in Agricola's language]. IN: Kalliokoski, J. (Ed.); Leino, P. (Ed.); Pyhtilä, P. (Ed.): Kieli 3. Helsinki: Helsingin yliopiston suomen kielen laitos, 1988. S.7-76

Kiuru, Silva: Agricolan teonnimijohdosten erikoispiirteitä. [Action noun derivatives in -mus, -mys in Agricola's language]. IN: KOIVUSALO 1988, S.133-179

Savijärvi, Ilkka: Passiivin ja monikon 3. persoonan suhteesta vepsän kielessä. [On passive in Vepsian]. IN: Suni, H. (Ed.): Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10.1990. Helsinki: VAPK, 1990. (Kotimaisten kielten tutkimuskeskuksen julkaisuja ; 60). S.40-60

GRAPHEMICS

Koivusalo, Esko; Suni, Helena: Autuahus taiuahisa. Jälkitavujen vokaalien välinen "h" Agricolan teosten kielessä. ["h" between vowels after syllables with main stress in Agricola's language]. IN: KOIVUSALO 1988, S.111-132

Leskinen, Heikki: Tietoja sananalkuisten grafeemien ja grafeemikombinaatioiden yleisyydestä. [Daten zur Allgemeinheit der Grapheme und Graphemkombinationen im Wortanlaut des Finnischen]. IN: VT 93(1989), S.401-419

LEXICOLOGY AND LEXICOGRAPHY

Anreiter, Peter: Akzelerierter Wortschatzzerfall durch Superstratdruck. IN: KLENK 1994, S.1-18

Čitašvili, R. Ja.; Baayen, R. Harald: Word frequency distributions. IN: HŘEBÍČEK 1993, S.54-135

Delcourt, Christian: Where have all the key words gone?. IN: CH 23(1989), S.285-291

Gibbon, Dafydd: ILEX: A linguistic approach to computational lexica. IN: KLENK 1992, S.32-53

Jussila, Raimo: Leksikografinen luokittelu kvantitatiivisesta näkökulmasta. [Lexicographical classification from the quantitative point of view]. IN: XV kielitieteen päivät Oulussa 13. - 14.5.1988. Oulu: Oulun yliopisto, 1989. (Acta Universitatis Ouluenis : Series B, Humaniora; 14). S.85-91

Nahkola, Kari; Saanilahti, Marja: Lekseemin esiintymistaajuuden vaikutus kielenmuutoksen leksikaaliseen diffuusioon. [On the effect of lexeme frequency on the lexical diffusion of a language change]. IN: VT 94(1990), S.196-217

Saarela, Leena: Neljäsluokkalaisten tyttöjen ja poikien sanaston eroista. [On differences in vocabulary used by boys and girls in fourth class]. Oulu: Oulun yliopisto, 1991. (Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja ; 33)

Särkkä, Tauno: Sanaston rikkaudesta ja sen mittaamisesta. [Über den Reichtum des Wortschatzes und seine Messung]. IN: VT 91(1987), S.129-137

Sanada, Haruko: Comparison of effectiveness of various basic vocabularies. IN: GLOTT 14(1993), S.122-138

MACHINE TRANSLATION

Guy, Jacques B. M.: An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. IN: JQL 1(1994), S.35-42

METHODOLOGY

Baumann, Klaus-Dieter: The significance of the statistical method for an interdisciplinary analysis of professionalism of texts. IN: HŘEBÍČEK 1993, S.280-296

Köhler, Reinhard: Methoden und Modelle der quantitativen Linguistik. IN: Faulbaum, F. (Ed.): Softstat '91. Advances in statistical software. 3. Stuttgart (u.a.): Gustav Fischer, 1992. S.489-495

Kubáček, Lubomír: Confidence limits for proportions of linguistic entities. IN: JQL 1(1994), S.56-61

Marcus, Solomon: The logical and semiotic status of the canonic formula of myth. IN: HŘEBÍČEK 1993, S.159-174

Rietveld, Toni; Van Hout, Roeland: Statistical techniques for the study of language and language behaviour. Berlin (u.a.): deGruyter, 1993 ISBN: 3-11-013663-5

Roos, Undine: Measuring text difficulty in Japanese. Different tools - same results? IN: HŘEBÍČEK 1993, S.239-252

Sankoff, David; Rosseau, Pascale: A test for mixed rules. IN: ECKERT 1991, S.45-62

MORPHOLOGY

Best, Karl-Heinz; Zhu, Jinyang: Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa. (Mit einem Ausblick auf das Chinesische). IN: KLENK 1994, S.19-30

Flenner, Gudrun: Ein quantitatives Morphsegmentierungssystem für spanische Wortformen. IN: KLENK 1994, S.31-62

Jäppinen, Harri: Finite state computational morphology. IN: KLENK 1992, S.96-109

Janssen, Axel: Segmentierung französischer Wortformen in Morphe ohne Verwendung eines Lexikons. IN: KLENK 1992, S.74-95

Klenk, Ursula: Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen. IN: KLENK 1992, S.110-124

Klenk, Ursula: Automatische morphologische Analyse arabischer Verbformen. IN: KLENK 1994, S.84-101

Leskinen, Heikki: Vieläkö nuoret nurisevat? Huomioita onomatopeettisten sanojen tuntemuksesta ja tulkinnasta. [Beobachtungen zur Kenntnis und Deutung onomatopoetischer Wörter]. IN: VT 95(1991), S.355-371

Losiewicz, B. L.: The effect of frequency on linguistic morphology (affix stripping). Austin, Tex.: University of Texas, 1992. Thesis (Ph.D.)

Nikkilä, Osmo: Agricolan kieli ja teokset loppuheiton valossa. [Agricola's language and work in the light of apocope]. IN: KOIVUSALO 1988, S.94-110

Zhu, Jinyang; Best, Karl-Heinz: Zum Monosyllabismus im Chinesischen. IN: ZPHSK 45(1992)4, S.341-355

ONOMASTICS

Kiviniemi, Eero: Kajahtiko Karjalasta? karjalaisten etunimimieltymykset tilastojen valossa. [Statistical study on Karelian given names]. IN: Suni, H. (Ed.):

Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10.1990. Helsinki: VAPK, 1990. (Kotimaisten kielten tutkimuskeskuksen julkaisuja ; 60). S.78-93

PARSING

Davis, Stuart: Investigating English phonotactic constraints using a computerized lexicon. IN: KLENK 1992, S.1-15

Naumann, Sven: Adaptives Parsen. IN: KLENK 1994, S.127-147

PHONETICS

Möbius, Bernd: Ein quantitatives Modell der deutschen Intonation. Analyse und Synthese von Grundfrequenzverläufen. Tübingen: Niemeyer, 1993 (Linguistische Arbeiten ; 305) ISBN: 3-484-30305-0

PHONOLOGY

Altmann, Gabriel: Phoneme counts. Marginal remarks to Pääkkönen's article. IN: GLOTT 14(1993), S.54-68

Habick, Timothy: Burnouts versus rednecks. Effects of group membership on the phonemic system. IN: ECKERT 1991, S.185-212

Mock, Carol C.: Impact of the Ozark drawl. Its role in the shift of the diphthong /ey/. IN: ECKERT 1991, S.233-250

PRAGMATICS

Těšitelová, Marie: On quantitative analysis of dialogue and monologue. IN: HŘEBÍČEK 1993, S.271-279

Uhlířová, Ludmila: Parts of the sentence. Evidence of their communicative significance in text structure. IN: HŘEBÍČEK 1993, S.263-270

PSYCHOLINGUISTICS

Kukemelk, Hasso; Mikk, Jaan: The prognosticating effectivity of learning a text in physics. IN: GLOTT 14(1993), S.82-103

Kukemelk, Hasso: The dependence of the learning time on the text characteristics. IN: GLOTT 14(1993), S.104-112

SEMANTICS

Gordesch, Johannes; Zapf, Antje: Computer-aided formation of concepts. IN: HŘEBÍČEK 1993, S.29-45

Lepik, Madis: Mathematical verbal problems: differences in solving difficulties. IN: GLOTT 14(1993), S.113-121

Mikk, Jaan; Elts, Jaanus: Comparison of texts on familiar or unfamiliar subject matter. IN: HŘEBÍČEK 1993, S.228-238

Nahkola, Kari; Saanilahti, Marja: Koululaisslangin semanttisia ja sosiolingvistisiä piirteitä. [Semantic and sociolinguistic features of school slang]. IN: VT 95(1991), S.123-140

Wildgen, Wolfgang: The distribution of imaginistic information in oral narratives. A model and its application to thematic continuity. IN: HŘEBÍČEK 1993, S.175-199

SYNTAX

Delcourt, Christian: About the statistical analysis of co-occurrence. IN: CH 26(1992), S.21-29

Langer, Hagen; Naumann, Sven: Syntaktische Hierarchie und lineare Abfolge. IN: KLENK 1992, S.125-145

Langer, Hagen; Thümmel, Wolfgang: Syntaxen mit multiplen Hierarchien. IN: KLENK 1994, S.102-126

Palander, Marjatta: Puhe- ja kirjakielen sanajärjestyseroista. [Word order differences in spoken and written language]. IN: VT 95(1991), S.235-254

Pfeiffer, Oskar E.; Půček, Michael; Sgall, Petr: Die Thema-Rhema-Gliederung im Deutschen und ihre automatische Analyse. IN: KLENK 1994, S.148-164

Savijärvi, Ilkka: Agricolan kieltolause. [Negative sentence in Agricola's language]. IN: KOIVUSALO 1988, S.69-93

Thümmel, Wolfgang: Über Strukturen- und Kategorienvielfalt in kombinatorischen Kategorialsyntaxen. IN: KLENK 1992, S.146-168

Tottie, Gunnel: Negation in English speech and writing. A study in variation. San Diego (u.a.): Academic Press, 1991 (Quantitative analyses of linguistic structure ; 4) ISBN: 0-12-696130-1

TEXTOLOGY

Baayen, R. Harald: Derivational productivity and text typology. IN: JQL 1(1994), S.16-34

Basili, Roberto; Marziali, Alessandro; Pazienza, Maria Teresa: Modelling syntactic uncertainty in lexical acquisition from texts. IN: JQL 1(1994), S.62-81

Boroda, Mojsej Grigor'evič: Complexity oscillations in a coherent text: Towards the rhythmic foundations of text organization. IN: JQL 1(1994), S.87-97

Delcourt, Christian: Aspects algorithmiques de l'analyse structurale. IN: LLC 3(1988), S.232-236

Fenk, August: Text-picture transinformation. IN: HŘEBÍČEK 1993, S.151-158

Hřebíček, Luděk; Altmann, Gabriel: Prospects of text linguistics. IN: HŘEBÍČEK 1993, S.1-28

Hřebíček, Luděk: Text as a strategic process. IN: HŘEBÍČEK 1993, S.136-150

Köhler, Reinhard; Galle, Matthias: Dynamic aspects of text characteristics. IN: HŘEBÍČEK 1993, S.46-53

Liiv, Heino; Tuldava, Juhan: On classifying texts with the help of cluster analysis. IN: HŘEBÍČEK 1993, S.253-262

Møller, Erik: The influence of context on narrative structure. IN: HŘEBÍČEK 1993, S.200-214

Tuldava, Juhan: The statistical structure of a text and its readability. IN: HŘEBÍČEK 1993, S.215-227

Tuldava, Juhan: Measuring text difficulty. IN: GLOTT 14(1993), S.69-81

THEORY

Myhill, John: Typological discourse analysis. Quantitative approaches to the study of linguistic function. Oxford (u.a.): Blackwell, 1992 ISBN: 0-631-17614-4

Nemcová, Emília: On two realizations of Menzerath's law. IN: JQL 1(1994), S.107-112

Silnitsky, George: Typological indices and language classes: a quantitative study. IN: GLOTT 14(1993), S.139-160

Smetáček, V.: K otázce následnosti informací o úsecích děje v různých prozaických žánrech. [Zur Frage der Folge von Information über Handlungsabschnitte in verschiedenen prosaischen Genres]. IN: BRJL 22(1993), S.217-230

Suihkonen, Pirkko: Korpustutkimus kielitypologiassa sovellettuna udmurttiin. [Computer corpus analysis in language typology applied to Udmurt]. Helsinki: SUS, 1990 (Suomalais-ugrilaisen Seuran toimituksia; 207)

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel: Towards a theory of word length distribution. IN: JQL 1(1994), S.98-106

Zörnig, Peter; Altmann, Gabriel: A model for the distribution of syllable types. IN: GLOTT 14(1993), S.190-196