QUANTITATIVE LINGUISTICS
Vol. 36

WIEDERHOLUNGEN IN TEXTEN

von

G. Altmann

Studienverlag Dr. N. Brockmeyer
Bochum 1988

CIP-Titelaufnahme der Deutschen Bibliothek

Altmann, Gabriel:
Wiederholungen in Texten / von G. Altmann. – Bochum : Studienverl. Brockmeyer, 1988
(Quantitative linguistics ; Vol. 36)
ISBN 3-88339-663-X

NE: GT

ISBN 3-88339-663-X

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

V O R W O R T

Das vorliegende Buch ist kein Lehrbuch der quantitativen Textanalyse, da es nur einen einzigen Aspekt der Textphänomene erfaßt. Dieser Aspekt scheint aber recht wichtig zu sein, und man sollte ihn bei keiner Analyse außer acht lassen. Obwohl die meisten Resultate bestimmte statistische (beschreibende oder inferenzielle) Methoden sind, habe ich mich bemüht, zu zeigen, daß ein Text nur oberflächlich ein durch Regeln gesteuertes Gebilde ist. Im Hintergrund wirken bestimmte stochastische Gesetze, die dem Autor bei der Verwendung der Regeln volle Freiheit lassen, aber alles andere, wofür es keine Regeln gibt, steuern. Wiederholungen sind ein derartiges gesetzesgesteuertes Textphänomen, auch wenn vorläufig nur wenige Hypothesen darüber bekannt sind. Die Wirkung der Gesetze im Text postulieren wir nicht nur apriori – als Voraussetzung jeglicher Forschung –, sondern auch aufgrund der Tatsache, daß Texte innerhalb einer Sprachgemeinschaft entstehen, die mit ihrem Texten stets bestimmte Bedürfnisse befriedigt. Diese Bedürfnisse stehen entweder in Einklang miteinander oder in Konkurrenz. Sie sind der Boden, aus dem die Gesetze erwachsen sind, die nun die Gestaltung der Texte steuern. Ein Text ist ein synergetisches Gebilde bzw. das Resultat eines synergetischen Mechanismus im Menschen, eines Mechanismus, der eine Selbstregulation im Text hervorruft.

Diesen Mechanismus zu entziffern, bedeutet nicht nur neue Hypothesen aufzustellen, sondern gleichzeitig auch Messungen an zahlreichen Texten durchzuführen, die sowohl den Ansporn zur Hypothesenbildung als auch Überprüfungsinstanzen liefern.

Die Textlinguistik ist trotz zahlreicher Begriffsbildungen eine zwar nicht allzu junge, jedoch noch eine embryonale Wissenschaft, die sich vorläufig mit empirischen Generalisierungen begnügt. Die Zeit ist vielleicht noch nicht reif für die Verwirklichung höher gesteckter Forschungsziele. Dennoch möchte dieses Buch einige Impulse bringen und Wege zeigen, die ohne allzuviel Mathematik gangbar sind.

An dieser Stelle sei vor allem der ***Stiftung Volkswagenwerk*** gedankt, die mir im Rahmen des Projekts "Sprachliche Synergetik" Hilfsmittel zur Verfügung gestellt hat, weiter meinen Kollegen P.Grzybek, L.Hřebíček, W.Lehfeldt und H.Bluhme, die die erste Version dieses Buches kritisch gelesen und zahlreiche Verbesserungen vorgeschlagen haben. Das Recht auf Irrtum möchte ich jedoch für mich allein in Anspruch nehmen.

G.A.

I N H A L T

1. E I N F Ü H R U N G

1.1. Der Gegenstandsbereich

Unter einem ***TEXT*** werden wir eine beliebige sinnvolle - schriftliche oder mündliche - Äußerung in einer natürlichen Sprache verstehen. Einfachheitshalber werden wir nur geschriebene Texte untersuchen, was auch eine bessere Überprüfbarkeit unserer Resultate durch den Leser gewährleistet.

Unter einer ***TEXTEINHEIT*** werden wir eine beliebige Erscheinung verstehen, die man im Text operational definieren kann, d.h. eine Entität, die man im Text mit Hilfe von Kriterien eindeutig identifizieren bzw. deren Eigenschaften man messen kann. Die bekanntesten Texteinheiten sind

Buchstabe	*Wortform*	*Semem*
Graphem	*Phrase*	*Metapher*
Phonem	*Takt*	*Versfuß*
Silbe	*Clause*	*poetische Figuren*
Morphem	*Satz*	*grammatische Funktion*
Lemma	*Absatz*	*Referenz*

usw. Zahlreiche andere Texteinheiten findet man in textlinguistischen Arbeiten wie beispielsweise Koch (1969, 1971), Daneš, Viehweger (1977), Gottman, Parkhurst (1980), v. Dijk (1980), Dressler, de Beaugrand (1981).

Alle derartigen Einheiten besitzen zahlreiche Eigenschaften, auf deren Basis sie in verschiedene Klassen eingeordnet werden können. So ist beispielsweise das Wort

" ***Haus***" :

einsilbig
vierbuchstabig
monomorphematisch
ein Nomen
es hat n Bedeutungen (je nach Wörterbuch)
es kann in k übertragenen Bedeutungen benutzt werden
es bildet m Komposita
es gehört zu einem bestimmten Nominalparadigma
es hat in Texten die relative Häufigkeit p
es ist germanischen Ursprungs
es kann in r verschiedenen Kontexten benutzt werden
es hat s Morphe

usw. Als Texteinheiten kann man also nicht nur "materielle" Einzeleinheiten wie die oben aufgeführten betrachten, sondern auch ihre Eigenschaften, wie etwa die des Wortes "Haus", d.h. praktisch Klassen von Texteinheiten und deren Kombinationen.

Unter ***WIEDERHOLUNG*** verstehen wir das mehrmalige Vorkommen einer Texteinheit in einem Text.

Texteinheiten wiederholen sich in Texten aus folgenden möglichen Gründen:

(*1*) ***Inventarbeschränkung***. Hat das Inventar der Einheiten einen geringen Umfang, dann müssen diese sich öfter wiederholen. So wiederholen sich in einem deutschen Satz, der aus 50 Buchstaben besteht, mindestens einige Buchstaben mehrmals. In einem Text braucht sich aber kein Satz zu wiederholen, da das Inventar der Sätze unendlich ist. Die Wiederholungen von Einheiten aus einem nicht allzu kleinen, aber endlichen Inventar (z.B. Morpheme) können als Testinstanz der sprachlichen Kreativität dienen.

(*2*) ***Grammatik***. Die grammatischen Regeln schreiben gewisse Arten von Wiederholungen vor, z.B. "Das, was schön ist, ist auch gut". Synsemantika wie Präpositionen, Artikel, Konjunktionen wiederholen sich im allgemeinen öfter als Autosemantika, weil sie zahlreiche grammatische Funktionen erfüllen.

(*3*) ***Thematischer Zwang***. Spricht man über ein bestimmtes Thema, so benutzt man die Wörter des gegebenen Themenkreises öfter als sonst. Dieser Grund ist besonders in fachsprachlichen Texten maßgebend.

(*4*) ***Emphase, poetische, stilistische, ästhetische Gründe***. Nicht alles, was sich im Text wiederholt, hat eine dieser Funktionen, jedoch können Einheiten wiederholt werden, um dem Gesagten einen größeren Nachdruck zu verleihen, um bestimmte stilistische Effekte oder Rhythmen hervorzurufen, Euphonie zu erzeugen usw.

(*5*) ***Perseveration***. Wiederholungen von Texteinheiten können auch durch psychische Ursachen, Erkrankungen, Selbststimulation zustandekommen. Daher kann man sie bisweilen auch in der psychiatrischen Diagnostik verwenden (vgl. Mittenecker 1953, Breidt 1973).

(*6*) ***Informationsfluß***. Erzeugt man einen Text, so muß man darauf achten, daß der Empfänger ihn versteht. Das heißt unter anderem, daß

die Menge der übertragenen Information nicht zu groß sein darf. Aus unserer Sicht bedeutet das, daß man nicht nur ständig neue Wörter benutzt, sondern manche wiederholt verwendet, daß bestimmte Gedankenkomplexe mit anderen Wörtern wiederholt werden, daß bestimmte Begriffe ausführlich expliziert werden.

Es läßt sich nicht immer eindeutig feststellen, ob Wiederholungen bewußt oder unbewußt erfolgten, eine Antwort darauf kann bestenfalls nur der Textautor geben (vgl. Jakobson 1971). Im Vordergrund unserer Untersuchung stehen die Fragen, ob die gegebenen Wiederholungen als zufällig oder als nichtzufällig zu betrachten sind, wie man die Zufälligkeit testen kann, in welchen Spielarten Wiederholungen vorkommen, wie man für einzelne Spielarten Modelle aufstellen kann usw.

Die Untersuchung von Wiederholungen wird aus vier Gründen betrieben:

(*i*) ***Charakterisierung*** der Texte mit Hilfe von Kenngrößen (Maßzahlen), die man entweder aus der Statistik übernimmt oder nach Bedarf mit einer Begründung konstruiert.

(*ii*) ***Vergleich*** von Texten aufgrund ihrer Charakteristiken und darauffolgende Klassifikation der Texte; die diskriminative Eigenschaft von Kenngrößen wird auch bei Schlüssen über die strittige Autorschaft von Texten benutzt.

(*iii*) ***Erforschung von Gesetzen***, die die Konstruktion von Texten steuern und darauffolgender Aufbau einer Theorie der Texte. Dies ist sicherlich das Endziel jeglicher quantitativer Textanalyse, aber von diesem Ziel sind wir noch weit entfernt. Es gibt schon zahlreiche Modelle, die die einzelnen Aspekte der Texterzeugung erfassen, aber von einer ausgereiften Texttheorie zu sprechen, wäre verfrüht.

(*iv*) ***Diagnose psychischer Zustände***, auf die man aus den Wiederholungen schließen kann. Diese Forschung wurde vor allem in der Psychologie und der Psychiatrie betrieben.

1.2. Wiederholungsformen

Die bekannteste Wiederholungsform ist der Reim. Der Reim ist jedoch deterministisch: Er ist obligatorisch vorhanden oder nicht vorhanden; seine Eigenschaften können aber ein stochastisches Verhalten aufweisen. Wir werden uns im weiteren vor allem mit solchen Spielarten der Wiederho-

lung beschäftigen, die man nur als Tendenz, als Abweichung von vollständiger Zufälligkeit und vollständiger Determiniertheit erfassen kann. Im folgenden geben wir eine kurze Übersicht über die Wiederholungsformen, später werden wir die meisten von ihnen mit unterschiedlicher Ausführlichkeit behandeln.

Folgende Wiederholungen sind möglich:

Formlose (absolute) Wiederholung

Sie äußert sich im freien Vorkommen einer Einheit, d.h. in ihrer einfachen Häufigkeit im Text. Damit eine erhöhte Vorkommenstendenz erkannt wird, muß diese Häufigkeit von der theoretischen Häufigkeit (der erwarteten Häufigkeit) signifikant abweichen, oder es muß gezeigt werden, daß sie dem Text eine spezielle Eigenschaft verleiht. Jede formlose Wiederholung folgt jedoch einem Gesetz, dessen Parameter für die Einheit oder für den Text (oder für beide gleichzeitig) charakteristisch sind.

Positionale Wiederholung

Eine Einheit kommt in einer bestimmten Position, etwa am Versanfang, am Versende, am Wortanfang u.a. häufiger als erwartet vor. Solche Erscheinungen sind z.B. der Stabreim, die Präferenz für vokalisch endende Wörter im Reim, Alliteration, positionsbedingte Assonaz u.a.

Assoziative (konfigurative) Wiederholung

Eine Einheit kommt zusammen mit einer anderen in einem bestimmten Rahmen, z.B. in Sätzen, häufiger als erwartet vor. Dadurch lassen sich unbewußte Assoziationen des Textproduzenten oder im allgemeinen die Konnotationen der Wörter ermitteln.

Iterative Wiederholung

Eine Einheit bildet nichtzufällige Iterationen, d.h. ununterbrochene Sequenzen. Dies ist am häufigsten bei formalen Einheiten wie Versfußarten, metrischen Mustern, Wortlängen, je nach Textart eventuell auch bei Satztypen u.a. zu beobachten.

Aggregative Wiederholung

Eine Einheit konzentriert sich an einigen Stellen des Textes, d.h. es entstehen "Klumpungen" einer Einheit im Text. Dies ist eventuell die

Folge der Skinnerschen "formalen Verstärkung", eine Art der Selbststimulation. Sie wird an den häufigen kleinen und den seltenen großen Abständen der Einheit im Text erkannt.

Ähnlichkeitsaggregative Wiederholung

Nicht nur identische Einheiten können die Tendenz aufweisen, eng nebeneinander zu stehen, sondern auch ähnliche Einheiten. So kann man z.B. beobachten – besonders in der Volkspoesie –, daß aufeinanderfolgende Verse phonetisch ähnlicher sind als weit auseinanderstehende.

Blockmäßige Wiederholung

Ein Wort zeigt eine gesetzesartige Verteilung auf Textblöcke, d.h., folgt einer Wahrscheinlichkeitsverteilung, deren Parameter beispielsweise mit der Wortart, der Häufigkeit des Wortes, seiner Bedeutungshaltigkeit usw. korrelieren können.

Parallele Wiederholung

Diese Erscheinung ist besonders aus der Volkspoesie bekannt. An bestimmten parallelen Stellen des Gedichts erscheinen die gleichen oder ähnliche formale oder inhaltliche Einheiten. Die bekannteste Form ist der Reim. In der Volkspoesie findet man auch parallele Lautassoziationen oder inhaltliche Parallelismen.

Zyklische Wiederholung

Die Wiederholung bestimmter Einheiten kann in einem so hohen Maße regulär sein, daß sie ganze Zyklen bildet. Sie ist am intensivsten an der Plazierung des Akzents im Vers untersucht worden.

Diese Aufzählung ist sicherlich nicht vollständig. Man hat im Laufe der Entwicklung der Wissenschaft in allen Forschungsobjekten immer wieder neue Aspekte, Eigenschaften, Verhaltensformen entdeckt oder konstruiert, keine Beschreibung oder Theorie ist zur Zeit vollständig. Unsere Untersuchung erfaßt diejenigen Aspekte, die bisher zumindest konzeptuell entdeckt worden sind, und stellt eine Art Handbuch dar. Sie erhebt keineswegs den Anspruch, eine ausgereifte Theorie zu sein, denn in einer Theorie muß das Wissen über das Objekt zu einem System von Gesetzen zusammengefaßt werden. Zwar werden hier einige Gesetze dargestellt, aber vorläufig lassen sich Texteigenschaften nur mühsam auf deduktivem Wege ermitteln. Einige der vorgeschlagenen Methoden sollten

als Anreiz dienen, weitere Untersuchungen durchzuführen, um die ganze Palette des Verhaltens einer Eigenschaft zunächst induktiv zu erfassen. Die meisten Methoden hier dienen der Überprüfung der Existenz von allgemeinen oder textspezifischen Tendenzen.

In den folgenden Kapiteln werden wir uns nie mit Einzelfällen beschäftigen, wie etwa dem sporadischen Vorkommen irgendwelcher poetischer Figuren, denen sich die qualitative Stilistik mit Vorliebe widmet (vgl. z.B. Gonda 1959; Mason 1961; Austerlitz 1961 u.a.), sondern immer mit solchen Wiederholungsformen, die sich als Tendenz erkennen lassen, die irgendeinem Gesetz folgen, deren latente Existenz mit objektiven Methoden ermittelt werden kann.

1.3. Auf dem Weg zu einer zukünftigen Texttheorie

Mit Recht spricht man heute lieber von einer Textlinguistik als von einer Texttheorie. Die Entwicklung von ganzen Systemen von Begriffen, die einen Text lückenlos abdecken, ist sicherlich eine theoretische Betätigung. Man pflegt diese Begriffe zu definieren, zu operationalisieren, man stellt Kriterien zur Textsegmentierung auf, man beschreibt und klassifiziert Texte. Dies sind alles notwendige, aber keineswegs hinreichende Voraussetzungen für eine Theorie. Eine Theorie besteht aus

Begriffen,
Konventionen (Definitionen, Kriterien, Operationen u.a.),
Hypothesen.

Man pflegt zu diesen Bestandteilen auch die ceteris-paribus-Bedingung zu zählen, ohne die kein Gesetz gilt.

Mit den Begriffen setzt man nur fest, ***WAS*** es in Texten gibt, d.h., man baut eine Ontologie auf, die auf einer bestimmten Art, Texte zu sehen, begründet ist. Die Theorie fängt aber erst dann an, zum Leben zu erwachen, wenn man auch Fragen nach dem ***WIE*** und dem ***WARUM*** zu beantworten wünscht, d.h., wenn man Hypothesen aufstellt. Dies läßt sich an einem einfachen Beispiel leicht zeigen: Über die Arten der Referenzen im Text sind ganze Bücher geschrieben worden, man hat sie detailliert erfaßt, klassifiziert, und man hat auch gezeigt, wie man sie im Text findet, d.h., man hat die Kriterien der Identifizierung aufgestellt. Die *Voraussetzungen* für eine Theorie der Referenzen sind damit geschaffen worden, die Theorie selbst aber keineswegs. Der erste Schritt in dieser Richtung wurde von L.Hřebíček unternommen (vgl. § 2.5.3), der das Anwachsen der Referenzen in Abhängigkeit von der aktuellen Textlänge und von dem Wortschatz theoretisch begründete, empirisch teilweise überprüfte, es mit den Hypothesen über Vokabularreichtum verknüpfte und dadurch zum ersten Referenzengesetz gelangte.

Um eine Theorie zu errichten, genügt es keineswegs, Hypothesen zu haben, denn diese können einen sehr unterschiedlichen gnoseologischen Status aufweisen. Bunge (1969: 256 f.) unterscheidet vier Stufen der Hypothesenbildung, je nachdem, wie sie mit Theorie und Empirie verbunden sind:

(i) Ratereien, die weder durch die Empirie bestätigt noch aus einer Theorie abgeleitet werden können.

(ii) ***Empirische Hypothesen*** oder ***empirische Generalisierungen*** oder ***induktive Hypothesen***, die durch Verallgemeinerung von Beobachtungen entstehen, durch die Empirie recht gut bestätigt sind und als isolierte, nicht systematisierte Aussagen am Anfang jeder Forschung stehen. Von dieser Art sind z.B. alle sprachlichen Universalien und typologischen Hypothesen, Aussagen von allgemeinen Grammatiken sowie fast alle textlinguistischen Verallgemeinerungen.

(iii) ***Plausible Hypothesen*** oder ***deduktive Hypothesen***, die zwar aus einer Theorie abgeleitet, jedoch noch nicht überprüft worden sind. Sie sollten aber grundsätzlich testabel sein.

(iv) ***Bestätigte Hypothesen***, die sowohl theoretisch begründet als auch durch die Empirie gestützt sind, d.h. abgeleitete und überprüfte Hypothesen. Wenn sie allgemein genug und in ein System eingebettet sind, dann nennt man sie ***Gesetze***.

Gesetze sind gerade diejenigen Hypothesen, die man braucht, um von einer Theorie sprechen zu können. Der Weg zu ihnen ist mühsam, denn ohne Mathematik kann man sich ihre Ableitung kaum vorstellen, ohne Mess- oder Zählmethoden kann man Texte nicht auswerten, und ohne die Statistik kann man keine Aussagen über den Bestätigungsgrad einer Hypothese treffen. Dies bedeutet aber nicht, daß empirische Generalisierungen, wie sie das Gros der Textlinguistik und Linguistik bilden, wertlos seien. Im Gegenteil, sie zeigen die Richtung an, in der man theoretisch arbeiten soll, sie liefern den Ansporn zur Theoriebildung, geben empirische Unterstützung und sind zu jeder Zeit in jeder empirischen Wissenschaft vorhanden.

Der Weg zur Theorie besteht also aus folgenden Stufen:

(a) Bildung von ***Begriffen*** als die elementare Tätigkeit zur Entdeckung dessen, "was es in den Texten gibt".

(b) Aufstellung von ***Hypothesen*** in Form von empirischen Generalisierungen über Befunde in Texten, z.B. über Verläufe, Abhängigkeiten, Wiederholungsmechanismen u.ä.

(c) ***Deduktion*** unserer Annahmen über Texterscheinungen. Auf dieser Stufe fängt die eigentliche Theoriebildung an, die im Grunde deduktiv ist. Erst hier kann man Antworten auf die wie- und warum-Fragen erhalten, d.h. adäquate Beschreibungen des Textverhaltens und vor allem Erklärungen der Mechanismen, die die Texterzeugung steuern. Sicherlich wird diese Stufe später die Einschaltung der Psychologie erfordern, aber große Teile der Texttheorie lassen sich rein im linguistischen Rahmen aufbauen.

(d) ***Überprüfung*** der theoretischen Ableitungen anhand von konkreten Texten, wozu am besten die Statistik geeignet ist.

(e) ***Systematisierung*** der Hypothesen, d.h. ihre Einbettung in einen größeren Rahmen, ihre Verknüpfung mit anderen Hypothesen, Aufbau eines Systems von Gesetzen, die für alle Texte aller Sprachen gelten.

Es ist schwer vorstellbar, daß man ohne Bewältigung dieser Stufen zu einer Texttheorie gelangen kann, es sei denn, man bezeichnet etwas anderes als Theorie. Da bisher in allen Erfahrungswissenschaften dieser Weg gegangen worden ist - und die Wissenschaftstheorie spiegelt ihn wider -, ist nicht ersichtlich, warum es gerade in der Textlinguistik oder der Lingustik anders sein sollte.

Bereits vor 50 Jahren hat man angefangen, die Sprache als System zu betrachten. Zu dieser Zeit war es eher eine Bezeichnung ohne Konsequenzen als eine methodologische Annahme, die man für Deduktion, Modellieren u.ä. benutzen konnte. Unter der Hegemonie der generativen Linguistik der letzten Jahrzehnte hat man zuletzt völlig darauf verzichtet, in der Linguistik in systemtheoretischen Begriffen zu denken. Außerhalb der Linguistik hat sich aber die Systemtheorie zu der mächtigsten Methodologie dieses Jahrhunderts entwickelt und alle empirischen Wissenschaften erfaßt. Auch die Linguistik blieb nicht ganz "verschont" von ihr. Nach Zipf (1949) haben sich zahlreiche Linguisten immer wieder darum bemüht, die Begrifflichkeit der Systemtheorie einzuführen (vgl. Koch 1974; Nöth 1974, 1975, 1977, 1978, 1983; Oomen 1971; Schweizer 1979; Wildgen 1985 u.a.; Köhler, Altmann 1983), in letzter Zeit wurden auch mathematische Modelle mit ausgesprochen systemtheoretischer Orientierung (vgl. Strauß 1980; Köhler 1986) verwendet. Der Status des Textes ist aber trotzdem noch unklar geblieben.

Was ist eigentlich ein Text? Wenn wir annehmen, daß Sprache ein System ist, dann können wir sofort nach Subsystemen suchen. Sprache hat zwei Arten von Subsystemen, erstens die der Einheiten auf allen Ebenen des Kodes und zweitens die der Idiolektträger. Der Sprecher/Hörer

ist also ein Subsystem der Sprache, genauso, wie die Leber ein Subsystem des Organismus ist. Alle Subsysteme produzieren irgendeinen output, der der Aufrechterhaltung und der Entwicklung des Systems dient. Texte sind output des Sprecher-Subsystems und gleichzeitg input für das Hörer-Subsystem. Sie erzeugen Spannung im Kode, die die Sprache durch Selbstregulation beseitigt. In den Sprecherprodukten (= Texten) kommen die Kreativität und der Autonomiedrang des Idiolektträgers zum Ausdruck, in der Rezeption des Textes durch den Hörer kommen die integrative Tendenz der Sprache, die Versklavung, die Dominanz des Systems, die Normierung oder wie man es auch immer nennt, zum Tragen.

Im Text wirken also zwei Arten von Systemkräften, die man in der Linguistik als "Zipfsche Kräfte" bezeichnet; sie kooperieren und konkurrieren, d.h. sie steuern die Gestaltung des Textes. Aus dem Chaos des unartikulierten Gedankenmaterials werden durch die kombinatorischen Einschränkungen der Grammatik Sätze gebildet. Grammatik (Morphologie, Syntax) ist die Lehre von kombinatorischen Freiheitsgraden der Subsysteme des Kodes (Morpheme, Wörter). Sie stellt also nur Einschränkungen (constraints) dar, die die Kreativität der Idiolektträger im Satzrahmen systemkonform halten. Oberhalb des Satzes gibt es nur semantische Einschränkungen und Referenzen, die mit der Grammatik weniger zu tun haben, jedoch oft als "Beziehungen zwischen Sätzen" dargestellt werden. Sowohl diese Beziehungen als auch der ganze "Rest" der Textbildung sind aber weder chaotisch noch deterministisch; oberhalb des Satzes bekommt der Text lediglich umfangreiche Freiheitsgrade, die durch Textgesetze so weit eingeschränkt werden, daß er als input akzeptiert wird. Der Text kann als input nur dann akzeptiert werden, wenn die ihm innewohnenden Strukturen als mit denen des Hörers identisch betrachtet werden können, d.h., wenn sie von denen des Hörers nicht signifikant abweichen: der Text muß also strukturiert sein. Da aber Strukturierung eine grundlegende Systemeigenschaft ist, sind Texte auch Systeme. Sie haben alle Eigenschaften, die ein System auszeichnen, und ihre Untersuchung unter diesem Blickwinkel würde zu einer Texttheorie führen.

Obwohl man heutzutage noch von keiner ausgereiften Systemtheorie der Texte sprechen kann (gute Ansätze findet man bei Nöth 1977, 1983), versuchen wir in fast allen Kapiteln die Möglichkeit zu zeigen, wie man Gesetzeskandidaten ableiten könnte. Wir sind der Überzeugung, daß oberhalb der einschränkenden, zum Teil durch Konventionen festgelegten grammatischen Ebene *alles* in Texten durch (stochastische) Gesetze gesteuert wird, und die Aufgabe einer künftigen Textanalyse wird darin liegen, nicht nur Begriffe anzuhäufen, sondern gleichzeitig auch zu versuchen, Hypothesen über das Verhalten der durch diese Begriffe identifi-

zierten Eigenschaften aufzustellen. Texte sind ein Forschungsgebiet, dessen Breite im Augenblick nicht einmal abgeschätzt werden kann, und die Wiederholungen stellen einen kleinen, aber nicht unwichtigen Teilbereich dar, an dem man die komplizierte Dynamik der Texterzeugung bereits ermessen kann.

2. FORMLOSE WIEDERHOLUNG

Wiederholt sich eine Einheit formlos, d.h. ohne jegliche Bedingungen, dann handelt es sich einfach um die Häufigkeit ihres Vorkommens. Soll dem Text durch ihre häufige Wiederholung eine bestimmte Eigenart verliehen werden, dann muß es möglich sein, zu zeigen, daß diese Häufigkeit an sich oder im Vergleich mit Häufigkeiten anderer Einheiten eine besondere, statistisch nachweisbare Qualität hat.

Wir wollen hier vier Fälle unterscheiden:

(1) Die Häufigkeit einer Einheit im Vergleich mit ihrer erwarteten Häufigkeit.

(2) Die Häufigkeit einer Einheit, bezogen auf die Häufigkeit einer anderen. Dieser Fall läßt sich leicht (auf mehrere andere Einheiten) verallgemeinern, so daß sich hier zahlreiche andere Untersuchungsmöglichkeiten ergeben.

(3) Die Häufigkeit aller Einheiten eines Typs, die den gesamten Text abdecken.

(4) Die Konstruktion von Textgesetzen, denen die empirischen Regularitäten von Texteinheiten folgen.

Bei allen Untersuchungen in diesem Kapitel und in den folgenden Kapiteln geht es immer darum, die Wahrscheinlichkeit der gegebenen Erscheinung zu ermitteln. Man kann oft direkt auf eine bekannte Wahrscheinlichkeitsverteilung zurückgreifen, häufig aber kann man die gesuchte Wahrscheinlichkeit nur "indirekt" finden, indem bestimmte Transformationen durchgeführt werden. Alle Methoden werden an konkreten "textlinguistischen" Problemen demonstriert.

2.1. Phonische Effekte

Betrachten wir das Vorkommen einer isolierten Einheit in einem Text(teil), einfachheitshalber einen Laut in einem Vers, und stellen uns die Frage, ob seine Häufigkeit (ohne Rücksicht auf die Position) einen (eu)phonischen Effekt erzeugen kann. Sollte das der Fall sein, dann muß diese Häufigkeit etwas Unerwartetes darstellen. Um dies festzustellen, kann man drei Wege gehen:

(i) Man fragt den Textautor, ob er eine "euphonische" Absicht hatte – was allerdings etwa bei Homer mit Schwierigkeiten verbunden wäre. Jedoch auch wenn der Autor noch lebt, kann er etwas spontan getan haben, ohne sich dessen bewußt gewesen zu sein.

(ii) Man befragt zahlreiche Textrezipienten, die die Textsprache als Muttersprache sprechen, nach ihrem subjektiven Urteil. Dies ist gleichfalls oft unmöglich und, falls möglich, dann mit großem Zeitaufwand verbunden. Das Resultat hängt auch davon ab, wen man befragt. Möglicherweise gibt es alters–, bildungs–, sozial– und andersbedingte Rezeptionsunterschiede, so daß zwei unabhängige Forscher zu völlig unterschiedlichen Resultaten kommen können.

(iii) Man objektiviert die Ermittlung und zieht die Statistik zur Hilfe heran. Man bekommt dann zwar keine direkte Antwort auf die Frage nach dem euphonischen Wert eines Lautes, sondern eine Zahl, eine Wahrscheinlichkeit, die man interpretieren muß. Ein derartiges Resultat hat mehrere Vorteile, auf die man bei subjektivem Vorgehen verzichten muß:

(a) Das Resultat ist durch andere Forscher nachvollziehbar, kontrollierbar, und man bekommt mit zulässiger Toleranz immer dasselbe Resultat.

(b) Zieht man aus der resultierenden Zahl Schlüsse, so ist das Risiko eines Fehlschlusses berechenbar.

(c) Aus dem Schluß und der anschließenden Interpretation im Lichte der aufgestellten Hypothese erfährt man nicht nur, daß beispielsweise ein Vers beim Rezipienten einen phonischen Effekt hervorgerufen hat, sondern auch, was diesen Effekt hervorgerufen hat und wie dieses Etwas gestaltet werden muß, um einen Effekt hervorzurufen.

Da ein poetischer Text keine Einheiten (z.B. Laute, Wörter oder ihre Kombinationen) enthält, die ein "nichtpoetischer" Text nicht enthalten könnte, können phonische Effekte nur dadurch entstehen, daß eine bestimmte Einheit an einer Stelle steht, an der sie bei zufälliger Verteilung dieser Einheiten nicht erwartet wird, oder dadurch, daß sie in einer Häufigkeit vorkommt, die "nicht zufällig" ist. An dieser Stelle interessiert uns nur die zweite Möglichkeit.

Die Nichtzufälligkeit können wir dadurch nachweisen, daß wir zeigen, daß die Wahrscheinlichkeit, die beobachtete oder eine noch größere Anzahl der fraglichen Einheit im gegebenen Teiltext zu finden, sehr klein ist. Es gibt jedoch keine objektive Prozedur, um festzustellen, was als "sehr klein" zu betrachten ist. Dies ist die einzige Stelle, wo eine subjektive Entscheidung getroffen werden muß, und dies geschieht durch Konvention oder nach Bedarf, je nach der Art des Problems. In der Linguistik muß man in diesem Fall genauso verfahren wie in den "härteren" Wissenschaften, wo derartige Entscheidungen lebenswichtig sein können.

Zerlegen wir unser Problem in Einzelschritte (vgl. Altmann 1973) und illustrieren dies gleich an einem Beispiel. Das Beispiel wählen wir aus einer "exotischen" Sprache, aus dem Indonesischen, um vom Inhalt ausgehende Einflüsse beim deutschen Leser auszuschalten. Die erste Strophe des Gedichts "Bunda dan anak" von Roestam Effendi lautet (in phonologischer Transkription) wie folgt:

Masaq jambaq
buah sĕbuah
dipĕram alam diujuŋ dahan
merah darah
bĕruris-uris
bĕndera masaq bagi sĕlera.

Unsere philologische Frage lautet:

Ruft /a/ in dem ersten Vers einen (eu)phonischen Effekt hervor?

Neutral, nicht "ganz statistisch" formuliert:

Ist das Vorkommen von /a/ im ersten Vers in der beobachteten Anzahl unerwartet?

Noch andere Formulierungen sind möglich.

Ein Statistiker würde sich unsere Frage folgendermaßen in seine Sprache übersetzen:

Wie groß ist die Wahrscheinlichkeit, daß am allen Stellen, wo ein Vokal vorkommt, ein /a/ zu finden ist?

Oder auch: Unterscheidet sich die beobachtete Häufigkeit von /a/ signifikant von der erwarteten ("theoretischen") Häufigkeit?

Um solche Fragen zu entscheiden, stellt der Statistiker üblicherweise Hypothesen auf. In unserem Fall würden die Hypothesen (nicht formal) folgendermaßen lauten:

Null-Hypothese: Die beobachtete Häufigkeit von /a/ im ersten Vers unterscheidet sich nicht von der erwarteten.
Alternative Hypothese: Die beobachtete Häufigkeit von /a/ im ersten Vers ist größer als die erwartete.

Nachdem die Hypothesen formuliert worden sind, stellt der Statistiker ein Modell auf, mit dem er das Verhalten der jeweiligen Entität simuliert und das erwünschte Resultat berechnet. In unserem Fall kann man folgendermaßen verfahren.

Die Wahrscheinlichkeit des Vorkommens von /a/ in einer Grundgesamtheit sei p. Dieses p kann unterschiedlich geschätzt werden (den "wahren" Wert kann man nicht ermitteln), je nachdem, welche Grundgesamtheit wir zugrundelegen. Es gibt mehrere Möglichkeiten: Wir nehmen

(a) das ganze gegebene Gedicht,
(b) alle Gedichte von Roestam Effendi,
(c) die gesamte indonesische Poesie,
(d) alle indonesischen Texte.

Was immer wir auch tun, so stellt die Frage der Grundgesamtheit in der Sprache ein sehr heikles Problem dar, wie Orlov (1982) gezeigt hat. Die Möglichkeit (a) fällt aus, denn wenn das ganze Gedicht "/a/-gefärbt" ist, führt dies zu keinem Resultat. Die Möglichkeiten (c) und (d) scheiden nach Orlov aus, so daß am plausibelsten noch (b) ist. Wenn aber ein Autor Lyrik und Epik geschrieben hat, so ist auch diese Grundgesamtheit nur illusorisch.

Da es uns hier nur um die Demonstration eines Verfahrens geht, nehmen wir die einzige "Grundgesamtheit", die uns zur Verfügung steht, nämlich eine Zählung von 23000 indonesischen Phonemen aus der Prosa. Hier ergibt sich die relative Häufigkeit von /a/ als $p_a^* = 0.2227$ (mit "*" werden wir immer Schätzungen bezeichnen).

Nehmen wir an, daß wir im Vers an Vokale vier Positionen zu vergeben haben. Sei die Wahrscheinlichkeit von /a/ gleich p und die Wahrscheinlichkeit eines beliebigen anderen Vokals /b/ gleich q (q = 1-p). Dann setzt sich das Ereignis, daß /a/ genau einmal im Vers vorkommt, aus folgenden Ereignissen zusammen

b a a a, a b a a, a a b a, a a a b

und die entsprechenden Wahrscheinlichkeiten berechnen sich als

p q q q, q p q q, q q p q, q q q p

oder zusammen

$$4pq^3.$$

Wir können also schreiben

$$P(X=1) = 4pq^3,$$

wobei P die Wahrscheinlichkeit des Ausgangs des Experiments bedeutet. Ebenso leicht kann man feststellen, daß es 6 Permutationen gibt, in denen 2 p und 2 q vorkommen, d.h.

$$P(X=2) = 6p^2q^2$$

usw. Allgemein können wir schreiben

$$P(X=x) = \binom{4}{x} p^x q^{4-x}$$

wobei $\binom{n}{x}$ die sogenannten Binomialkoeffizienten sind. Man kann auch schreiben

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

wo $k! = k(k-1)(k-2)\ldots3\cdot2\cdot1$ die Fakultät bezeichnet. Die allgemeine Formel lautet dann

$$P(X=x) = \binom{n}{x} p^x q^{n-x} \qquad (2.1.1)$$

und gibt die Wahrscheinlichkeit an, daß bei n Ereignissen das günstige Ereignis x-mal und das ungünstige Ereignis (n-x)-mal vorkommen. Hier haben wir stillschweigend angenommen, daß alle Vorkommen von Vokalen voneinander unabhängig sind, was in Wirklichkeit niemals zutrifft. Daher kann man das Modell nur als eine Approximation benutzen.

Fragen wir nun, wie groß die Wahrscheinlichkeit ist, daß an allen vier Vokalstellen des Verses ein /a/ steht, so ergibt sich

$$P(X=4) = \binom{4}{4} p^4 q^{4-4} = p^4 = 0.2227^4 = 0.0025.$$

Rechnungen dieser Art kann man nur für Einzeleinheiten durchführen. Man kann aber auch nach der euphonischen "Kraft" etwa des Morphems /buah/, der Silbe /rah/ usw. fragen.

Aus unserem Resultat kann man unschwer verschiedene Indizes der Euphonie bilden, was wir hier jedoch unterlassen werden.

Es wäre nicht verkehrt, sich auch die Frage zu stellen, ob es eine "eusemische" Wiederholung von Semen geben kann, z.B. bei Gradationen. Auch den Reim kann man als eine Erscheinung mit "eulexischer" Wirkung bezeichnen, weil sich hier morphologisch oder semantisch nicht segmentierbare Teile des Wortes wiederholen können.

Man kann sich auch fragen, welche Effekte überhaupt durch freie Wiederholung einer Einheit im Text hervorgerufen werden können.

2.2. Zwei Variablen

Im vorigen Abschnitt haben wir die formlose Wiederholung einer einzigen Einheit untersucht. Die hier angewandte Methode erlaubt es, zahlreiche Indizes zu konstruieren, was in der Textanalyse routinemäßig geschieht und bei einer Variablen keine besonderen Probleme hervorruft.

Zählt man jedoch mehr als eine Einheit gleichzeitig, so ergeben sich Probleme sowohl bei der Indexbildung als auch bei der Anwendung der Indizes.

2.2.1. Der Aktionsquotient

In diesem Abschnitt beschränken wir uns auf zwei Variablen und illustrieren die Problematik an dem wohlbekannten Busemannschen "Aktionsquotienten" (vgl. Busemann 1925; Boder 1940; Schlismann 1948; Antosch 1953; Fischer 1969; Altmann 1978). Man pflegt diesen Index als

$$Q = \frac{v}{a} \qquad (2.2.1)$$

zu definieren, wobei

v = Zahl der "aktiven" Verben im Text
a = Zahl der Adjektive im Text

bedeutet. (Man kann den Quotienten auch als $Q = a/v$ definieren, wie es einige Autoren tun.) Wenn man mehr Verben als Adjektive findet, dann soll der Text eher "aktiv" sein; wenn es sich umgekehrt verhält, dann ist er eher "deskriptiv".

Schauen wir uns zuerst die linguistische Interpretation dieses Indexes an. Das erste Problem ist, was man unter einem "aktiven" Verb versteht. Gehören "schlafen", "ruhen", "leiden" u.a. zu den Verben, die Aktivität ausdrücken? Zählen Adverbien, die die Qualität von Verben angeben, auch zu den "deskriptiven" Mitteln oder nicht? Denn in Goethes "Erlkönig" beispielsweise gibt es mehr "deskriptive" Adverbien als Adjektive.

Betrachtet man den Index rein quantitativ, dann sieht man, daß bei gleicher Anzahl von Verben und Adjektiven im Text $Q = 1$ ist, ganz gleich, ob es von beiden viele oder von beiden wenige gibt. Daher beschreibt dieser Index im Grunde einen Gleichgewichtszustand zwischen Verben und Adjektiven und keine isolierte Eigenschaft. Denn der Text kann gleichzeitig sehr "aktiv" und sehr "deskriptiv" sein. Weiter drückt er nur einen einzigen Aspekt der Dimension "Aktivität-Deskriptivität" aus, man kann diese Eigenschaften auch anders charakterisieren.

Wenn man den Wertebereich des Indexes, nämlich das Intervall $\langle 0,\infty)$, betrachtet, so sieht man, daß $Q = 0$ minimale Aktivität bedeutet; der Wert ∞ heißt aber nicht "unendliche Aktivität", denn falls es im Text keine Adjektive gibt, dann würde schon ein Verb die "unendliche Aktivität" signalisieren. Daher bedeutet ∞ "minimale Deskriptivität" ohne jegliche Aussage über die Stärke der Aktivität.

Die Entscheidung darüber, ob ein Text "aktiv" oder "deskriptiv" ist, folgt also nicht direkt aus dem Index. Man weiß nicht, wie groß Q sein muß, damit man den Text als "aktiv" oder "deskriptiv" bezeichnen darf. Man kann nur sagen, daß beim Überwiegen der Adjektive Q im Intervall $\langle 0,1\rangle$ liegt, beim Überwiegen von Verben hingegen im Intervall $(1,\infty)$ - wahrhaft ein sehr asymmetrisches Verhältnis!

Ein weiterer Umstand ist beim Vergleich von Texten gravierend und läßt sich an einem einfachen Beispiel leicht demonstrieren. Betrachten wir zwei Texte (T_1, T_2)

T_1 mit $v_1 = 100$, $a_1 = 2$, woraus $Q = 50$
und T_2 mit $v_2 = 100$, $a_2 = 4$, woraus $Q = 25$.

Der zweite Text hat nur um zwei Adjektive mehr als der erste, aber der Unterschied zwischen Q_1 und Q_2 scheint extrem groß zu sein und ver-

leitet leicht zu eventuell falschen subjektiven Schlüssen. Man kann zwei Quotienten nur dann direkt vergleichen, wenn $a_1 = a_2$. Wann aber kann man sagen, daß ein Unterschied zwischen Q_1 und Q_2 signifikant, systematisch und nicht durch eine zufällige Schwankung der Stichprobe hervorgerufen worden ist?

Da Indizes dieser Art in der Textanalyse recht häufig sind und sinnvoll eingesetzt werden können, werden wir im weiteren einige Aspekte des Aktionsquotienten untersuchen, wobei wir die Analyse von Altmann (1978) zugrundelegen werden.

Es ist immer ratsam, jedoch nicht unbedingt nötig, die Werte eines Indexes auf das Intervall <0,1> zu beschränken. Dies kann man mit einer geeigneten Transformation durchführen, z.B

$$I' = \frac{I}{I + 1}$$

oder

$$I' = \frac{I - I_{min}}{I_{max} - I_{min}}$$

oder

$$I' = \frac{I}{I_{max}}$$

usw. je nach den Umständen. Im Falle des Aktionsquotienten wäre eine Möglichkeit

$$Q' = \frac{v}{(a + v)}. \qquad (2.2.2)$$

In diesem Fall würden wir a + v = n als die ganze Stichprobe betrachten, und Q' wäre lediglich die Proportion von v, d.h. $Q' = p_v$. Daher würde sich Q' im Intervall <0,1> bewegen, und man könnte die ganze Statistik für das Testen der Proportionen anwenden. Q' mißt ebenso wie Q das "aktiv-deskriptive Gleichgewicht", das bei Q' = 0.5 vorhanden ist.

Von (2.2.2) ausgehend sind folgende Hypothesen möglich:

(a) Die Hypothese der ***erhöhten Aktivität***:

$H_A: p > 0.5$

gegen die Nullhypothese, daß keine solche Tendenz besteht:

$H_0: p = 0.5.$

In diesem Fall ist zu berechnen

$$P(X \geq v) = \sum_{x=v}^{n} \binom{n}{x} 0.5^n. \qquad (2.2.3)$$

Setzt man das Signifikanzniveau z.B. auf $\alpha = 0.05$ fest, so kann man aus (2.2.3) folgende Schlüsse ziehen:

Wenn $P(X \geq v) \leq 0.05$, dann ist die Nullhypothese abzulehnen. Man kann dann annehmen, daß eine signifikant erhöhte Aktivität vorliegt.

Wenn $P(X \geq v) > 0.05$, dann kann man nicht sinnvoll auf erhöhte Aktivität schließen, d.h. man nimmt H_0 an.

(b) Hypothese der ***erhöhten Deskriptivität***:

$H_b: p < 0.5$

gegen die Nullhypothese, daß keine solche Tendenz besteht:

$H_0: p = 0.5.$

In diesem Falle ist zu berechnen

$$P(X \leq v) = \sum_{x=0}^{v} \binom{n}{x} 0.5^n \qquad (2.2.4)$$

Wenn $P(X \leq v) \leq 0.05$, dann ist die Nullhypothese abzulehnen,und man kann annehmen, daß eine signifikant erhöhte Deskriptivität vorliegt.

Wenn $P(X \leq v) > 0.05$, dann kann man nich sinnvoll auf eine erhöhte Deskriptivität schließen.

(c) Die Hypothesen (a) und (b) sind einseitig. Jedoch sind einseitige Fragestellungen nur dann korrekt, wenn *vor* der Datenerhebung eine begründete Richtungshypothese aufgestellt werden kann. Daher handelt es sich in den meisten Fällen um die Hypothese des "***aktiv-deskriptiven Ungleichgewichts***":

$$H_U : p \neq 0.5$$

gegen $H_0 : p = 0.5.$

Wenn $Q' > 0.5$, d.h., wenn $v > a$, dann ist (2.2.3) zu berechnen und wenn $Q' < 0.5$, d.h., wenn $v < a$, dann ist (2.2.4) zu berechnen.
Wenn nun $P(X \geq v) \leq 0.025$ oder $P(X \leq v) \leq 0.025$, dann kann man H_0 ablehnen und von einem signifikanten aktiv-deskriptiven Ungleichgewicht sprechen. Andernfalls nimmt man Gleichgewicht an.

Dies ist der übliche ***Binomialtest***, wie man ihn in jedem Lehrbuch der Statistik findet.

Beispiele. (i) Man vermutet in einem gegebenen Text erhöhte Aktivität (in bezug auf Verben und Adjektive). Die Zählung ergibt $v = 20$, $a = 10$. Um die Hypothese zu testen, berechnet man

$$P(X \geq v) = P(X \geq 20) = \sum_{x=20}^{30} \binom{30}{x} 0.5^n = 0.0494.$$

Da $0.0494 < 0.05$, lehnen wir H_0 ab und nehmen an, daß im Text eine signifikant erhöhte Aktivität vorhanden ist.

(ii) In einem Text, für den eine erhöhte Deskriptivität angenommen wird, findet man $v = 13$ und $a = 17$. Um diese Hypothese zu testen, berechnen wir

$$P(X \leq v) = P(X \leq 13) = \sum_{x=0}^{13} \binom{30}{x} 0.5^{30} = 0.2933.$$

Da $0.2933 > 0.05$, nehmen wir die Null-Hypothese an, daß keine signifikant erhöhte Deskriptivität vorliegt.

(iii) In einem Text wurde $v = 20$, $a = 10$ gefunden. Man teste, ob hier Ungleichgewicht vorhanden ist. In (i) oben haben wir festgestellt, daß $P(X \geq 20) = 0.0494$. Da dieser Wert größer als 0.025 ist, nehmen wir H_0 an, d.h., es besteht kein Ungleichgewicht. An diesem Beispiel sieht man, daß sich bei Unkenntnis der Richtung des Ungleichgewichts das Signifikanzniveau halbiert.

2.2.2. Exkurs

Indizes wie den Aktionsquotienten findet man des öfteren in der Textanalyse, meistens verbunden mit Ziehung falscher Schlüsse wegen der nichtkorrekten Einschätzung ihres Verhaltens. Daher scheint es ratsam, einen Exkurs einzufügen, in dem für den Praktiker nur die resultierenden Formeln (2.2.19) bis (2.2.22) relevant sind.

Die folgende Überlegung zeigt, daß bei ziemlich einfach aussehenden Indizes wie Q die Ableitung eines Tests nicht ganz einfach ist. Manchmal, wenn wir Glück haben, läßt sich die ganze Prozedur beträchtlich vereinfachen.

Sei v die Zahl der Verben in einem gegebenen Text, a die Zahl der Adjektive und r die Zahl aller restlichen Wortarten. Weiter sei die (unbekannte) Wahrscheinlichkeit des Vorkommens von Verben p_v, die von Adjektiven p_a und die der restlichen Wortarten p_r. Es gilt

$$v + a + r = n$$
$$p_v + p_a + p_r = 1. \qquad (2.2.5)$$

Philologisch gesehen, stellen wir die Frage, ob im Text ein Gleichgewicht zwischen Deskriptivität und Aktivität besteht. Übersetzt in die Sprache der Statistik testen wir die Hypothese

$$H_0 : p_v = p_a$$

gegen

$$H_U : p_v \neq p_a.$$

Zu diesem Zweck stellen wir den sogenannten Likelihood-Quotienten

$$\lambda = \frac{L(w)}{L(\Omega)}$$

auf, wobei L(w) die Maximum-Likelihood-Funktion mit den Schätzern p_v^* und p_a^* von p_v und p_a unter der Nullhypothese, und $L(\Omega)$ die Maximum-Likelihood-Funktion mit den Schätzern p_v^* und p_a^* von p_v und p_a im ganzen Parameterraum (unter der Menge der Alternativhypothesen) ist. Die Wahrscheinlichkeit, daß wir bei der Untersuchung eines gegebenen Texts genau v Verben, a Adjektive und r restliche Wortarten in der gegebenen Reihenfolge vorfinden, ist

$$L = p_v^v p_a^a p_r^r \qquad (2.2.6)$$

Unter der Nullhypothese "$p_v = p_a$" folgt daraus

$$L = p_v^{v+a} p_r^r \; . \qquad (2.2.7)$$

Da laut (2.2.5) $p_r = 1 - p_v - p_a$, so wird aus (2.2.7)

$$L = p_v^{v+a}(1 - 2p_v)^r \qquad (2.2.8)$$

Zur Bestimmung eines "besten Schätzers" p_v^* von p_v geht man von dem Bestreben aus, p_v^* so zu bestimmen, daß die in der Stichprobe vorgefundene Wertekonstellation (hier: a,v,r) die wahrscheinlichste ist. Dies ist gerade dann der Fall, wenn L bezüglich p_v ein Maximum annimmt, was mit der üblichen Extremalaufgabe $\delta L/\delta p_v = 0$ zu lösen ist. Es ist jedoch einfacher, stattdessen mit $\delta \ln L/\delta p_v = 0$ zu rechnen, was erlaubt ist, da der Logarithmus seine Extrema an denselben Stellen wie die Funktion hat:

$$\ln L = (v+a)\ln p_v + r \ln(1-2p_v)$$

$$\frac{\delta \ln L}{\delta p_v} = \frac{v+a}{p_v} - \frac{2r}{1-2p_v} = 0,$$

woraus

$$p_v^* = \frac{v+a}{2n} \qquad (2.2.9)$$

folgt. Dasselbe gilt auch für p_a^*, wenn man in (2.2.7) statt p_v überall p_a schreibt und die Schätzung berechnet.

Setzt man (2.2.9) in (2.2.8) ein, so kann man schreiben

$$L(w) = \left[\frac{v+a}{2n}\right]^{v+a} \left[1 - \frac{v+a}{n}\right]^r \; . \qquad (2.2.10)$$

Die Schätzung von p_v und p_a unter der Annahme der Alternativhypothesen finden wir auf dieselbe Weise. Durch Logarithmieren von (2.2.6) bekommen wir

$$\ln L = v \ln p_v + a \ln p_a + r \ln p_r$$

$$= v \ln p_v + a \ln p_a + r \ln(1 - p_v - p_a).$$

Die Ableitung nach p_v bzw p_a ergibt

$$\frac{\delta \ln L}{\delta p_v} = \frac{v}{p_v} - \frac{r}{1-p_v-p_a} = \frac{v}{p_v} - \frac{r}{p_r}$$

bzw.

$$\frac{\delta \ln L}{\delta p_a} = \frac{a}{p_a} - \frac{r}{1-p_v-p_a} = \frac{a}{p_a} - \frac{r}{p_r} \; .$$

Setzt man diese Ausdrücke gleich 0, so erhält man laut (2.2.5)

$$p_v^* = \frac{vp_r^*}{r} \qquad (2.2.11)$$

bzw.

$$p_a^* = \frac{ap_r^*}{r} . \qquad (2.2.12)$$

Addiert man (2.2.11) und (2.2.12), so bekommt man

$$p_v^* + p_a^* = 1 - p_r^* = \frac{(v + a)p_r^*}{r} , \qquad (2.2.13)$$

woraus

$$\frac{p_r^*}{r} = \frac{1}{n} \quad \text{folgt.} \qquad (2.2.14)$$

Setzt man (2.2.14) in (2.2.11) und (2.2.12) ein, so wird

$$p_v^* = \frac{v}{n} \qquad (2.2.15)$$

und

$$p_a^* = \frac{a}{n}. \qquad (2.2.16)$$

Daraus folgt nach Einsetzung von (2.2.15) und (2.2.16) in (2.2.6)

$$L(\Omega) = (\frac{v}{n})^v (\frac{a}{n})^a (1 - \frac{v + a}{n})^r . \qquad (2.2.17)$$

Aus (2.2.10) und (2.2.17) bilden wir den Likelihood-Quotienten als

$$\lambda = \frac{L(w)}{L(\Omega)} = \frac{(\frac{v+a}{2n})^{v+a} (1 - \frac{v+a}{n})^r}{(\frac{v}{n})^v (\frac{a}{n})^a (1 - \frac{v+a}{n})^r}$$

$$= \frac{((v + a)/2)^{v+a}}{v^v a^a} \qquad (2.2.18)$$

Die Funktion $-2 \ln \lambda$ ist ungefähr wie Chiquadrat, hier mit einem Freiheitsgrad verteilt (vgl. Kendall, Stuart 1967,II:240f.), d.h.

$$-2 \ln \lambda = 2v \ln v + 2a \ln a - 2v \ln((v+a)/2) - 2a \ln((v+a)/2)$$

$$= 2\Sigma x \ln(2x/(v+a)) \approx X^2_{(1)}, \quad x = a,v. \qquad (2.2.19)$$

Diese Funktion kann leicht für Testzwecke verwendet werden, da $2x \ln x$ tabelliert ist (vgl. Kullback, Kupperman, Ku 1962), so daß man sie auch ohne Computer benutzen kann. Setzt man jedoch

$$\frac{x - (v+a)/2}{(v + a)/2} = \delta,$$

so ist

$$\frac{x}{(v+a)/2} = 1 + \delta,$$

d.h. (2.2.19) wird zu

$$2\Sigma x \ln(1 + \delta).$$

Entwickelt man $\ln(1+\delta)$ in eine Taylorreihe, und nimmt man nur die ersten zwei Glieder, so bekommt man

$$2\Sigma x \ln(1+\delta) = 2\Sigma x(\delta - \delta^2/2)$$

$$= 2\Sigma[(x - \frac{v+a}{2}) + \frac{v+a}{2}](\delta - \delta^2/2),$$

woraus unter der Bedingung, daß *v und a groß sind*, folgt

$$X^2_{(1)} = \sum_x \frac{(x - (v+a)/2)^2}{(v + a)/2} , \quad x = a,v. \qquad (2.2.20)$$

Entwickelt man das Binom und führt die Summation durch, so ergibt sich schließlich

$$X^2_{(1)} = \frac{(v - a)^2}{v + a} \quad . \qquad (2.2.21)$$

Dieses einfache Kriterium ist für einen zweiseitigen Test bei großen Stichproben leicht zu verwenden.

Beispiel. Wenn v = 20 und a = 10, so ergibt sich nach (2.2.21)

$$X^2 = \frac{(20 - 10)^2}{20 + 10} = 3.33.$$

Der kritische Wert $X^2_{0.05(1)} = 3.84$. Da das berechnete X^2 kleiner als 3.84 ist, nehmen wir die Hypothese des "aktiv-deskriptiven Gleichgewichts" an.■■■

Man beachte, daß wir den Test ohne die Bildung des Quotienten durchgeführt haben. Will man jedoch den Quotienten Q' bilden und dann testen, so kann man bei großen Stichproben auch folgendermaßen verfahren: Da $Q'^* = p^*$, so ist $E(Q'^*) = p$ und $V(Q'^*) = pq/n$. Transformiert man Q'^* auf eine Normalvariable, dann ergibt sich

$$z = \frac{Q'^* - E(Q'^*)}{[V(Q'^*)]^{1/2}} = \frac{Q'^* - p}{[pq/n]^{1/2}} \quad .$$

Da unter der Nullhypothese p = 0.5 ist, so bekommt man

$$\frac{Q'^* - 0.5}{[0.5(0.5)/n]^{1/2}} = (2Q'^* - 1)\sqrt{n} = z, \qquad (2.2.22)$$

was mit der Wurzel von (2.2.21) identisch ist, wenn man $Q'^* = v/(v+a)$ und n = v+a einsetzt.

Beispiel. Sei wieder v = 20 und a = 10, so wird laut (2.2.22)

$$[2(20/30) - 1]\sqrt{30} = 1.83,$$

d.h. ein nichtsignifikantes Resultat. Es ist leicht nachzuprüfen, daß $(1.83)^2 = 3.33 = X^2_{(1)}$, was wegen $z^2 = X^2_{(1)}$ übereinstimmen muß. Bei diesem kleinen Stichprobenumfang weicht jedoch die berechnete Wahrscheinlichkeit vom exakten Wert beträchtlich ab.■■■

2.2.3. Vergleich von Aktionsquotienten

Eines der oft aufgegriffenen Probleme bei der Textanalyse ist der Vergleich zweier Indizes. Die einfache Feststellung, daß beispielsweise bei einem Verfasser $Q_1 = 6.75$ und bei einem anderen $Q_2 = 7.05$ ist, reicht nicht aus, um den zweiten Verfasser als "aktiver" einzustufen. Ein statistischer Test ist hier unentbehrlich. Wiederum zeigen wir, wie man in diesem einfachen Fall verfahren kann.

Ohne einen Index zu bilden, bewerten wir die Proportionen von Verben in zwei Stichproben. Wir wissen aus § 2.2.1, daß das Vorkommen von Verben einer Binomialverteilung folgt. Unsere Nullhypothese lautet

$$H_0: p_1 = p_2 ,$$

wobei p_1 den Parameter der Binomialverteilung im Text T_1 darstellt. Es kann wiederum drei sinnvolle Alternativhypothesen (wie oben) geben, und zur Entscheidung kann man z.B. mit Hilfe des Fischerschen exakten Tests gelangen (vgl. z.B. Siegel 1956:96).

Bei großen Stichproben ergibt der Vergleich zweier Proportionen, d.h. von $Q_1'^*$ und $Q_2'^*$, einfach

$$z = \frac{(v_1 n_2 - v_2 n_1)(n_1 + n_2)^{1/2}}{[n_1 n_2 (v_1 + v_2)(a_1 + a_2)]^{1/2}} \quad , \qquad (2.2.23)$$

wo $n_1 = v_1 + a_1$, $n_2 = v_2 + a_2$. Formel (2.2.23) folgt aus

$$z = \frac{Q_1'^* - Q_2'^*}{[V(Q_1'^* - Q_2'^*)]^{1/2}} \qquad (2.2.23a)$$

durch Einsetzung von $Q_i'^* = p_i^* = v_i/(v_i + a_i)$ und

$$V(Q_1'^* - Q_2'^*) = V(p_1^* - p_2^*) = p^* q^* (1/n_1 + 1/n_2),$$

wo $p^* = (v_1 + v_2)/(n_1 + n_2)$.

Benutzt der Philologe jedoch Q und testet die Differenz $Q_1^* - Q_2^*$, so muß er vorsichtig verfahren. Will er bei großen Stichproben zur Normalverteilung übergehen und unter der Nullhypothese das Kriterium

$$\frac{Q_1^* - Q_2^*}{[V(Q_1^* - Q_2^*)]^{1/2}} = z \qquad (2.2.24)$$

verwenden, so muß er die Varianz einsetzen, die mit $V(Q_1'^* - Q_2'^*)$ nicht identisch ist. Es ist

$$V(Q_1^* - Q_2^*) = V(Q_1^*) + V(Q_2^*).$$

$V(Q^*)$ leitet man durch Linearisierung in der Umgebung des Erwartungswertes E als

$$V(Q) = V\left(\frac{v}{a}\right) = \left(\frac{\delta Q}{\delta v}\right)^2_E V(v) + \left(\frac{\delta Q}{\delta a}\right)^2_E V(a) + 2\left(\frac{\delta Q}{\delta v}\right)\left(\frac{\delta Q}{\delta a}\right)_E \mathrm{Cov}(v,a) \qquad (2.2.25)$$

ab, wenn man nur die ersten Glieder der Entwicklung benutzt. Greift man auf die ursprüngliche Multinomialverteilung zurück, wo

$$E(v) = np_v, \quad E(a) = np_a$$

$$V(v) = np_v q_v, \quad V(a) = np_a q_a$$

$$\mathrm{Cov}(v,a) = -np_v p_a,$$

so wird (2.2.25) zu

$$V(Q^*) = \left(\frac{1}{np_a}\right)^2 np_v q_v + \left[\frac{-np_v}{(np_a)^2}\right]^2 np_a q_a + \frac{2n^2 p_v^2 p_a}{(np_a)^3}$$

$$= \frac{Q^2}{n}\left(\frac{1}{p_v} + \frac{1}{p_a}\right). \qquad (2.2.26)$$

Setzt man hier die Schätzung $p_v^* = v/n$ und $p_a^* = a/n$ ein, so bekommt man

$$V(Q^*) = Q^2\left(\frac{1}{v} + \frac{1}{a}\right). \qquad (2.2.27)$$

Daraus ergibt sich

$$V(Q_1^* - Q_2^*) = Q_1^2\left(\frac{1}{v_1} + \frac{1}{a_1}\right) + Q_2^2\left(\frac{1}{v_2} + \frac{1}{a_2}\right). \qquad (2.2.28)$$

Da unter der Nullhypothese $Q_1 = Q_2$ ist, schreiben wir

$$V(Q_1^* - Q_2^*) = Q^2\left(\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}\right),$$

so daß (2.2.24) zu

$$z = \frac{Q_1^* - Q_2^*}{Q\left(\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}\right)^{1/2}} \qquad (2.2.29)$$

wird. Q selbst kann man als das geometrische Mittel beider Quotienten abschätzen, d.h. als

$$Q^* = (Q_1^* Q_2^*)^{1/2},$$

so daß schließlich

$$z = \frac{Q_1^* - Q_2^*}{(Q_1^* Q_2^*)^{1/2}\left(\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}\right)^{1/2}} \cdot \qquad (2.2.30)$$

Beispiele. (i) Sei im Text T_1 $v_1 = 60$, $a_1 = 40$ und in T_2 $v_2 = 50$, $a_2 = 50$. Daraus ergeben sich $Q_1 = 60/40 = 1.5$ und $Q_2 = 50/50 = 1$. Laut (2.2.23) ergibt sich

$$z = \frac{[60(100) - 50(100)]\sqrt{200}}{[100(100)110(90)]^{1/2}} = 1.42$$

und laut (2.2.30)

$$z = \frac{1.5 - 1}{[1.5(1)]^{1/2}(1/60+1/40+1/50+1/50)^{1/2}} = 1.43.$$

Im Falle einer zweiseitigen Hypothese wäre die Gleichheit beider Q_i^* anzunehmen, da $1 - P(-1.42 \leq z \leq 1.42) = 2\Phi(-1.42) = 0.15560$.■■■

(ii) Setzt man jedoch $v_1 = 40$, $a_1 = 60$, $v_2 = a_2 = 60$, so bekommt man $Q_1^* = 40/60 = 0.67$, $Q_2^* = 50/50 = 1$ und

$$|Q_1^* - Q_2^*| = 0.33,$$

d.h. rein "optisch" einen anderen Unterschied als in Beispiel (i), wo $|Q_1^* - Q_2^*| = 0.5$ war. Berechnet man jedoch z, so bekommt man genau ***dasselbe Resultat*** wie im Beispiel (i).■■■

Diese Überlegungen sollten zeigen, daß die Charakterisierung eines Textes durch einen Index erst dann sinnvoll und weiterführend ist, wenn

(a) der Index eine ***eindeutige philologische Interpretation*** hat, d.h., wenn man weiß, was die Zahlen aussagen;

(b) man weiß, welche ***Werte*** der Index annehmen kann und was die einzelnen Werte bedeuten;

(c) falls der Index für Vergleichszwecke verwendet wird, man eine statistische Prozedur zur Verfügung hat, mit deren Hilfe man die ***Signifikanz von Unterschieden*** beurteilen kann.

Betrachten wir noch einige konkrete Beispiele aus deutschen Texten, wie sie von H. Fischer (1969) ausgezählt worden sind (vgl. Tabelle 2.1).

Bis auf den Text (8) zeigen alle Texte eine erhöhte "Aktivität" ($Q' > 0.5$). Ob diese Aktivität nicht nur zufällig ist, überprüfen wir mit dem Kriterium (2.2.21). So erhalten wir für den Text (1)

$$X^2 = \frac{(81-22)^2}{81+22} = 33.80,$$

was von einer starken Aktivität zeugt. Die Resultate einzelner Tests sind in der Tabelle 2.2 angegeben (Spalte 3).

Tabelle 2.1

Aktionsquotienten deutscher Texte

nach H.Fischer(1969)

Text	Verben	Adjektive	Q	Q'
1.G.Schwab, Des Odysseus Heimkehr nach Ithaka	81	22	3.68	0.79
2.W.von Ebner-Eschenbach,Die Nachbarn	81	32	2.53	0.72
3.J.G.Herder, Die drei Freunde	37	4	9.25	0.90
4.M.Pestalozzi, Der Abend vor einem Festtage im Hause einer rechtschaffenen Mutter	54	12	4.50	0.82
5.J.Gotthelf, Jakobs Lehrjahre	93	30	3.10	0.76
6.W.Raabe, Bekenntnis einer alten Mutter	43	11	3.91	0.80
7.J.P.Heber, Kannitverstan	88	48	1.83	0.65
8.M.Waser, Auf der See	49	82	0.60	0.37
9.A.Stifter, Die Lawine	33	28	1.18	0.54
10.G.Keller, Karl Hedigers Schützenfestrunde	70	40	1.75	0.64

In der vierten Spalte findet man die entsprechende Wahrscheinlichkeit, mit der man den berechneten oder einen noch extremeren Chiquadrat-Wert erwarten würde. Wie man leicht sieht, hängt die Beurteilung

der Größe eines Aktionsquotienten nicht nur von seiner absoluten Größe ab (nach der die Texte in der Tabelle 2.2 geordnet wurden), sondern auch von der Größe der erhobenen Stichprobe. Bis auf Text 8, der eine signifikante "Deskriptivität" zeigt, und Text 9, der im "Gleichgewicht" steht, weisen alle anderen eine signifikante Aktivität auf. Jedoch sieht man gleichzeitig, daß sich die größte Signifikanz nicht bei Text 3 mit dem größten Q', sondern bei Text 1 zeigt. Dies bedeutet nicht, daß Text 1 eine größere "Aktivität" hätte als Text 3, sondern daß unser Urteil über die "Aktivität" des Textes 1 mit einem viel kleineren Risiko eines Fehlurteils gefällt wurde als bei Text 3.

Tabelle 2.2

Tests für den Aktionsquotienten

Text	Q'	X^2	P	z für Unterschiede	P einseitig
8	0.37	8.31	0.0039	2.19	0.014
9	0.54	0.41	0.52	1.22	0.111
10	0.64	8.18	0.0042	0.17	0.433
7	0.65	11.76	0.0006	1.18	0.119
2	0.72	21.25	$4(10^{-6})$	0.69	0.245
5	0.76	32.27	10^{-8}	0.54	0.295
1	0.79	33.80	$6(10^{-9})$	0.15	0.440
6	0.80	18.96	10^{-5}	0.30	0.382
4	0.82	26.73	$2(10^{-7})$	1.20	0.115
3	0.90	26.57	$3(10^{-7})$		

Zwar haben wir die Texte nach ihrer "Aktivität" geordnet, aber daraus folgt noch nicht, daß sie sich in dieser Hinsicht auch signifikant (= nichtzufällig) voneinander unterscheiden. Zur Beurteilung dieses Problems kann man das Kriterium (2.2.23) oder (2.2.29) benutzen und den Unterschied zweier nach der Größe von Q' (oder Q) benachbarte Texte testen. So ergibt sich z.B. für Text 8 und Text 9 nach (2.2.23a) (für Q')

$$p^* = \frac{v_8 + v_9}{n_8 + n_9} = \frac{49 + 33}{131 + 61} = 0.4271$$

$$q^* = 1 - p^* = 0.5729$$

$$[V(p_1^* - p_2^*)]^{1/2} = [0.4271(0.5729)(1/131 + 1/61)]^{1/2}$$

$$= 0.0767$$

und

$$z = \frac{0.54 - 0.37}{0.0767} = 2.22.$$

Tabelle 2.3

Tests für Unterschiede der "Aktivität–Deskriptivität" zwischen den Texten

(erste Zeile = z, zweite Zeile = P)

2	3	4	5	6	7	8	9	10	Text
1.18 0.12	1.65 0.05	0.50 0.31	0.54 0.295	0.15 0.44	2.38 0.009	6.89 $2.8(10^{-12})$	3.41 0.0003	2.44 0.007	1
	2.45 0.007	1.53 0.063	0.69 0.245	1.10 0.14	1.18 0.12	5.67 $7(10^{-9})$	2.36 0.009	1.29 0.10	2
		1.20 0.115	2.03 0.02	1.42 0.08	3.24 0.0006	6.60 $2.1(10^{-11})$	4.17 0.00002	3.31 0.0005	3
			0.98 0.16	0.30 0.38	0.30 0.38	6.47 $4.9(10^{-11})$	3.52 0.0002	2.61 0.005	4
				0.58 0.28	1.93 0.03	6.62 $1.8(10^{-11})$	3.03 0.001	2.01 0.02	5
					2.03 0.02	5.64 $8(10^{-9})$	2.99 0.001	2.11 0.02	6
						4.61 $2(10^{-4})$	1.41 0.08	0.17 0.43	7
							2.19 0.014	4.18 10^{-5}	8
								1.22 0.11	9

Nach (2.2.29) bekämen wir (für Q)

$$z = \frac{1.18 - 0.60}{[1.18(0.60)]^{1/2}[1/49+1/82+1/33+1/28]^{1/2}}$$

$$= \frac{0.58}{0.2642} = 2.19.$$

Die Resultate der Tests für "benachbarte" Texte, durchgeführt mit (2.2.29), sind in Tabelle 2.2, Spalte 5 zu finden. Der einzige signifikante Wert ist der zwischen den Texten 8 und 9, alle anderen sind nichtsignifikant.

Tabelle 2.3 enthält alle z-Werte für Unterschiede zwischen den Texten 1 bis 10. Eine Clusteranalyse der Texte kann mit einer geeigneten taxonomischen Methode durchgeführt werden (vgl. z.B. Bock 1979).

2.3. Alle Variablen

Im vorigen Abschnitt haben wir den Umgang mit zwei Variablen betrachtet. Statistische Modelle muß man nach der Art und Weise einsetzen, wie man die Variablen verknüpft, und es ist ratsam, die Interpretation, den Wertebereich, die Wahrscheinlichkeitsverteilung bzw. eine Transformation zu kennen, bevor man zu zählen anfängt.

In der Textanalyse hat man früher auch kompliziertere Indizes eingeführt, ohne jedoch die obigen Fragen zu beantworten. So findet man etwa bei F.Schmidt (1972) eine philologisch sinnvolle Abstufung der Prädikate im Satz in Haupt- (a),in Neben- (b) und in Zusatzprädikate (c), woraus der Autor einen Index $\alpha = ac/b^2$ bildet. Summiert man α_i für alle Sätze (i) des Textes und bildet die Größe $n/\Sigma\alpha_i$, wobei n die Anzahl aller Prädikate im Text ist, dann erhält man eine Größe, die man als den Prädikationskoeffizienten bezeichnen kann (Schmidt nennt ihn Stilquotient). Hier sind drei Variablen vorhanden, der Wertebereich ist $\langle 0,\infty)$, aber wie die einzelnen Werte zu interpretieren sind und wie man Vergleiche und Schlüsse ziehen kann, ist unbekannt.

Noch mehr Variablen enthält z.B. Birkhoffs Index für die musikalische Qualität eines Gedichts: M = (aa+2r+2m-2ae-2ce)/C, wobei die einzelnen Bestandteile der Formel Zahlen bedeuten, z.B. "ce = Anzahl der Konsonanten - 2·Anzahl der Vokale", oder "ae = Anzahl derjenigen Laute, die mit mehr als zwei vorangehenden Leitlauten direkt verbunden sind,

vermehrt durch die Anzahl der Laute einer Silbe, die einer ihr identischen Silbe unmittelbar folgt, sofern sie nicht zum selben Wort gehört, und vermehrt durch die Anzahl derjenigen Laute, die zu einer Periode gehören, wobei dieser Folge mindestens zwei vorangehende Laute angehören" (zitiert nach Gunzenhäuser 1969:302-303). Die Stichprobenverteilung dieses Indexes wäre so kompliziert, daß es fraglich ist, ob er sinnvoll angewendet werden kann.

Wir werden derartige Indizes hier nicht analysieren, weil man gleichwertige, einfachere Indizes aufstellen kann. Wir widmen uns im weiteren der Erfassung der formlosen Wiederholung aller Elemente einer Menge von Texteinheiten, d.h. solchen, die den gesamten Text abdecken.

Es haben sich in der Textanalyse im Laufe der Zeit zwei Arten beschreibender Mittel kristallisiert, nämlich ***globale Indizes*** wie Mittelwert, Entropie, Wiederholungsrate, "Yulesche Charakteristik" u.a., und ***Wahrscheinlichkeitsverteilungen***, die die Verteilung einer (meistens diskreten) Eigenschaft im Text erfassen. Im folgenden werden beide nur auswahlsweise behandelt, die Literatur ist sehr umfangreich.

2.3.1. Die Entropie

Seit der Etablierung der Informationstheorie wurde die Entropie ständig als eine Charakteristik einer Sprache, eines Textes oder auch anderer künstlerischer Produkte betrachtet. Man hat das Entropiemaß an zahlreichen unterschiedlichen Texteinheiten berechnet und ihm so diverse Interpretationen gegeben, daß allein die Aufzählung der Resultate einige Dutzende von Seiten in Anspruch nehmen würde. Die Interpretationen sind poetischer, stilistischer, ästhetischer, kommunikationstheoretischer u.a. Art und stellen sekundäre, sachbezogene Umdeutungen der elementaren Bedeutung der Entropie, nämlich der ungleichmäßigen Verteilung der relativen Häufigkeiten einer Menge von Texteinheiten dar.

Die Entropie berechnet sich nach der Formel

$$H_1' = -\Sigma p_i \, ld \, p_i,$$

wo p_i die Wahrscheinlichkeit der i-ten Texteinheit der untersuchten Menge von n Texteinheiten ist und ld der Logarithmus zur Basis 2. Die p_i-Werte pflegt man mit den relativen Häufigkeiten zu schätzen, d.h. mit

$$p_i^* = f_i/N,$$

wo f_i = die absolute Häufigkeit der Einheit i und
N = Anzahl aller Einheiten im Text, d.h. $N = \Sigma f_i$ ist.

Statt des Logarithmus zur Basis 2 werden wir im folgenden den natürlichen Logarithmus benutzen, was uns Umrechnungen erspart und Transformationen erleichtert, d.h., wir definieren

$$H_1 = -\Sigma p_i \ln p_i. \qquad (2.3.1)$$

In diesem Fall ergibt sich aus (2.3.1)

$$H_1 = -\Sigma \frac{f_i}{N} \ln \frac{f_i}{N}$$

$$= \ln N - \frac{1}{N}\Sigma f_i \ln f_i. \qquad (2.3.2)$$

Das Entropiemaß läßt sich sowohl auf qualitative als auch auf quantitative Variablen anwenden.

Beispiele. (1) Grotjahn (1979:176) hat in Goethes "Erlkönig" die Wortlängen, gemessen in Silben, gezählt und erhielt die Resultate in Tabelle 2.4.

Tabelle 2.4

Verteilung der Wortlängen in Silben in Goethes Erlkönig"
nach Grotjahn (1979)

Zahl der Silben im Wort i	1	2	3	4	Σ
Zahl der Wörter der Länge i: f_i	152	55	6	2	215

Mit Hilfe der Formel (2.3.2) erhalten wir

$$H_1 = \ln 215 - \frac{1}{215}(152 \ln 152 + 55 \ln 55 + 6 \ln 6 + 2 \ln 2).$$

Den Wert x ln x findet man tabelliert in Kullback, Kupperman, Ku (1962). So erhalten wir

$$H_1 = 5.3706 - 0.0047(763.6298 + 220.4033 + 10.75056 + 1.3863)$$

$$= 0.7373.$$

In diesem Fall kann man diese Zahl als die ***Diversität*** der Wortlängen interpretieren. Die Zahl selber sagt uns nicht, wie groß diese Diversität ist, und deswegen ist es vernünftig, H_1 irgendwie zu normieren.

In welchem Intervall bewegt sich H_1? Nehmen wir an, daß wir n Klassen haben (beim Erlkönig war n = 4) und jede Klasse die gleiche Wahrscheinlichkeit hat, d.h. $p_i = 1/n$. In diesem Fall ergibt (2.3.1)

$$H_0 = -\Sigma \frac{1}{n} \ln \frac{1}{n} = -\frac{1}{n}\Sigma(\ln 1 - \ln n) = \ln n. \qquad (2.3.3)$$

Wenn aber eine Klasse die Wahrscheinlichkeit p = 1 hat und alle anderen $p_i = 0$, dann bekommen wir

$$H_1 = -1 \ln 1 - 0 \ln 0 - \ldots - 0 \ln 0 = 0,$$

wobei wir 0 ln 0 = 0 setzen. Daher bewegt sich H_1 im Intervall $\langle 0, \ln n \rangle$. Man pflegt ln n als H_0 oder als H_{max} zu bezeichnen. Setzen wir also

$$H_{1rel} = \frac{H_1}{H_0},$$

dann erhalten wir die ***relative Entropie.*** In unserem Beispiel wäre

$$H_{1rel} = \frac{0.7373}{\ln 4} = \frac{0.7373}{1.36829} = 0.5318.$$

Zum Vergleich bringen wir noch einige Entropiewerte für die Wortlängenverteilungen, die von Grotjahn (1979:174–177) ausgezählt wurden (vgl. Tabelle 2.5, Spalte 2 und 3).

Tabelle 2.5

Entropiewerte für Wortlängenverteilungen
nach Grotjahn

Text	H_1	H_{1rel}	Var H
Goethe: Erlkönig	0.7373	0.5318	0.00268760
Der Totentanz	0.8693	0.6270	0.00015440
Brief Nr. 589	1.0561	0.6562	0.00150811
Nr. 591	0.9730	0.7019	0.00005172
Nr. 596	1.1580	0.7195	0.00167903
Nr. 605	0.9716	0.5423	0.00086099
Nr. 612	1.1118	0.6908	0.00149714
Nr. 641	1.0027	0.6230	0.00120948
Nr. 644	1.2021	0.6177	0.00088160
Nr. 647	1.0872	0.6068	0.00155308
Nr. 659	0.9599	0.5940	0.00269397
Nr. 667	1.2652	0.7861	0.00270826
Schiller: Die Kraniche des Ibycus	0.9990	0.6207	0.00052240
Lukrez: De Rerum Natura	1.3521	0.7546	0.00004468
De Arte Poetica	1.3454	0.8359	0.00008340
Caesar: De Bello Gallico	1.5964	0.8204	0.00033559
Sallust: Bellum Iugurthinum	1.4431	0.8054	0.00039049

(ii) Drobisch (1866) hat die Häufigkeit der einzelnen Verstypen bei Vergil ausgezählt und erhielt Daten, wie in Tabelle 2.6 mit Grotjahns (1979:207) Korrektur dargestellt. Hier bedeutet S = Spondeus, D = Daktylus, wobei die letzten zwei Versfüße unbeachtet geblieben sind.

Die Entropie ergibt sich wieder als

$$H_1 = \ln 1760 - \frac{1}{1760}(118 \ln 118 + \ldots + 38 \ln 38) = 2.5958$$

und

$$H_{1rel} = 2.5958/2.7726 = 0.9362.$$

Tabelle 2.6

Verstypen bei Vergil
(Drobisch 1866; Grotjahn 1979)

Verstypen	Anzahl
SSSS	118
SSSD	51
SSDS	89
SDSS	178
DSSS	263
SSDD	32
SDSD	66
SDDS	99
DSSD	110
DSDS	199
DDSS	208
SDDD	36
DSDD	65
DDSD	80
DDDS	128
DDDD	38

S=Spondeus, D=Daktylus

Je größer die Entropie, desto gleichmäßiger sind die Häufigkeiten verteilt. Wir sehen zwar, daß im ersten Beispiel bei den lateinischen Schriftstellern etwas höhere Werte als bei den deutschen erscheinen, aber allein aus der Kenntnis des Wertebereichs kann man noch keine Schlüsse ziehen.

Bei den Hexametern von Vergil ist $H_{1rel} = 0.9362$, d.h. so hoch, daß man "optisch" geneigt wäre, auf eine Gleichverteilung von Häufigkeiten zu schließen. Daß dies nicht der Fall ist, werden wir unten zeigen.

Hier drängt sich aber eine andere, sekundäre Interpretation der Entropie auf. Ist sie nämlich klein, dann trägt eine einzige Klasse den größten Teil der Häufigkeiten, während die anderen sehr niedrig oder 0 sind. Wenn beispielsweise alle Hexameter das Muster DDDD hätten, dann wäre ein Gedicht rhythmisch monoton. Wechseln sich die Muster ab, dann wirkt es nicht monoton. Daher ist die Entropie ein Maß der ***Monotonie***, der ***Stereotypie*** : Je niedriger die Entropie, desto größer die Monotonie; je größer die Entropie, desto heterogener ist der Text gestaltet, desto mehr tendieren die Klassen zur Gleichverteilung.

Ob nun eine Gleichverteilung angenommen werden kann oder nicht, überprüft man mit einem Chiquadrat-Test für Homogenität. Die Formel lautet

$$X^2 = \sum_{i=1}^{n} \frac{(f_i - E_i)^2}{E_i}, \qquad (2.3.5)$$

wo f_i die beobachteten und E_i die erwarteten Häufigkeiten sind. Da wir Gleichverteilung testen, ist jedes $E_i = N/n$, so daß aus (2.3.5)

$$X^2 = \frac{n}{N} \sum_i f_i^2 - N \qquad (2.3.6)$$

wird.

Berechnen wir diese Größe für die Verstypen von Vergil, so erhalten wir

$$X^2 = \frac{16}{1760}(118^2 + 51^2 + \ldots + 128^2 + 38^2) - 1760 = 650.85.$$

Dieses X^2 hat 15 Freiheitsgrade, so daß wir sicher schließen können, daß es sich hier um keine Gleichverteilung handelt. Hat man jedoch H_1 berechnet, dann kann man sich die Berechnung von (2.3.5) sparen, da

$$X^2 \approx 2N(H_0 - H_1) \qquad (2.3.7)$$

bzw., wenn man (2.3.5) hat, kann man die Entropie als

$$H_1 \approx H_0 - \frac{X^2}{2N} \qquad (2.3.8)$$

berechnen (vgl. Altmann, Lehfeldt 1980:176-178). Im Beispiel von Vergils Hexameter erhalten wir X^2 aus (2.3.7) als

$$X^2 \approx 2(1760)(2.7726 - 2.5958) = 622.34,$$

während sich umgekehrt, nach (2.3.8), $H_1 = 2.5877$ ergibt.

2.3.2. Vergleich zweier Entropien

Will man zwei Texte in bezug auf ihre Entropien vergleichen, so benutzt man einen t-Test, wie er von Hutcheson (1970) vorgeschlagen wurde. Man berechnet das Kriterium

$$t = \frac{H_1 - H_2}{(\text{Var } H_1 + \text{Var } H_2)^{1/2}}, \qquad (2.3.9)$$

wo H_1, H_2 die betreffenden Entropien sind; Var H berechnet man nach der Formel (vgl. Miller, Madow 1954/1963; Basharin 1959; Bowman, Hutcheson, Odum, Shenton 1969; Hutcheson 1970):

$$\text{var } H = \frac{\sum p_i \ln^2 p_i - H^2}{N} + O(1/N^2), \qquad (2.3.10)$$

und die Freiheitsgrade ergeben sich als

$$FG = \frac{(\text{Var } H_1 + \text{Var } H_2)^2}{\frac{(\text{Var } H_1)^2}{N_1} + \frac{(\text{Var } H_2)^2}{N_2}}. \qquad (2.3.11)$$

Die Glieder der Ordnung $O(1/N^2)$ in der Entwicklung von Var H pflegt man (besonders bei großem N) auszulassen.

Beispiel. Betrachten wir die Entropien der Wortlängen im "Totentanz" und im "Erlkönig" von Goethe. Der Tabelle 2.5 entnehmen wir

$$t = \frac{0.8693 - 0.7373}{(0.0001544 + 0.0026876)^{1/2}} = 2.48.$$

Die Freiheitsgrade ergeben sich aus der Formel (2.3.11) als

oder nach McIntosh (1967)

$$R_{rel} = \frac{1 - \sqrt{R}}{1-1/\sqrt{n}} \quad . \qquad (2.3.18)$$

2.3.4. Kenngrößen von Häufigkeitsverteilungen

Ist die untersuchte Texteinheit eine quantitative Variable, die ganzzahlige oder reelle Werte annehmen kann, dann besteht meistens die Möglichkeit, die Häufigkeiten ihrer einzelnen Werte im Text zu ermitteln. In der Linguistik und der Textanalyse haben sich sogar Rangvariablen (Ordinalvariablen) beheimatet und werden besonders im Zipf-Mandelbrotschen Gesetz verwendet.

Tabelle 2.7

Verteilung der Wortlängen
(gemessen in Silbenzahl in vier Texten)
(nach Grotjahn 1979:177)

Zahl der Silben im Wort x	Zahl der Wörter mit x Silben			
	CAESAR De bello gallico f_x	SALLUST Bellum Iugurthinum f_x	GOETHE Der Totentanz f_x	SCHILLER Die Kraniche des Ibycus f_x
1	184	122	218	580
2	204	249	99	296
3	194	196	21	97
4	125	110	4	24
5	54	24	–	1
6	13	1	–	–
7	1	–	–	–
Σ	775	702	342	998

Die Häufigkeitsverteilungen erlauben es, weitere globale Indizes aufzustellen, Texte zu charakterisieren und zu vergleichen. Einige von ihnen sollen im weiteren dargestellt werden. Zur Illustration verwenden

wir vier von Grotjahn (1979) ermittelte Verteilungen, wie sie in Tabelle 2.7 dargestellt sind.

Der Mittelwert

Der Mittelwert einer diskreten Verteilung berechnet sich nach der Formel

$$\bar{x} = \frac{1}{N} \sum_{x=1}^{n} x f_x \quad , \qquad (2.3.19)$$

wo $N = \Sigma f_x$ ist.

So ist z.B. für die Goethe-Daten in Tabelle 2.7

$$\bar{x} = [1(218) + 2(99) + 3(21) + 4(4)]/342$$
$$= 495/342 = 1.4474.$$

Der Mittelwert ist ein Lokationsmaß, das die Lage der Verteilung auf der x-Achse angibt.

Die Varianz

Dieses Maß gibt die Streuung der Verteilung um den Mittelwert herum an. Man berechnet die Varianz nach der Formel

$$s^2 = \frac{1}{N} \sum_{x=1}^{n} (x - \bar{x})^2 f_x . \qquad (2.3.20)$$

Für rechnerische Zwecke wertet man das Binom aus und erhält

$$s^2 = \frac{1}{N}\left[\sum_{x=1}^{n} x^2 f_x - \frac{(\Sigma x f_x)^2}{N} \right] . \qquad (2.3.21)$$

So ergibt sich für die Goethe-Daten in Tabelle 2.7

$\Sigma x f_x = 495$, wie oben schon berechnet,

$\Sigma x^2 f_x = 1(218) + 2^2(99) + 3^2(21) + 4^2(4) = 867.$

Daraus ergibt sich

$$S^2 = \frac{1}{342}\left(867 - \frac{495^2}{342}\right) = 0.4402 .$$

Für Testzwecke benutzt man üblicherweise die sogenannte erwartungstreue Schätzung der Varianz

$$s^2 = \frac{1}{N-1}\left[\Sigma x^2 f_x - \frac{(\Sigma x f_x)^2}{N}\right] , \qquad (2.3.22)$$

was bei großem N wenig Unterschied ausmacht. In unserem Beispiel wäre $S^2 = 0.4415$.

Zur Charakterisierung der Texte benutzt man meistens die sogenannte ***Standardabweichung***, die die Wurzel aus S^2 oder s^2 darstellt. In unserem Beispiel hätten wir $\sqrt{S^2} = S = 0.66$, $\sqrt{s^2} = s = 0.66$, d.h. kein Unterschied auf zwei Dezimalstellen.

Momente

Bevor wir zwei weitere Maße einführen, definieren wir noch die Momente einer Verteilung.

Die ***Anfangsmomente*** werden definiert als

$$m_r' = \frac{1}{N}\sum_{x=1}^{n} x^r f_r \qquad (2.3.23)$$

und die ***Zentralmomente*** als

$$m_r = \frac{1}{N}\sum_{x=1}^{n} (x-\bar{x})^r f_x . \qquad (2.3.24)$$

Man sieht sofort, daß $m_1' = \bar{x}$ und daß $m_2 = S_2$. Die Zentralmomente kann man auch mit Hilfe der Anfangsmomente ausdrücken. So sieht man, daß

$$m_2 = S^2 = \frac{1}{N}\left[\Sigma x^2 f_x - \frac{(\Sigma x f_x)^2}{N}\right]$$

$$= \frac{1}{N}\Sigma x^2 f_x - \left(\frac{\Sigma x f_x}{N}\right)^2$$

$$= m_2' - m_1'^2 .$$

Ebenso einfach ergibt sich

$$m_3 = \frac{1}{N}\Sigma (x-\bar{x})^3 f_x$$

$$= \frac{1}{N}\Sigma (x^3 - 3x^2\bar{x} + 3x\bar{x}^2 - \bar{x}^3) f_x$$

$$= \frac{1}{N}\Sigma x^3 f_x - 3\bar{x}\,\frac{1}{N}\Sigma x^2 f_x + 3\bar{x}^2\,\frac{1}{N}\Sigma x f_x - \bar{x}^3\,\frac{1}{N}\Sigma f_x$$

$$= m_3' - 3m_1'm_2' + 3m_1'^2 m_1' - m_1'^3$$

$$= m_3' - 3m_1'm_2' + 2m_2'^3 , \qquad (2.3.25)$$

weil man $\bar{x} = \frac{1}{N}\Sigma x f_x = m_1'$ usw. nach der Formel für die Anfangsmomente einsetzen kann.

Auf die gleiche Weise erhält man

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4 \qquad (2.3.26)$$

Mit Hilfe der Zentralmomente definieren wir

die ***Schiefe*** oder ***Asymmetrie***

als

$$b_1 = \frac{m_3}{S^3} = \frac{m_3}{m_2^{3/2}} \qquad (2.3.27)$$

und

die ***Wölbung, Steilheit*** oder ***Exzess***

als

$$b_2 = \frac{m_4}{S^4} - 3 = \frac{m_4}{m_2^2} - 3. \qquad (2.3.28)$$

Beispiel. Illustrieren wir die Berechnung an den Goethe-Daten aus Tabelle 2.7. Es ergeben sich die folgenden Summen

$$\Sigma x f_x = 495 \quad \text{wie oben}$$
$$\Sigma x^2 f_x = 876 \quad \text{wie oben}$$
$$\Sigma x^3 f_x = 1(218) + 2^3(99) + 3^3(21) + 4^3(4) = 1833$$
$$\Sigma x^4 f_x = 1(218) + 2^4(99) + 3^4(21) + 4^4(4) = 4527.$$

Daraus folgen die Anfangsmomente

$$m_1' = 495/342 = 1.4474 \quad \text{wie oben}$$
$$m_2' = 867/342 = 2.5351$$
$$m_3' = 1833/342 = 5.3596$$
$$m_4' = 4527/342 = 13.2368$$

und aus diesen die Zentralmomente

$$m_2 = 0.4402 \quad \text{wie oben,}$$
$$m_3 = 5.3596 - 3(2.5351)1.4474 + 2(1.4474)^3 = 0.4162$$
$$m_4 = 13.2368 - 4(5.3596)1.4474 + 6(2.5351)1.4474^2 - 3(1.4474)^4$$
$$= 0.9059.$$

Die Schiefe und der Exzess können wir jetzt leicht nach Einsetzung berechnen:

$$b_1 = \frac{0.4162}{0.4402^{3/2}} = 1.42$$

$$b_2 = \frac{0.9059}{0.4402^2} - 3 = 1.67 .$$

Die resultierenden Werte interpretiert man nach dem folgenden Schema:

	= 0	< 0	> 0
b_1 b_2	symmetrisch normal	linksasymmetrisch flach	rechtsasymmetrisch steil

Goethes "Totentanz" ist also leicht rechtsasymmetrisch und steil, was man auch an den f_x – Werten leicht ersehen kann.

In Tabelle 2.8 findet man die vier Charakteristika für die Texte der Tabelle 2.7.

Tabelle 2.8

Verteilungscharakteristika der Texte in Tabelle 2.7

	x	S^2	b_1	b_2
Caesar	2.6181	1.6786	0.4923	–0.4546
Sallust	2.5271	1.1325	0.3619	–0.4972
Goethe	1.4474	0.4402	1.4248	1.6749
Schiller	1.5671	0.5962	1.2834	1.1127

"Optisch" gewinnt man den Eindruck, daß die Charakteristika der lateinischen Autoren "anders" sind als die der deutschen. Ein derartiges subjektives Urteil läßt sich jedoch mit Hilfe statistischer Tests objektivieren.■■■

2.3.4.1. Vergleich zweier Mittelwerte

Für den Vergleich zweier Mittelwerte gibt es zahlreiche Methoden, die das Problem von verschiedenen Seiten angehen. Man kann nicht in jedem Fall gleich verfahren, aber bei großen Stichprobenumfängen, mit denen wir es in der Textanalyse üblicherweise zu tun haben, kann man die einzelnen Tests mit Sicherheit verwenden.

(1) Im Falle, daß man zwei Texte eines Autors, geschrieben in identischem Genre, oder von zwei Autoren mit gleichem Genre in einer Sprache vergleicht, kann man annehmen, daß die ***Varianzen gleich*** (jedoch unbekannt) sind, und benutzt den Test

$$\frac{\bar{x}_1 - \bar{x}_2}{s\left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{1/2}} = t, \qquad (2.3.29)$$

wo

$$s^2 = \frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2} = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

und t eine Studentsche Variable mit $N_1 + N_2 - 2$ Freiheitsgraden darstellt. Wenn das berechnete t größer ist als das theoretische t mit den gegebenen Freiheitsgraden, dann lehnen wir die Hypothese der Gleichheit ab.

Wenn wir auf diese Weise die Texte von Caesar und Sallust vergleichen, so bekommen wir aufgrund der Zahlen in Tabelle 2.7 und 2.8

$$s^2 = \frac{775(1.6786) + 702(1.1325)}{775 + 702 - 2} = 1.4210$$

$$s = 1.1920 \quad \text{und}$$

$$t = \frac{|2.6181 - 2.5271|}{1.1920\left(\frac{1}{775} + \frac{1}{702}\right)^{1/2}} = 1.47$$

Die Zahl der Freiheitsgrade (= 1475) ist hier praktisch unendlich, so daß man auf die Normalverteilung zurückgreifen kann. Die dazugehörige Wahrscheinlichkeit ist $P = 0.14$, so daß man die Gleichheit der Mittelwerte annehmen kann.

(2) Im Falle, daß man die Gleichheit der Varianzen nicht annehmen kann, berechnet man

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^{1/2}} \qquad (2.3.30)$$

und die Freiheitsgrade nach der Methode von Welch als

$$FG = \frac{\left[\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right]^2}{\frac{\left[\frac{s_1^2}{N_1}\right]^2}{N_1 - 1} + \frac{\left[\frac{s_2^2}{N_2}\right]^2}{N_2 - 1}} . \qquad (2.3.31)$$

Zahlreiche andere Methoden findet man bei Sachs (1972).

Wenn wir mit diesem Ansatz die Mittelwerte der Texte von Caesar und Sallust vergleichen, so bekommen wir

$$s_1^2 = \frac{N_1 S_1^2}{N_1 - 1} = \frac{775(1.6786)}{774} = 1.6808,$$

$$s_2^2 = 1.1341$$

$$t = \frac{|2.6181 - 2.5271|}{\left(\frac{1.6808}{775} + \frac{1.1341}{702}\right)^{1/2}} = 1.48.$$

Die Freiheitsgrade ergeben sich aus (2.3.31) als

$$FG = \frac{\left(\frac{1.6808}{775} + \frac{1.1341}{702}\right)^2}{\frac{\left(\frac{1.6808}{775}\right)^2}{774} + \frac{\left(\frac{1.1341}{702}\right)^2}{701}} \approx 1461$$

Beide Resultate sind fast identisch. Testen wir noch den Unterschied zwischen Caesar und Goethe mit der zweiten Methode, so erhalten wir

$$t = \frac{|2.6181 - 1.4474|}{[0.002169 + 0.441491]^{1/2}} = 1.76$$

$$FG = \frac{(0.002169 + 0.441491)^2}{\frac{0.002169^2}{774} + \frac{0.441491^2}{341}} \approx 344.$$

Auch hier kann man eine unendliche Zahl von Freiheitsgraden annehmen, jedoch ist der Unterschied der Mittelwerte auch hier nicht signifikant (P = 0.078).

Die Unterschiede zwischen den anderen Charakteristika (S^2, b_1, b_2) zu testen, ist, wie Grotjahn (1982) zeigt, linguistisch gesehen weniger sinnvoll. Bei so großen Stichproben, wie wir sie in der Textanalyse bekommen, erweist sich auch der kleinste Unterschied zwischen zwei Varianzen als signifikant, wenn wir ihn mit dem klassischen F-Test überprüfen. Auch eine Transformation auf eine Normalvariable hilft nicht, wenn die Streuung der Daten an sich zu klein ist, wie es in unseren Beispielen der Fall ist. Es wäre eventuell zweckmäßiger, die obigen Charakteristika als Elemente eines Vektors zu betrachten und andere geeignete Vergleichsmethoden zu benutzen (Taxonomie, Diskriminanzanalyse usw.)

2.3.4.2. Vergleich zweier Verteilungen

Anstatt die Differenzen einzelner Charakteristika zu testen, kann man ganze Verteilungen miteinander vergleichen. Man kann nämlich untersuchen, ob die Aufteilung der Häufigkeiten auf die einzelnen Klassen in den beiden empirischen Verteilungen homogen ist, d.h., man vergleicht die Häufigkeiten in den parallelen Klassen zweier Verteilungen. An dieser Stelle werden wir zwei einfache, gleichwertige Homogenitätstests zeigen und an dem Vergleich der Daten von Goethe und Schiller aus der Tabelle 2.7 illustrieren. Die notwendigen Zahlen sind in Tabelle 2.8 enthalten. Hier bezeichnen wir die Häufigkeiten in der Tabelle als f_{ij}, wobei i = 1,2,...,n (n = 5) und j = 1,2, (oder 1 = Goethe, 2 = Schiller). Die Randsumme rechts bezeichnen wir als $f_{i.}$; so ist $f_{1.}$ = 798, $f_{2.}$ = 395 usw. Die Randsumme unter der Tabelle ist $f_{.j}$; so ist $f_{.1}$ = f(Goethe) = 342, $f_{.2}$ = f(Schiller) = 998. Die Gesamtsumme bezeichnen wir als N.

Den Chiquadrat-Test für Homogenität führen wir nach der Formel

$$X^2 = \sum_{j=1}^{2} \sum_{i=1}^{n} \frac{(f_{ij} - f_{i.} f_{.j}/N)^2}{f_{i.} f_{.j}/N} \qquad (2.3.37)$$

durch. Für rechnerische Zwecke läßt sich diese Formel auf

$$X^2 = \frac{N^2}{f_{.1} f_{.2}} \sum_{i=1}^{n} \frac{f_{i1}^2}{f_{i.}} - \frac{N f_{.1}}{f_{.2}} \qquad (2.3.38)$$

Tabelle 2.8

Daten für den Homogenitätstest

Goethe f_{i1}	Schiller f_{i2}	$f_{i.}$
218	580	798
99	296	395
21	97	118
4	24	28
-	1	1
342 = $f_{.1}$	998 = $f_{.2}$	1340 = N

umformen, und die resultierende Größe ist wie ein Chiquadrat mit n-1 Freiheitsgraden verteilt.

Setzen wir die Werte aus Tabelle 2.8 ein, so erhalten wir

$$X^2 = \frac{1340^2}{342(998)} \left(\frac{218^2}{798} + \frac{99^2}{395} + \frac{21^2}{118} + \frac{4^2}{28} + \frac{0^2}{1} \right) - \frac{1340(342)}{998}$$

$$= 466.5040 - 459.1984$$

$$= 7.31 .$$

Ein X^2 mit 4 Freiheitsgraden auf $\alpha = 0.05$ hat den Wert 9.49. Da unser berechneter Wert kleiner ist, schließen wir daraus, daß die beiden Verteilungen homogen sind.

Führt man jedoch den Test für Caesar und Sallust durch, so erhält man $X^2 = 33.30$, was mit 6 Freiheitsgraden einen sehr hoch signifikanten Unterschied andeutet ($P = 9 \times 10^{-6}$). Dies zeigt, daß bei Werken im gleichen Genre nicht alle Eigenschaften homogen sein müssen.

Eine andere Art, diesen Test durchzuführen, ergibt sich mit Hilfe der Informationsstatistik nach der Formel

Auch hier kann man eine unendliche Zahl von Freiheitsgraden annehmen, jedoch ist der Unterschied der Mittelwerte auch hier nicht signifikant (P = 0.078).

Die Unterschiede zwischen den anderen Charakteristika (S^2, b_1, b_2) zu testen, ist, wie Grotjahn (1982) zeigt, linguistisch gesehen weniger sinnvoll. Bei so großen Stichproben, wie wir sie in der Textanalyse bekommen, erweist sich auch der kleinste Unterschied zwischen zwei Varianzen als signifikant, wenn wir ihn mit dem klassischen F-Test überprüfen. Auch eine Transformation auf eine Normalvariable hilft nicht, wenn die Streuung der Daten an sich zu klein ist, wie es in unseren Beispielen der Fall ist. Es wäre eventuell zweckmäßiger, die obigen Charakteristika als Elemente eines Vektors zu betrachten und andere geeignete Vergleichsmethoden zu benutzen (Taxonomie, Diskriminanzanalyse usw.)

2.3.4.2. Vergleich zweier Verteilungen

Anstatt die Differenzen einzelner Charakteristika zu testen, kann man ganze Verteilungen miteinander vergleichen. Man kann nämlich untersuchen, ob die Aufteilung der Häufigkeiten auf die einzelnen Klassen in den beiden empirischen Verteilungen homogen ist, d.h., man vergleicht die Häufigkeiten in den parallelen Klassen zweier Verteilungen. An dieser Stelle werden wir zwei einfache, gleichwertige Homogenitätstests zeigen und an dem Vergleich der Daten von Goethe und Schiller aus der Tabelle 2.7 illustrieren. Die notwendigen Zahlen sind in Tabelle 2.8 enthalten. Hier bezeichnen wir die Häufigkeiten in der Tabelle als f_{ij}, wobei i = 1,2,...,n (n = 5) und j = 1,2, (oder 1 = Goethe, 2 = Schiller). Die Randsumme rechts bezeichnen wir als $f_{i.}$; so ist $f_{1.}$ = 798, $f_{2.}$ = 395 usw. Die Randsumme unter der Tabelle ist $f_{.j}$; so ist $f_{.1}$ = f(Goethe) = 342, $f_{.2}$ = f(Schiller) = 998. Die Gesamtsumme bezeichnen wir als N.

Den Chiquadrat-Test für Homogenität führen wir nach der Formel

$$X^2 = \sum_{j=1}^{2} \sum_{i=1}^{n} \frac{(f_{ij} - f_{i.} f_{.j}/N)^2}{f_{i.} f_{.j}/N} \qquad (2.3.37)$$

durch. Für rechnerische Zwecke läßt sich diese Formel auf

$$X^2 = \frac{N^2}{f_{.1} f_{.2}} \sum_{i=1}^{n} \frac{f_{i1}^2}{f_{i.}} - \frac{N f_{.1}}{f_{.2}} \qquad (2.3.38)$$

Tabelle 2.8

Daten für den Homogenitätstest

Goethe f_{i1}	Schiller f_{i2}	$f_{i.}$
218	580	798
99	296	395
21	97	118
4	24	28
-	1	1
342 = $f_{.1}$	998 = $f_{.2}$	1340 = N

umformen, und die resultierende Größe ist wie ein Chiquadrat mit n-1 Freiheitsgraden verteilt.

Setzen wir die Werte aus Tabelle 2.8 ein, so erhalten wir

$$X^2 = \frac{1340^2}{342(998)} \left(\frac{218^2}{798} + \frac{99^2}{395} + \frac{21^2}{118} + \frac{4^2}{28} + \frac{0^2}{1} \right) - \frac{1340(342)}{998}$$

$$= 466.5040 - 459.1984$$

$$= 7.31 .$$

Ein X^2 mit 4 Freiheitsgraden auf $\alpha = 0.05$ hat den Wert 9.49. Da unser berechneter Wert kleiner ist, schließen wir daraus, daß die beiden Verteilungen homogen sind.

Führt man jedoch den Test für Caesar und Sallust durch, so erhält man $X^2 = 33.30$, was mit 6 Freiheitsgraden einen sehr hoch signifikanten Unterschied andeutet ($P = 9 \times 10^{-6}$). Dies zeigt, daß bei Werken im gleichen Genre nicht alle Eigenschaften homogen sein müssen.

Eine andere Art, diesen Test durchzuführen, ergibt sich mit Hilfe der Informationsstatistik nach der Formel

$$2I = 2 \sum_i \sum_j f_{ij} \ln \frac{Nf_{ij}}{f_{i.} f_{.j}}, \qquad (2.3.39)$$

die man für Rechenzwecke einfacher als

$$2I = 2\sum_{ij} f_{ij} \ln f_{ij} + 2N \ln N - 2\sum_i f_{i.} \ln f_{i.} - 2\sum_j f_{.j} \ln f_{.j} \qquad (2.3.40)$$

darstellt.

Entwickelt man diese Formel in eine Taylorreihe, so ergibt sich ungefähr (2.3.37).

Für die Daten in Tabelle 2.8 erhalten wir

$$2\sum_i \sum_j f_{ij} \ln f_{ij} = 2(218 \ln 218 + 580 \ln 580 + 99 \ln 99 + \ldots + 24 \ln 24 + 1 \ln 1)$$
$$= 15186.27979$$

$$2N \ln N = 2(1340) \ln 1340 = 19279.13871$$

$$2 \sum_i f_{i.} \ln f_{i.} = 2(798 \ln 798 + 395 \ln 395 + \ldots + 1 \ln 1)$$
$$= 16700.4501$$

$$2 \sum_j f_{.j} \ln f_{.j} = 2(342 \ln 342 + 998 \ln 998) = 17774.89408.$$

Daher ist

$$2I = 15186.27979 + 19297.13871 - 16700.4501 - 17774.89408$$
$$= 8.07$$

Wie man sieht, unterscheidet sich 2I etwas von X^2. Nach dem Vorschlag von Ku (1963) pflegt man 2I für jedes leere Feld um 1 zu verringern. Da bei Goethe die Klasse i = 5 leer ist, korrigieren wir 2I auf

$$2I_{kor} = 8.07 - 1 = 7.07,$$

wodurch 2I dem X^2 etwas näher rückt.

2.4. Modellierung von Wahrscheinlichkeitsverteilungen

Die Charakterisierung eines Textes durch Indizes (Kenngrößen, Maßzahlen) und der Vergleich zweier Indizes sind induktive Verfahren, mit deren Hilfe wir versuchen, die Erscheinungsformen des Textes zu erfassen. Wir können dadurch ein phänomenologisches Bild des Textes aufstellen, oft auch seine Oberflächenstruktur rekonstruieren. Zählen (messen) wir mehrere unterschiedliche Variablen (Texteinheiten), dann können wir mit Hilfe der Faktoranalyse oder anderer multivariater Verfahren die Zusammenhänge zwischen den Variablen erforschen (vgl. Carroll 1960; Sommers 1962), die zur Anregung für die Aufstellung von Hypothesen von Nutzen sein können, aus denen sich später eine Theorie der Texte aufbauen ließe.

Jedoch auch dann, wenn wir die "Phänomenologie" der Texte erforscht und eine umfangreiche Beschreibung der Eigenschaften und ihrer Korrelationen untereinander erhalten haben, sind wir von einer Theorie noch weit entfernt, da wir nicht wissen, welche Mechanismen der Erzeugung der vorhandenen Konfigurationen von Eigenschaften zugrundeliegen. Das heißt, wir haben noch immer keine Gesetze, mit deren Hilfe wir Erklärungen leisten können.

Der Weg zur Entdeckung und Formulierung von Gesetzen ist beschwerlich und fängt meistens mit der "phänomenologischen" Erfassung einer Regularität an. Es gibt hier drei Möglichkeiten:

(i) Einem Verlauf von Werten werden tentativ mehrere Kurven angepaßt, und die "beste", d.h. beispielsweise diejenige, die die kleinste Summe der Abweichungsquadrate aufweist, wird gewählt. Dieses induktive Verfahren hat sozusagen nur lokale Bedeutung, da das Resultat wenig Chancen hat, in eine Theorie eingegliedert zu werden, es sei denn, man hat eine glückliche Hand gehabt. Von dieser Art ist z.B. eine von Piotrovskaja, Piotrovskij (1974) vorgeschlagene Kurve für die Erfassung der Entwicklung von Sprachentitäten, die später theoretisiert wurde (vgl. Altmann, v.Buttlar, Rott, Strauss 1983). Kurven dieser Art können aber schon vom Anfang an der Beschreibung gut dienen.

(ii) Man hat eine Kurve, die sich in anderen Bereichen gut bewährt hat, oder die allgemein genug ist, und versucht, sie auch im betreffenden Bereich anzuwenden. Falls sie im neuen Bereich gut interpretierbar ist, dann hat diese Kurve bessere Überlebenschancen. Von dieser Art ist das Zipf-Mandelbrotsche Gesetz, das aus der Linguistik in die Musikologie und in die Theorie der bildenden Künste übertragen worden ist (vgl. Orlov, Boroda, Nadarejšvili 1982).

(iii) Ideal ist natürlich eine "repräsentationale" (vgl. Bunge 1967: 248) Erfassung der Regularität, bei der man von einem Hintergrundwissen ausgeht, den Erzeugungsmechanismus in Betracht zieht und z.B. eine Kurve ableitet, die mit den Daten gut verträglich ist. Läßt sich eine solche Kurve in ein System von derartigen Aussagen einordnen, dann kann man sie als Gesetz bezeichnen (Bunge 1967). Dieser Weg ist in der Textanalyse äußerst schwierig, weil man hier erst wenige Gesetze kennt und weil hier zahlreiche Wege für die Erforschung einer einzigen Erscheinung offenstehen.

Fraglich ist auch, ob eine einzige Kurve für alle Texteinheiten zuständig ist; dies ist kaum anzunehmen, obwohl es sehr allgemeine Verteilungen gibt, die mehrere Erscheinungen erfassen, z.B. die Sichel-Verteilung (vgl. Sichel 1971, 1974, 1975). Solche Verteilungen sind aber besonders schwer zu interpretieren. Wir nehmen an, daß sogar die Häufigkeitsverteilung einer einzigen Texteinheit durch mehrere Wahrscheinlichkeitsverteilungen modelliert werden muß, weil in unterschiedlichen Genres zusätzliche "Kräfte" wirken können, die sich durch bloße Unterschiede in Parametern nicht adäquat erfassen lassen.

Wir werden hier nur eine Texteinheit in Betracht ziehen, nämlich die Wortlänge (in Silbenzahl), und versuchen, einen Modellierungsweg zu zeigen, der von linguistischen Überlegungen ausgeht, für andere Texteinheiten modifizierbar und ohne komplizierte Mathematik leicht erweiterbar ist.

Die Verteilung der Wortlängen ohne Bezug auf andere Worteigenschaften wurde von Fucks (1955) und Čebanov (1974) erforscht, die zu einer verschobenen Poisson-Verteilung kamen. Häufige Unstimmigkeiten mit den Daten führten Grotjahn (1982) dazu, den Parameter der Poisson-Verteilung als eine Gamma-Variable zu betrachten, wodurch er zu einer negativen Binomialverteilung kam. Diese Möglichkeit hat bereits Fucks (1970) erwogen, und sie erbrachte eine bessere Übereinstimmung mit den Daten. Wir werden zeigen, wie man auf einem anderen Weg zu dem gleichen Resultat kommt, und werden es noch etwas verallgemeinern.

Orlov hat an mehreren Stellen gezeigt, daß das Zipf-Mandelbrotsche-Gesetz nur für Texte gilt, nicht aber für Teile des Textes. Die gleiche Erkenntnis gewann Boroda aus musikalischen Texten (vgl. Orlov, Boroda, Nadarejšvili 1982). Orlov schloß daraus, daß der Verfasser des Textes eine geplante Länge des Textes im Sinne hat und den Informationsfluß auf diese Gesamtlänge zerlegt. Dadurch erreicht er eine reguläre Rang-häufigkeitsverteilung der Wörter, die dem Zipf-Mandelbrotschen-Gesetz folgt. Diese theoretische Textlänge bezeichnete Orlov als "Zipfsche Zahl", aber aufgrund der von ihm vorgelegten Begründung verdient sie eher den Namen "Zipf-Orlovsche Zahl" oder "Zipf-Orlovsche Länge".

Aus dieser Erkenntnis folgt, daß ein Autor aufgrund seiner Zipf-Orlovschen Länge für den gegebenen Text im Grunde nicht die Worthäufigkeiten selbst, sondern die Abstände, die Differenzen zwischen den rangbenachbarten Häufigkeiten unbewußt so steuert, daß sie dem Zipf-Mandelbrotschen Gesetz folgen. Daraus folgt aber weiter, daß wir uns bei der Modellierung von Verteilungen gerade auf die Gestaltung dieser Differenz $P_x - P_{x-1} = \triangle P_{x-1}$ konzentrieren können, um die Wahrscheinlichkeitsverteilung abzuleiten.

Sei hier also P_x die Wahrscheinlichkeit, daß die Zufallsvariable, z.B. die Wortlänge, den Wert x annimmt, und $\triangle$ der Differenzoperator, wie oben definiert. Wir werden im weiteren immer die relative Differenz $D = \triangle P_{x-1}/P_{x-1}$ in Betracht ziehen und überlegen, wie sie gestaltet werden kann.

Zu diesem Zweck gehen wir davon aus, daß jeder Text, jede Texteinheit, von den Zipfschen Kräften, die in der Sprache durchgängig wirken, gestaltet wird. Es sind die Kräfte der Unifikation und der Diversifikation, die je nach den Umständen von Sprecher oder Hörer ausgehen (vgl. Zipf 1949). So will der Sprecher beispielsweise die Bedeutungen eines Wortes diversifizieren, damit er mit einem Wort möglichst viel ausdrücken kann; der Hörer möchte aber, daß jedes Wort nur eine Bedeutung hat, damit er mit der Dekodierung der Nachricht weniger Mühe hat. Der Sprecher will kurze Wörter haben, um Kodierungsanstrengung zu sparen, der Hörer wünscht lange Wörter, um Dekodierungsanstrengung zu sparen usw. usw. Sie pendeln sich in allen Bereichen der Sprache in ein Fließgleichgewicht ein.

In geschriebenen Texten kann der Autor auf D unterschiedlichen Einfluß nehmen, je nachdem, wie die gegebene Eigenschaft (hier Wortlänge) in der Sprache überhaupt gestaltet ist (es gibt z.B. Sprachen, in denen öfter zweisilbige als einsilbige Wörter benutzt werden), welche Beschränkungen die Textart ihm auferlegt (z.B. Metrum), welche Art der ästhetischen Wirkung er beabsichtigt (z.B. Euphonie), welche Rücksicht er

auf den Leser nimmt (z.B. Informationsfluß im Gedicht vs. im wissenschaftlichen Text), d.h., auf welche Art der Hörer (Leser) an der Gestaltung des Textes teilnimmt, usw. .

Veranschaulichen wir diese Überlegung an einem einfachen Beispiel. Bezeichnen wir die "Kraft", den "Anteil", den "Einfluß", die "Proportionalität" o.ä. des Sprechers mit S und die des Hörers mit H, wobei beide reelle Zahlen darstellen können. Wirken beide Kräfte additiv, konstant und negativ auf D, dann bekommen wir

$$\frac{\triangle P_{x-1}}{P_{x-1}} = -(H + S) \quad . \qquad (2.4.1)$$

Bezeichnet man $H + S = A$ und löst (2.4.1) für P_x auf, dann erhält man

$$P_x - P_{x-1} = -AP_{x-1}$$

$$P_x = (1 - A)P_{x-1} \, .$$

Die Lösung ist dann

$$P_x = (1 - A)^x P_0 \, .$$

Wenn $0 < A < 1$, dann schreiben wir $A = p$, $1 - A = q$, d.h. $P_x = P_0 q^x$. Wegen $\Sigma P_x = 1$ bekommen wir

$$1 = P_0 \sum_{x=0}^{\infty} q^x = P_0 \frac{1}{1-q} = P_0 \frac{1}{p} \, ,$$

woraus $P_0 = p$ folgt, so daß wir aus diesem Ansatz die geometrische Verteilung

$$P_x = pq^x \, , \quad x = 0,1,... \qquad (2.4.2)$$

erhalten. Andere Ansatzmöglichkeiten für dasselbe Resultat wären $D = -H/S$, $D = -S/H$, $D = H-S$ mit $S>H$ usw., mit der Bedingung, daß die Konstante auf der rechten Seite kleiner als 1 ist. Ist sie größer als 1, dann bekommen wir eine oszillierende Kurve. Ist D positiv, dann muß x kleiner als eine endliche Zahl sein, sonst divergiert die Reihe.

Subtrahiert man von der Konstante A die proportionale Wirkung des Hörers (bzw. des Sprechers) und dividiert durch dieselbe Größe, d.h.

$$D = \frac{A - Hx}{Hx} \qquad \text{oder} \qquad D = \frac{A - Sx}{Sx} \, , \qquad (2.4.3)$$

dann erhält man als Lösung die Poissonverteilung mit dem Parameter A/S, die Fucks als Wortlängenverteilung auf eine andere Art abgeleitet hat.

Diese Technik bietet erstens eine bessere linguistische Interpretationsmöglichkeit, zweitens einen Anschluß an die Systeme von Katz (1965) und Ord (1967), deren Verteilungen mit den hier ableitbaren teilweise übereinstimmen, und drittens eine Möglichkeit, neue Verteilungen abzuleiten oder bekannte zu modifizieren.

Wir sind davon überzeugt, daß man eine Eigenschaft für alle Texte aller Sprachen nicht mit einer einzigen Verteilung modellieren kann, es sei denn, die Verteilung ist so allgemein, daß sie für Daten aller Art paßt. In dem Falle wäre sie aber völlig immun, unfalsifizierbar und dadurch wissenschaftlich nicht fruchtbar.

Weiter besteht der Verdacht, daß eine Verteilung zwar für Wortlängenverteilungen kurzer Texte ausreichen, aber bei längeren Texten versagen kann. Dies kann dadurch verursacht werden, daß längere Texte nicht auf einmal, sondern mit zeitlichen Pausen geschrieben werden, wobei der "Wortlängenrhythmus" des Autors sich ändern kann. Daher wären bei längeren Texten eher gemischte oder zusammengesetzte Verteilungen geeignet. Dies gilt natürlich auch für Texte, an denen viel korrigiert wurde, von denen es mehrere Varianten gibt, die von mehreren Autoren stammen usw.

Im weiteren werden wir den obigen Ansatz dazu benutzen, um die Wortlängenverteilung zu erhalten und um ein Modell für die Satzlängenverteilung zu bekommen.

(1) Wir nehmen an, daß die relative Differenz D bei Wortlängen aus der Summe einer linearen Funktion des Höreranteils und eines Textsortenparameters besteht, die mit einem inversen linearen Anteil des Sprechers (ohne Konstante) multipliziert wird, d.h.

$$\frac{\triangle P_{x-1}}{P_{x-1}} = \frac{A - Hx}{Sx} \, . \qquad (2.4.4)$$

Wenn wir nach P_x auflösen, bekommen wir

$$P_x = (1 + \frac{A - Hx}{Sx}) P_{x-1}$$

$$= \frac{Sx + A - Hx}{Sx} P_{x-1}$$

$$= \frac{A + (S - H)x}{Sx} P_{x-1} \quad . \qquad (2.4.5)$$

Klammert man $S - H$ aus und bezeichnet $K = A/(S - H) + 1$, dann wird (2.4.5) zu

$$P_x = \frac{(S - H)}{S} \cdot \frac{K + x - 1}{x} P_{x-1}$$

$$= q(\frac{K + x - 1}{x}) P_{x-1} \quad , \qquad (2.4.6)$$

wenn man $(S - H)/S = q$ setzt. Hier ist $0<q<1$, weil wir vorausgesetzt haben, daß $S > H$. Gleichung (2.4.6) kann man schrittweise lösen:

$$P_1 = q \frac{K}{1} P_0$$

$$P_2 = q \frac{K + 1}{2} P_1 = \frac{K}{1} \cdot \frac{K + 1}{2} q^2 P_0$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$P_x = \frac{K(K + 1)(K + 2)\ldots(K + x - 1)}{x!} q^x P_0 \quad ,$$

was man als

$$P_x = (\begin{matrix} K + x - 1 \\ x \end{matrix}) q^x P_0 \qquad (2.4.7)$$

schreiben kann. Aus $\Sigma P_x = 1$ ergibt sich

$$1 = P_0 \sum_{x=0}^{\infty} (\begin{matrix} K + x - 1 \\ x \end{matrix}) q^x = P_0 \sum_{x=0}^{\infty} (\begin{matrix} -K \\ x \end{matrix}) (-q)^x$$

$$= P_0 (1 - q)^{-K} .$$

Setzt man $p = 1-q$, dann bekommt man $P_0 = p^K$. Daraus ergibt sich

$$P_x = (\begin{matrix} K + x - 1 \\ x \end{matrix}) p^K q^x \quad , \qquad x = 0,1,\ldots \qquad (2.4.8)$$

die ***negative Binomialverteilung***, die Grotjahn (1982) auf eine andere Weise für die Wortlängenverteilung bekommen hat. Die Konstanten hier sind nicht nur irgendwelche Funktionen der Sprecher-Hörer-Interaktion, sondern auch Textparameter, jedoch kann man nur durch eine sehr breite empirische Untersuchung ihre numerischen Werte ermitteln. So ist $q = (S-H)/S$, $p = 1-q = 1 - (S-H)/S = H/S$, A ist eine der oben angedeuteten Interaktionen, z.B. H/S, und $K = A/(S-H)+1$. Möglicherweise setzt sich A aus mehreren interpretierbaren Größen zusammen.

In Tabelle 2.9 findet man die Anpasssung der negativen Binomialverteilung an die von Grotjahn ausgewerteten Briefe von Goethe. Wie man sieht, sind alle Anpassungen sehr gut. Es ist jedoch keineswegs zwingend, daß die ermittelten Parameter K und p für Goethes Briefe oder für Goethe charakteristisch sind. Sie sind nur das Resultat einer iterativen Anpassung, wobei keineswegs gesichert ist, daß wir das globale Minimum gefunden haben. Andere Anfangswerte der Optimierung können eventuell zu noch besseren Anpassungen führen. So erhält man z.B. für den Brief Nr. 596 mit den Parametern $K = 59.1903$, $p = 0.9871$ ein $X^2 = 1.07$, dem mit 2 FG die Wahrscheinlichkeit $P = 0.5857$ entspricht.

(2) Will man die Satzlängenverteilung in einem Text modellieren, so kann man den gleichen Ansatz benutzen. Mißt man jedoch die Satzlänge nicht in der Anzahl der Clauses, sondern in der Anzahl der Wörter, dann muß man die Störung, den Faktor der intervenierenden Ebene in Betracht ziehen. Bezeichnet man diesen Faktor als B, so kann man die Gleichung als

$$\frac{\triangle P_{x-1}}{P_{x-1}} = \frac{A - Hx}{Sx + B} \qquad (2.4.9)$$

Tabelle 2.9

Anpassung der negativen Binomialverteilung an einige Briefe von Goethe

	Nr. 612		Nr. 647		Nr. 659		Nr. 667	
x	fx	NPx	fx	NPx	fx	NPx	fx	NPx
1	164	162.61	259	259.16	151	150.91	77	76.15
2	105	104.38	132	125.65	68	65.64	51	53.37
3	35	38.81	37	46.65	16	19.81	26	24.56
4	15	10.94	19	15.55	7	5.10	10	9.33
5	1	3.26	6	4.89	1	1.54	4	4.59
6	-	-	1	2.10	-	-	-	-
k		6.3120		1.8819		2.5790		3.1954
p		0.8983		0.7424		0.8314		0.7807
X^2		3.47		3.91		1.71		0.32
FG		2		3		2		2
P		0.1764		0.2713		0.4253		0.8521

ansetzen. Daraus folgt

$$P_x = (1 + \frac{A - Hx}{Sx + B})P_{x-1}$$

$$= \frac{Sx + A + B - Hx}{Sx + B} P_{x-1}$$

$$= \frac{(A + B) + (S - H)x}{B + Sx} P_{x-1}.$$

Klammern wir im Zähler (S - H) und im Nenner S aus, dann erhalten wir

$$P_x = \frac{S - H}{S} \cdot \frac{\frac{A + B}{S - H} + x}{\frac{B}{S} + x} P_{x-1}. \qquad (2.4.10)$$

Bezeichnen wir nun

$$\frac{S - H}{S} = q, \qquad \frac{A + B}{S - H} = K - 1, \qquad \frac{B}{S} = R - 1,$$

dann wird (2.4.10) zu

$$P_x = \frac{K + x - 1}{R + x - 1} qP_{x-1}, \qquad (2.4.11)$$

und die rekursive Lösung lautet

$$P_x = \frac{K(K + 1)(K + 2)\ldots(K + x - 1)}{R(R + 1)(R + 2)\ldots(R + x - 1)} q^x P_0, \qquad (2.4.12)$$

ein Resultat, daß sich auf verschiedene Weisen schreiben läßt. P_0 ergibt sich wieder aus der Bedingung, daß $\Sigma P_x = 1$, und läßt sich mit Hilfe der hypergeometrischen Funktion als

$$P_0 = \frac{1}{{}_2F_1(K, 1; R; q)} \qquad (2.4.13)$$

symbolisieren. Die Verteilung (2.4.12) ist ein Spezialfall des Ordschen Systems und stellt eine ***Hyperpascal-Verteilung*** dar. In der Tabelle 2.10 findet man einige Anpassungen der Hyperpascal-Verteilung an die Satzlängen in Texten von Herodot, wie sie von Morton (1965) bzw. Morton, Levison (1966) ermittelt wurden.

Gute Anpassungen lassen sich auch mit anderen Parametern erreichen, möglicherweise noch bessere. Für unsere Demonstration sind die Resultate sehr befriedigend.

Mit dieser Technik läßt sich beliebig fortfahren, jedoch soll man sich möglichst an die folgenden unverbindlichen Regeln halten.

(a) Man wähle für den gegebenen Fall die einfachste Verteilung, d.h., eine mit minimaler Anzahl von Parametern. Das Prinzip der Einfachheit widerspricht zwar den meisten Kriterien der Wissenschaft (vgl. Bunge 1961), aber ein einfaches Modell gibt einen besseren Ausgangspunkt als ein kompliziertes. Wenn man mit einem einfachen Modell auskommt, das ein Spezialfall eines komplizierten ist, soll man es beibehalten. So kann man beispielsweise die Wortlängenverteilungen zuerst mit

Tabelle 2.10

Anpassung der Hyperpascal-Verteilung an die Satzlängen von Herodot

(nach Daten von Morton 1965)

	Buch							
	Nr 1		Nr. 2		Nr. 3		Nr. 4	
x	fx	NPx	fx	NPx	fx	NPx	fx	NPx
0	16	16.08	6	4.75	7	6.87	8	8.07
1	35	36.19	33	30.81	48	45.44	38	37.05
2	47	43.05	34	38.99	37	40.54	49	45.05
3	39	37.82	39	35.90	34	31.62	38	38.71
4	24	27.62	26	28.56	21	23.44	24	27.99
5	16	17.77	22	20.88	23	16.92	15	18.23
6	11	10.42	19	14.43	6	12.02	15	11.06
7	5	5.69	5	9.59	11	8.44	8	6.38
8	4	2.93	4	6.19	6	5.88	2	3.54
9	1	1.44	6	3.91	2	4.08	0	1.90
10	1	0.68	1	2.42	3	2.81	1	1.00
11	1	0.31	0	1.48	2	1.93	0	0.51
12	-	-	0	0.89	-	-	1	0.26
13			0	0.53			0	0.13
14			1	0.31			0	0.06
15			2	0.18			0	0.03
16			1	0.11			0	0.02
17			1	0.06			0	0.01
18			-	-			1	0.00
K	5.9987		1.7383		0.3997		2.5904	
R	0.8277		0.1414		0.0401		0.2360	
q	0.3106		0.5275		0.6630		0.4185	
X^2	1.72		5.33		7.94		6.18	
FG	6		6		8		7	
P	0.94		0.50		0.44		0.52	

der geometrischen Verteilung modellieren, falls die Häufigkeiten monoton abnehmen, denn die geometrische Verteilung ist ein Spezialfall der negativen Binomialverteilung, wenn K = 1. Ist kein monotoner Verlauf vorhanden, dann soll man die Poissonverteilung anwenden, da diese einen Grenzfall der negativen Binomialverteilung darstellt (wenn K -> ∞, q -> 0, Kq = a). Passen die beiden Verteilungen nicht, dann soll man die negative Binomialverteilung verwenden. Die negative Binomialverteiung läßt

sich auf sehr viele verschiedene Arten verallgemeinern (vgl. z.B. Patil, Yoshi 1968), eine der Möglichkeiten haben wir oben gezeigt. Beim Testen mit dem Chiquadrattest kann es jedoch vorkommen, daß bei mehreren Parametern keine Freiheitsgrade übrigbleiben.

(b) Überprüft man ein derartiges "Textgesetz" an einer ganzen Sprache, d.h. an einer großen gemischten Stichprobe, dann muß man in Betracht ziehen, daß die Parameter der Grundverteilung (hier die der negativen Binomialverteilung oder der Hyperpascal-Verteilung), nicht in allen erhobenen Stichproben gleich sind. In solchen Fällen ist es legitim, eine Mischung von gleichen Verteilungen zu erwägen, d.h.

$$P_x = \sum_{i=1}^{k} \alpha_i f(x,\theta_i) \quad , \qquad (2.4.14)$$

wo $f(x,\theta)$ die Wahrscheinlichkeitsfunktionen der Komponenten sind, α_i reelle Zahlen, so daß $\Sigma\alpha_i = 1$, und θ_i alle Parameter einer Komponente vertritt; oder man betrachtet die Parameter der Grundverteilung als Zufallsvariablen mit eigenen Verteilungen und bildet eine zusammengesetzte Verteilung der Form

$$P_x(\theta) = \int f(x|\theta)g(\theta)d\theta$$

oder

$$P_x(\theta) = \Sigma f(x|\theta)g(\theta)$$

oder

$$P_x(\theta_1,\theta_2) = \iint f(x|\theta_1,\theta_2)g(\theta_1)h(\theta_2)d\theta_1 d\theta_2 \qquad (2.4.15)$$

usw. usw. Auf diese Art kann man immer zu einem numerisch befriedigenden Resultat gelangen, das aber theoretisch nicht immer gut interpretierbar ist.

Ist der Text sehr lang, dann kann man annehmen, daß die unterbewußte Steuerung der Wortlängen (bzw. anderer Eigenschaften) beträchtlich gestört wird. Man erwartet also bei großem N, auch wenn das Modell sonst gut funktioniert, Abweichungen, die die konventionellen Grenzen (0.05, 0.01) stark unterschreiten. Es wäre daher empfehlenswert, die kri-

tischen Bereiche für die Linguistik überhaupt und für einzelne Texteigenschaften separat neu zu überdenken.

(c) Manche Verteilungen erlauben unter bestimmten Bedingungen auch negative Parameter, die zwar in unserem Ansatz linguistisch problemlos interpretierbar sind, die man aber, wenn möglich, meiden soll. Üblicherweise existiert immer eine gleichwertige Anpassung mit possitiven Parametern.

(d) Die klassischen Schätzmethoden wie maximum likelihood, Momentenmethode, Methode der kleinsten Quadrate u.a., versagen oft bei Anpassungen. Daher ist es immer empfehlenswert, die Anpassung durch iterative Methoden zu optimieren (was hier in jedem Fall getan wurde).

(e) Die Ableitung der Verteilungen wurde hier so dargestellt, daß $P_0 \neq 0$. Bei unserer Verteilung der Wortlängen fangen die Verteilungen immer bei $x = 1$ an. Man muß daher in jedem Fall eine formale Verschiebung vornehmen, d.h. $P_x = f(x-1,\Theta)$, für $x = 1,2,....$ Bei der Hyperpascal-Verteilung haben wir die Intervalle 1-5, 6-10, 11-15,... der Variablen "Satzlänge" so transformiert, daß wir eine Skala 0,1,2... erhielten, nämlich $x = (y_0 - 5)/5$, wo y_0 die obere Grenze des Intervalls war.

Auf diese Art lassen sich wohl alle Verteilungen der Eigenschaften von Spracheinheiten modellieren.

(f) Man soll möglichst vermeiden, laut Orlovs Empfehlung, Grundgesamtheiten (z.B. für die ganze Sprache) zu bilden (vgl. Orlov, Boroda, Nadarejšvili 1982). Gilt ein Gesetz für einen geschlossenen Text, dann braucht es nicht unbedingt für eine imaginäre Grundgesamtheit aller Texte dieser Art zu gelten (vgl. Punkt b). Die Auswirkung der Zusammensetzung von Texten zeigen wir an einem Beispiel. In Tabelle 2.9 wurde die negative Binomialverteilung dem Brief Nr. 647 mit

$k = 1.8819$ $\quad X^2 = 3.91$
$p = 0.7424$ $\quad FG = 3$
$P = 0.27$

und dem Brief Nr. 612 mit

$k = 6.3120$ $\quad X^2 = 3.47$
$p = 0.8983$ $\quad FG = 2$
$P = 0.18$

angepaßt. Es wäre ein Irrtum, anzunehmen, daß eine Zusammensetzung beider Briefe eine Stabilisierung der Verteilung mit sich bringt. Addiert man nämlich die entsprechenden Häufigkeiten, dann bekommt man die Resultate in Tabelle 2.11, die eine schlechtere Anpassung liefern als die beiden Briefe separat.

Tabelle 2.11

Anpassung der negativen Binomialverteilung an die Zusammensetzung der Briefe Nr. 612 und 647 von Goethe.

x	1	2	3	4	5	6
f_x	423	237	72	34	7	1
NP_x	422.36	230.40	84.96	26.32	7.38	2.58

$k = 2.842$ | $p = 0.808$ | $X^2 = 5.39$ | $FG = 3$ | $P = 0.15$

2.5. Einige Textgesetze

Im folgenden werden vier bekannte Modelle vorgestellt, bei denen die Wiederholungen formlos sind. Die beiden ersten beruhen auf Wahrscheinlichkeitsverteilungen, das dritte und das vierte sind Kurvenmodelle. Die Problematik der ersten zwei ist sehr komplex, das dritte ist verhältnismäßig neu und ebenso wie das vierte scheint es eine akzeptable Lösung darzustellen.

2.5.1. Das ZIPF-MANDELBROTSCHE Gesetz

Keine Errungenschaft der Linguistik, einschließlich des Strukturalismus und der generativen Linguistik, hat ein derartig nachhaltiges Echo in den Wissenschaften gefunden, wie das heftig umstrittene Zipf-Mandelbrotsche Gesetz. Man findet es in allen Humanwissenschaften, in Geographie, Bio-

logie, Musikologie, Dokumentationswissenschaft, Mathematik, Ökonomie usw. usw. bis hinauf in die Systemtheorie. Ursprünglich wurde es von Estoup (1916) aufgestellt, eine tiefgreifende Analyse erfuhr es aber erst in zahlreichen Publikationen von G.K.Zipf, der diese Regularität mit der Neigung des Menschen zur geringsten Anstrengung begründete (bes. 1935, 1949). Es folgten zahlreiche weitere Begründungen, Ableitungen und Varianten, die Literatur zählt einige Hunderte von Arbeiten (s. Guiter, Arapov 1982). Die bekanntesten sind die von Mandelbrot (1953, 1954a,b, 1957), Arapov, Efimova, Šrejder (1975a,b), Orlov (s. Orlov, Boroda, Nadarejšvili 1982). Eine einfache Darstellung findet man bei Piotrowski (1984:125 f.), Rapoport (1982), Mandelbrot (1966), Miller, Chomsky (1963) und Brookes (1982).

Die Behauptung des Gesetzes kann man einfach folgendermaßen formulieren: Stellt man die Häufigkeiten einzelner Einheiten eines Textes fest und ordnet sie so, daß das häufigste Element den Rang 1 bekommt, das zweithäufigste den Rang 2 usw., dann folgen die relativen Häufigkeiten (mit denen man die Wahrscheinlichkeiten P_r schätzt) der Verteilung

$$P_r = \frac{K}{(A+r)^B}, \qquad r = 1,2,\ldots,n . \qquad (2.5.1)$$

Hier ist

n der Inventarumfang,
A und B bestimmte Konstanten,
r der Rang ,
K die Normierungsgröße $K = [\Sigma(A+r)^{-B}]^{-1}$,
die von A und B abhängt.

Es ergeben sich sofort mehrere Probleme, die man nicht immer befriedigend lösen kann.

(1) Der Rang ist keine Zufallsvariable, sondern nur eine Hilfsvariable. Mehrere Ansätze klären dieses Problem zufriedenstellend. Wir werden es mit Hilfe des weniger bekannten Ansatzes von Miller (1957) unten zeigen.

(2) Gilt das Gesetz für Stichproben aus einem Text oder nur für ganze Texte? Gilt es nur für Texte oder auch für die Sprache als ganze? Orlov und Boroda (s. Orlov, Boroda, Nadarejsvili 1982) haben gezeigt, daß

es nur für ganze Texte gilt, ganz gleich ob sprachliche oder musikalische. Dies liegt etwa in folgendem begründet: Ein Verfasser "plant" die Textlänge im voraus – die man als Zipf–Orlovsche Länge bezeichnen kann –, und in Abhängigkeit davon steuert er den Informationsfluß im Text, der dann unter anderem zu einer speziellen Ranghäufigkeitsverteilung führt. Nimmt man also nur eine, beispielsweise eine zufällige Stichprobe, dann erhält man ein völlig verzerrtes Bild dieser Verteilung. Dies gibt aber auch auf die zweite Frage Antwort. Mischt man viele Texte zusammen und ermittelt daraus eine Ranghäufigkeitsverteilung, dann erhält man eine Mischung von Verteilungen mit genau so vielen unterschiedlichen Parametern, die nur bei viel Glück dem einfachen Zipf–Mandelbrotschen Gesetz folgen kann. Daraus folgt, daß man nicht versuchen sollte, diese Verteilung Häufigkeitswörterbüchern einer (ganzen) Sprache anzupassen. Für die ganzsprachlichen Häufigkeiten gelten womöglich andere Gesetze.

(3) Für welche Einheiten gilt dieses Gesetz? Für Laute, Silben, Morpheme, Wörter, Wortklassen, Versfüße u.ä.? Dies ist noch nicht geklärt. Wenn es aber für unterschiedliche Einheiten gilt, wie sind jeweils die Parameter? Hier ergibt sich ein ergiebiges Forschungsfeld.

(4) Bei einigen Einheiten ist die Segmentierung und Klassifikation einfach, z.B. bei Phonemen. Aber wie ist es bei Wörtern? Soll man Wortformen oder Lexeme zählen? Was davon folgt dem Zipf–Mandelbrotschen Gesetz? Aus Millers Interpretation (s. unten) folgt, daß es die Wortformen sind, nicht die Lexeme (Lemmata). Das Bedürfnis nach optimalem Informationsfluß, d.h. nach geregelter Zunahme neuer Wörter, bezieht sich aber eher auf Lexeme als auf Wortformen. Das Problem ließe sich sicherlich durch empirische Untersuchungen klären.

(5) Das Zipf–Mandelbrotsche Gesetz wurde sehr oft benutzt, aber sehr selten wurde mit einem Test gezeigt, ob es tatsächlich akzeptierbar ist. Wir zeigen an drei Beispielen die Anpassung an die Daten.

Betrachten wir zunächst die geordneten Häufigkeiten der Phoneme im "Erlkönig", wie sie von Grotjahn (1979:182–3) ausgezählt wurden (s. Tabelle 2.12). Für die Schätzung der Parameter benutzen wir einfach die Optimierungsmethode von Nelder und Mead (1964) und erhalten die Resultate, die in der dritten Spalte von Tabelle 2.12 zu sehen sind. Der Chiquadrat–Test ergibt $X^2 = 30.90$, was mit 36 Freiheitsgraden einem $P = 0.71$ entspricht. Wir können also die Adäquatheit des Zipf–Mandelbrotschen Gesetzes für diesen Fall akzeptieren.

Grotjahn unterschied kurze und lange Vokale und betrachtete Diphthonge als separate Phoneme. Eine andere phonematische Analyse hätte sicherlich etwas andere Parameter ergeben.

Tabelle 2.12

Ranghäufigkeitsverteilung der Phoneme in Goethes "Erlkönig"

nach GROTJAHN (1979)

Rang r	Häufigkeit f_r	NP_r (2.5.1)	Rang r	Häufigkeit f_r	NP_r (2.5.1)
1	111	101.85	21	14	10.92
2	97	86.16	22	14	10.13
3	66	73.59	23	13	9.41
4	66	63.40	24	13	8.77
5	51	55.04	25	12	8.18
6	41	48.12	26	10	7.64
7	35	42.33	27	9	7.15
8	34	37.45	28	7	6.70
9	33	33.31	29	4	6.29
10	27	29.77	30	4	5.92
11	27	26.72	31	4	5.57
12	25	24.09	32	3	5.25
13	23	21.80	33	3	4.96
14	21	19.79	34	2	4.68
15	17	18.03	35	2	4.43
16	17	16.47	36	2	4.20
17	17	15.10	37	1	3.98
18	16	13.87	38	1	3.78
19	15	12.78	39	1	3.59
20	15	11.80			

N = 873; A = 14.8743; B = 2.7385; K =226.4899
X^2 = 30.90; FG = 36; P = 0.71

Betrachten wir nun die Ranghäufigkeitsverteilung der Wörter. Hier ergeben sich mehrere Probleme. Nimmt man Wortformen, dann stellen die Wörter "Erlkönig" und "Erlkönigs" zwei unterschiedliche Formen dar, aber man ist unsicher, ob "Erlenkönig" in der vierten Strophe als unterschiedlich von "Erlkönig" betrachtet werden sollte oder nicht. Weiter, was tut man mit abgetrennten Präfixen, die der Computer mechanisch als separate Wörter zählt? Zählt man nur Lexeme, dann muß man entscheiden, ob "sein/bin/bist/ist/sind/sei/war" usw. ein Lexem oder mehrere Lexeme

Tabelle 2.13

Ranghäufigkeitsverteilung (a) der Wortformen, (b) der Lexeme in Goethes "Erlkönig"

Rang	Häufigkeit	NP_r
1	11	11.42
2	9	9.14
3	9	7.72
4	7	6.74
5	6	6.01
6	6	5.45
7	5	4.99
8	5	4.62
9	4	4.31
10	4	4.05
11	4	3.82
12	4	3.62
13	4	3.44
14	4	3.29
15	4	3.15
16	3	3.02
17	3	2.90
18	3	2.80
19	3	2.70
20	3	2.61
21	3	2.53
22	2	2.46
23	2	2.39
24	2	2.32
25	2	2.26
26	2	2.20
27	2	2.14
28	2	2.09
29	2	2.04
30	2	2.00
31	2	1.95
32	2	1.91
33	2	1.87
34	2	1.84
35	2	1.80
36	2	1.77
37	2	1.73
38	2	1.70
39	2	1.67
40-124	85	90.53

N = 225; A = 1.7139
B = 0.7090; K = 0.1030
X^2 = 5.91; FG = 99
P ≈ 1.00

Rang	Häufigkeit	NP_r
1	24	23.10
2	14	15.14
3	10	11.47
4	10	9.32
5	9	7.89
6	9	6.87
7	6	6.11
8	6	5.50
9	5	5.02
10	5	4.62
11	4	4.29
12	4	4.00
13	4	3.75
14	4	3.54
15	4	3.35
16	3	3.18
17	3	3.03
18	3	2.89
19	3	2.77
20	3	2.66
21	3	2.55
22	2	2.46
23	2	2.37
24	2	2.29
25	2	2.22
26	2	2.15
27	2	2.08
28	2	2.02
29	2	1.97
30	2	1.91
31	2	1.86
32	2	1.81
33	2	1.77
34	2	1.73
35-97	62	67.05

N = 225; A = 0.5033
B = 0.8276; K = 0.1439
X^2 = 5.69; FG = 78
P ≈ 1.00

sind, ob man "ich" und "mein" unter ein Lexem subsumiert usw. Im Grunde kann man aber immer eine Entscheidung treffen.

Die Ranghäufigkeitsverteilung der Wortformen im "Erlkönig" ist in Tabelle 2.13(a) dargestellt, die der Lexeme in Tabelle 2.13(b). Die unterste Zahl bedeutet, daß in der Tabelle 2.13(a) auf den Rängen 4 bis 124 lauter 1 stehen; analog in Tabelle 2.13(b) für Ränge 35 bis 97. Beim Chiquadrat-Test wurden die Klassen mit der Häufigkeit 1 nicht so zusammengesetzt, wie in der Tabelle dargestellt, sondern so, daß die theoretische Häufigkeit mindestens 1 betragen mußte. Dadurch haben wir die gegebenen Freiheitsgrade erhalten. Bessere Anpassungen sind praktisch unmöglich.

Wie man sieht, sind die Parameter in den drei untersuchten Fällen recht unterschiedlich, obwohl im lexikalischen Bereich einander ähnlicher. Für texttheoretische Zwecke wäre es sehr nützlich, zu untersuchen, wie einzelne Sprachebenen die beiden Parameter gestalten, ob es auch Unterschiede in bezug auf Textsorten gibt oder eventuell auch zwischen Sprachen, wie sich die Gesamttextlänge auf die Parameter auswirkt u.ä.

(6) Von Mandelbrot wurde die Zipfsche Form des Gesetzes aufgrund von Ökonomie-Argumenten auf die Form (2.5.1) modifiziert, und die meisten Beschreibungen halten sich an seine Begründung. Anders verfährt Miller (1957), dessen Interpretation in der Literatur kaum erwähnt wird. Wir werden sie hier aufführen, um auch einen anderen Aspekt der Ranghäufigkeitsverteilung zu zeigen.

Man stelle sich vor, daß ein Affe auf einer Schreibmaschine "schreibt", d.h. zufällig auf Tasten drückt, wobei er (i) die leere Taste mit der Wahrscheinlichkeit p, alle anderen mit der Wahrscheinlichkeit q = 1-p drückt und (ii) die leere Taste nie zweimal hintereinander drückt. Sein "Text" besteht dann aus Sequenzen von jeweils i Buchstaben (i = 1,2,3,...), getrennt durch eine Leerstelle. Man kann also erwarten, daß man Sequenzen der Länge i mit der Wahrscheinlichkeit

$$P_i = pq^{i-1}, \qquad i = 1,2,...$$

erhält, so daß man eine monotone, geometrische Abnahme der Häufigkeiten mit wachsender Länge der Wörter bekommt.

Seien auf der Schreibmaschine n Tasten außer der leeren Taste. (iii) Wenn alle Kombinationen der Buchstaben erlaubt und gleichwahrscheinlich sind, dann kann man aus n Buchstaben genau n^i Wörter der Länge i bilden.

Die Wahrscheinlichkeit eines bestimmten Wortes beträgt dann P_i, dividiert durch die Zahl aller möglichen Wörter der Länge i, d.h. P_i/n^i, so daß man die Gleichung

$$P_i/n^i = pq^{i-1}n^{-i} \tag{2.5.2}$$

erhält. Bezeichnet man die linke Seite als p(w,i) (Wahrscheinlichkeit des Wortes der Länge i), dann kann man aufgrund der Beziehung

$$a^x = e^{x \ln a}$$

(2.5.2) schreiben als

$$\begin{aligned} p(w,i) &= \frac{p}{q} q^i n^{-i} \\ &= pe^{i \ln q} e^{-i \ln n}/q \\ &= pe^{-i(\ln n - \ln q)}/q . \end{aligned} \tag{2.5.3}$$

Wegen n Tasten gibt es n Wörter der Länge 1, n^2 Wörter der Länge 2, n^3 Wörter der Länge 3 usw. Insgesamt gibt es also

$$\Sigma n^i = \frac{n(1-n^k)}{1-n} \tag{2.5.4}$$

Wörter, die höchstens die Länge k haben.

Man ordnet nun die einzelnen Wörter nach ihrer Länge und schreibt ihnen Ränge zu. Die Wörter der Länge 1 (1 Buchstabe) stehen auf den Rängen 1 bis n, die zweibuchstabigen fangen mit Rang n+1 an und gehen bis zum Rang $n(1-n^2)/(1-n)$ (laut 2.5.4), die dreibuchstabigen gehen vom Rang $n(1-n^2)/(1-n) + 1$ bis $n(1-n^3)/(1-n)$ usw.

Ein bestimmtes Wort w der Länge i erhält den durchschnittlichen Rang r(w,i), der sich als die Mitte eines Rangintevalls für die Länge i ergibt:

$$r(w,i) = \frac{1}{2}\left[\frac{n(1-n^{i-1})}{1-n} + 1 + \frac{n(1-n^i)}{1-n}\right]$$

$$= n^{i\,\frac{n+1}{2(n-1)}} - \frac{n+1}{2(n-1)} \quad . \qquad (2.5.5)$$

Ordnet man diese Gleichung um und benutzt die obige Beziehung $n^i = e^{i \ln n}$, dann erhält man

$$\frac{2(n-1)}{n+1}\left[r(w,i) + \frac{n+1}{2(n-1)}\right] = e^{i \ln n} \quad . \qquad (2.5.6)$$

Dieses Resultat kann man in (2.5.3) einsetzen. Da dort der Exponent negativ ist, schreiben wir (2.5.3) etwas um und erhalten

$$p(w,i) = p(e^{i \ln n})^{-(1-\ln p/\ln n)/q} \quad . \qquad (2.5.7)$$

In diese Gleichung kann man nun (2.5.6) einsetzen. Man erhält dann

$$p(w,i) = \frac{p}{q}\left\{\frac{2(n-1)}{n+1}\left[r(w,i) + \frac{n+1}{2(n-1)}\right]\right\}^{-(1-\ln p/\ln n)} \quad . \qquad (2.5.8)$$

Schreibt man

$$A = (n+1)/[2(n-1)],$$
$$B = 1 - \ln p/\ln n,$$

und

$$K = p\{(n+1)/[2(n-1)]\}^B/q,$$

dann kann man die obige Formel als

$$p(w) = \frac{K}{[r(w) + A]^B}$$

darstellen, was mit (2.5.1) identisch ist.

Millers Ableitung hat den Vorteil, daß die Parameter bereits interpretiert sind und daß man keine Ökonomie-Begründungen benötigt. Der Nachteil liegt darin, daß diese Ableitung nur für Wortformen gilt, nicht aber für andere Spracheinheiten. Weiter müßten die Parameter für alle Texte einer Sprache gleich sein, was keineswegs zutrifft; und schließlich gilt, daß keine Sprache alle Kombinationen von Phonemen zuläßt, so daß Bedingung (iii) nicht erfüllt ist.

Empfehlenswerte Lektüre über diese Probleme ist noch Woronczak (1967), Kalinin (1966, 1964), Segal (1961), Mandelbrot (1961), Belonogov (1962), Simon (1955), Carroll (1968).

2.5.2. Das Simon-Herdan Modell

Im vorigen Abschnitt haben wir die Häufigkeiten von Wörtern in Form einer Ranghäufigkeitsverteilung untersucht. Die andere Art der Präsentation ist die nach Häufigkeitsklassen. Dies ist nichts anderes als eine "umgekehrte" Darstellung, in der man Klassen von Wörtern bestimmt, die alle die gleiche Häufigkeit haben. Betrachtet man Tabelle 2.13(a), dann kann man feststellen, daß es im "Erlkönig" genau 85 Wörter gibt, die jeweils genau 1-mal im Text vorkommen, es gibt 18 Wörter, die jeweils 2-mal vorkommen usw., d.h., in der Spalte "Häufigkeit" zählt man die Anzahl der 1, der 2,..., und man bekommt eine neue empirische Häufigkeitsverteilung, die in Tabelle 2.14 dargestellt ist. Die Zufallsvariable (X) ist hier die Häufigkeit, und die Häufigkeit (f_x) wird durch die Zahl der Wörter, die x-mal vorkommen, repräsentiert.

Tabelle 2.14

Häufigkeitsklassenverteilung der Wortformen in Goethes "Erlkönig"

x	f_x	NP_x (Yule)	NP_x (Waring)
1	85	85.00	84.98
2	18	20.34	20.57
3	6	7.85	7.90
4	7	3.81	3.80
5	2	2.12	2.10
6	2	1.30	1.27
7	1	0.85	0.82
8	0	0.58	0.56
9	2	0.42	0.40
10	0	0.31	0.29
≥11	1	1.42	1.21
N = 124		b = 2.1795 X^2 = 4.21 FG = 6 P = 0.65	b = 2.2881 n = 1.0507 X^2 = 4.39 FG = 5 P = 0.49

Die ersten umfangreichen Untersuchungen wurden von Yule (1944) durchgeführt, jedoch noch ohne ein Modell. Simon (1955) leitete eine Verteilung aus einem stochastischen Prozess ab, die er als Yule-Verteilung bezeichnete.

Diese Verteilung, die man für unsere Zwecke als

$$P_x = \frac{b(x-1)!}{(b+x)^{(x)}} = \frac{b(x-1)!}{(b+1)^{(x)}}, \quad x = 1,2,... \qquad (2.5.9)$$

schreiben könnte, wo

$$c^{(x)} = c(c-1)(c-2)...(c-x+1)$$

bzw. $$c^{(x)} = c(c+1)...(c+x-1)$$

ist, paßt auf zahlreiche empirische Verteilungen. Es gibt aber viele, für die sie sich als ungeeignet erwiesen hat. Den Parameter b kann man z.B. aus dem Erwartungswert schätzen, da

$$\mu'_1 = b/(b-1) \qquad (2.5.10)$$

so daß

$$b^* = \frac{\bar{x}}{\bar{x} - 1}, \qquad (2.5.11)$$

oder aus der Häufigkeit der ersten Klasse als

$$b^* = \frac{f_1}{N - f_1}. \qquad (2.5.12)$$

In unserem Beispiel bekommen wir aus (2.5.10)

$$b^* = 85/(124-85) = 2.1795.$$

So erhalten wir nach (2.5.9)

$$P_1 = \frac{2.1795(1)}{2.1795 + 1} = 0.6855$$

und $$NP_1 = 124(0.6855) = 85.00.$$

Die weiteren Werte berechnen wir mit der Rekursionsformel

$$NP_x = \frac{x - 1}{b + x} NP_{x-1}, \qquad (2.5.13)$$

z.B.

$$NP_2 = \frac{1}{2.1795 + 2} 85.00 = 20.34$$

$$NP_3 = \frac{2}{2.1795 + 3} 20.34 = 7.85$$

usw. Alle Werte sind in der dritten Spalte von Tabelle 2.14 aufgeführt. Die Anpassung ist sehr gut. Ein Optimierung verbessert sie auf b = 1.9807, X^2 = 3.72, FG = 6, P = 0.71.

Auch wenn das Modell hier paßt, läßt sich zeigen, daß dies nicht immer der Fall ist. Um diesen Nachteil zu beseitigen, hat man mehrere neue Modelle entwickelt. Haight und Jones (1974) haben den stochastischen Prozess von Simon verallgemeinert und eine neue Klasse von Verteilungen erhalten. Die Yule-Verteilung ist ein Spezialfall dieser Klasse (vgl. auch Lánský, Radil-Weiss 1980). Sichel (1975) verallgemeinerte die Poisson-Verteilung und erhielt gleichfalls eine Klasse, in der die Yule-Verteilung einen Spezialfall bildet. Orlov (vgl. Orlov, Boroda, Nadarejšvili 1982) brachte die Häufigkeitsklassendarstellung in einen Zusammenhang mit dem Zipf-Mandelbrotschen Gesetz. Obwohl theoretisch äußerst fruchtbar, scheinen die Anpassungen nicht befriedigend zu sein.

Einen anderen Weg ging Herdan (1964), der für die Anpassung die ***Waring-Verteilung*** benutzte, die wir in der Form

$$P_x = \frac{b \; n^{(x-1)}}{b + n \; (b+n+1)^{(x-1)}}, \quad x = 1,2,... \qquad (2.5.14)$$

schreiben können. Sowohl die Yule- als auch die Waring-Verteilung sind hier in der sogenanten 1-verschobenen Form geschrieben. Diese Verteilung ist oft erfolgreich für die Anpassung benutzt worden (vgl. Muller 1965, 1968, 1977; Tešitelová 1967).

Wie man leicht sehen kann, ist die Yule-Verteilung ein Spezialfall der Waring-Verteilung mit n = 1, denn in dem Fall ist (2.5.14)

$$P_x = \frac{b}{b+1} \cdot \frac{1^{(x-1)}}{(b+2)^{(x-1)}}$$

$$= \frac{b(x-1)!}{(b+1)(b+2)\ldots(b+2+x-2)}$$

$$= \frac{b(x-1)!}{(b+1)(b+2)\ldots(b+x)} ,$$

was mit (2.5.9) identisch ist.

Die Schätzung der Parameter kann z.B. folgendermaßen erfolgen:

$$b^* = \frac{(\bar{x} - 1)f_1}{xf_1 - N} \qquad (2.5.15)$$

$$n^* = \frac{(\bar{x}-1)(N-f_1)}{xf_1 - N} .$$

Die Rekursionsformel lautet

$$P_x = \frac{n + x - 2}{b + n + x - 1} P_{x-1} \qquad (2.5.16)$$

In unserem Beispiel ist $\bar{x} = 1.8145$, so daß

$$b^* = \frac{0.8145(85)}{1.8148(85)-124} = 2.2881$$

$$n^* = \frac{0.8145(124-85)}{1.8145(85)-124} = 1.0507$$

Die berechneten NP_x sind in der vierten Spalte von Tabelle 2.14 angegeben. Hier ergibt sich $X^2_5 = 4.39$, $P = 0.49$, was schlechter ist als die Anpassung der Yule-Verteilung. Durch Optimierung und geeignete Klassenzusammenfassung erhalten wir jedoch etwas bessere Resultate, nämlich $n = 0.9536$, $b = 1.9179$, $X^2_5 = 3.46$, $P = 0.64$, was aber wegen eines verlorenen Freiheitsgrades dennoch schlechter ist als die Yule-Verteilung.

Nutzt man diesen Ansatz für die Erfassung der Verteilung der Klassenhäufigkeiten, dann kann man sich folgende Probleme stellen:

(1) Für welche Texte genügt die Yule-Verteilung, wo muß man die Waring-Verteilung einsetzen?

(2) In welchem Verhältnis stehen die Parameter dieser Verteilung zu dem Vokabular der Texte?

(3) In welchen Intervallen bewegen sich die Werte dieser Parameter für bestimmte Textsorten?

(4) Falls beide Verteilungen versagen und man bei diesem Ansatz bleiben will, wie verallgemeinert man die Waring-Verteilung? Gemäß unserem Vorschlag in § 2.4 folgen diese Verteilungen aus dem Ansatz

$$D = - \frac{a}{x + d} , \qquad (2.5.17)$$

wobei für die Yule-Verteilung $a = b+1$, $d = b$ und für die Waring-Verteilung $a = b+1$, $d = b+n-1$, so daß man eine Verallgemeinerung auf verschiedene Weisen durchführen kann. Die bekannteste ist vielleicht die verallgemeinerte hypergeometrische Verteilung Typ IV von Kemp und Kemp (1956) (vgl. auch Johnson, Kotz 1969:158-160), aber der obige Ansatz liefert hinreichend viele linguistisch gut interpretierbare Erweiterungen.

2.5.3. Das Referenzgesetz von Hřebíček

Im einem laufenden Text werden nicht nur Einheiten wiederholt, sondern auch Bezüge auf bestimmte Inhalte hergestellt. So bezieht sich in den Sätzen "Es war einmal ein König. Er hatte drei Söhne" das Pronomen "er" auf den König. In Goethes "Erlkönig" beziehen sich die Wörter "Sohn", "Kind", "Knabe" immer auf dieselbe Person. Das Adverb "danach" in einem Text bezieht sich immer auf ein Ereignis im Vortext. Solche Bezüge nennt man Referenzen, die in der Textlinguistik bereits ausgiebig untersucht worden sind (s. z.B. Harweg 1974; Palek, Fischer 1977; Halliday, Hasan 1976 usw.).

Ein Text ist so konstruiert, daß er immer Referenzen enthält. Die Effektivität der Kommunikation verlangt, daß die Referenzen nicht chao-

tisch, sondern in Abhängigkeit von anderen Eigenschaften des Textes erscheinen.

Hřebíček (1985) geht von zwei Annahmen aus:

(1) Je wortreicher der Text, desto weniger Referenzen,
(2) je mehr Sätze im Text, desto mehr Referenzen gibt es.

Daraus folgt, daß die Zahl der Referenzen nur von diesen beiden Größen abhängen soll. Bezeichnet man als

R = Zahl der Referenzen
S = Zahl der Sätze
W = Vokabularreichtum des Textes (types)
N = Zahl der Wörter des Textes (tokens, Textlänge),

dann verändert sich die Zahl der Referenzen - bezogen auf die Veränderung der Satzzahl - proportional zu dem Vokabular des Textes d.h.

$$\frac{\delta r}{\delta s} = aw, \qquad (2.5.18)$$

und gleichzeitig verändert sich die Zahl der Referenzen - bezogen auf die Veränderung des Vokabulars - proportional zu der Zahl der Sätze, d.h.

$$\frac{\delta r}{\delta w} = bs, \qquad (2.5.19)$$

woraus sich die Funktion

$$r = csw \qquad (c = ab) \qquad (2.5.20)$$

ergibt. Hřebíček erwähnt mehrere Möglichkeiten der Interpretation von W, dem Vokabularreichtum, und wählt die folgende: Sei V die Zahl der Wort-types im Text, dann ist

$$w = \frac{v}{n},$$

woraus

$$r = cs\,\frac{v}{n} \qquad (2.5.21)$$

folgt. Nach Herdan (1966:76) gilt aber

$$v = n^a . \qquad (2.5.22)$$

Setzt man dies in (2.5.21) ein, so erhält man

$$r = csn^{a-1}$$

oder einfach

$$r = csn^b, \qquad (2.5.23)$$

wo b und c bestimmte Konstanten sind, s die Zahl der Sätze und n die Zahl der Worttokens im Text.

Hřebíček überprüfte sein Modell an 10 türkischen Texten, von denen wir hier ein Beispiel übernehmen (s. Tabelle 2.15).

Tabelle 2.15

Anwachsen der Referenzenzahl in einem türkischen Text*

nach Hřebíček (1986)

Zahl der Sätze s_i	Zahl der Worttokens n_i	Zahl der Referenzen r_i	Berechnete Zahl der Referenzen r_i^*
10	58	24	23.82
20	100	46	43.87
30	149	62	61.96
40	246	81	76.59
50	299	82	92.95
60	403	96	106.63
70	491	117	120.74
80	601	141	133.84
90	676	156	147.92
100	786	168	160.66

[1] Bügünün Diliyle Atatürk'ün Söylevleri (ed. B.K. Çağlar) Ankara 1968:85 ff.

Mit Hilfe der Methode der kleinsten Quadrate, die wir auf die logarithmierte Form von (2.5.23) anwenden, d.h. durch Minimierung von

$$\sum_{i=1}^{k} (\ln r_i - \ln c - \ln s_i - b \ln n_i)^2 \qquad (2.5.24)$$

erhalten wir

$$D = k\Sigma(\ln n_i)^2 - (\Sigma \ln n_i)^2$$

$$\ln c^* = [(\Sigma \ln n_i)^2(\Sigma \ln r_i - \Sigma \ln s_i) - \qquad (2.5.25)$$

$$-\Sigma \ln n_i(\Sigma \ln r_i \ln n_i - \Sigma \ln s_i \Sigma \ln n_i)]/D$$

$$b^* = [k(\Sigma \ln r_i \ln n_i - \Sigma \ln s_i \ln n_i) -$$

$$- \Sigma \ln n_i(\Sigma \ln r_i - \Sigma \ln s_i)]/D$$

Daraus ergibt sich

$$\ln c^* = 1.4810, \qquad c^* = 4.3973, \qquad b^* = -0.1510.$$

Die mit diesen Parametern berechneten Werte findet man in der vierten Spalte von Tabelle 2.15.

Die Anpassungsgüte bewerten wir einfach mit dem Determinationskoeffizienten

$$R = 1 - \frac{\Sigma(r_i - r_i^*)^2}{\Sigma(r_i - \bar{r})^2}, \qquad (2.5.26)$$

wo r_i die beobachteten, r_i^* die berechneten Referenzen und $\bar{r}$ die durchschnittliche Referenzenzahl ist (alle logarithmisch). So erhalten wir

$$R = 1 - \frac{0.0407}{3.2852} = 0.9876.$$

R bewegt sich in dem Intervall <0,1>. Je größer R, desto mehr Variabilität wird von den unabhängigen Variablen erklärt, d.h., desto besser ist die

Anpassung. Ein derartig hoher Determinationskoeffizient zeigt, daß das Modell gut geeignet ist. Der F-test liefert $F_{2,5} = 7.9$, was mit $P = 0.013$ ein signifikantes Resultat bedeutet.

Mit diesem Modell kann man folgende Probleme untersuchen:

(1) Wie verhalten sich spezielle Referenzen, z.B. wie verläuft die Pronominalisierung im Text?

(2) Wie verhalten sich einzelne Textsorten, d.h. wie verlaufen die Referenzen in unterschiedlichen Textsorten? Die Unterschiede kann man an den Parametern ersehen, so daß man eventuell auch eine Referenztypologie der Texte aufstellen kann.

(3) Wie ist der Verlauf der Referenzen in Texten aus verschiedenen Sprachen, jedoch von gleicher Textsorte?

2.5.4. Type-token Modelle

Eines der beliebtesten textanalytischen Probleme ist die Messung der Beziehung zwischen der Anzahl der types (= unterschiedliche Wörter) eines Textes und der Zahl der tokens (= alle Wörter), d.h. zwischen dem Wortschatz und der Textlänge.

Bei mechanischer Textverarbeitung pflegt man als Wort die Wortform zu nehmen, d.h. "Haus" und "Hauses" als zwei unterschiedliche Wörter. Dies läßt sich schwerlich als Wortschatzmessung begründen, man kann es eventuell beim Sprachlernen der Kinder anwenden, wobei man prüft, wie stark die Analogiebildung nach Alter fortgeschritten ist.

In der type-token-Problematik sollte man also Lemmata als unterschiedliche Wörter betrachten. Aber auch hier gibt es Probleme, die nur durch festgesetzte Kriterien zu lösen sind, wobei diese sprachspezifisch gestaltet werden müssen. Für das Deutsche werden wir in einer Zählung folgende Kriterien verwenden:

Zu einem Lemma gehören

(i) Alle Formen eines Nomens
(ii) Alle Formen eines Verbs (z.B. sehen, siehst, hat gesehen, ist gesehen worden,...), auch bei Suppletivismus, (z.B. sein, bin, bist, sind, wäre,...).

(iii) Alle Formen eines Adjektivs (z.B. schön, schöner, am schönsten, schönen, schönes,...), auch wenn sie in adverbialer Funktion benutzt werden.

(iv) Alle Formen eines Pronomens (z.B. ich, meiner, mir, mich). Zu einem Lemma gehören auch er/sie/es und ihre Formen.

(v) Alle Formen des Artikels (z.B. der, die, das, dessen, deren, dem,...), d.h. auch die Pluralformen und die Anwendung als Relativpronomina.

(vi) Alle Formen eines Zahlworts (z.B. drei, dritter, dritt).

(vii) Abtrennbare Affixe bilden mit dem Grundwort ein Lemma.

Man kann die Kriterien natürlich auch anders wählen. Die grundlegende Unterscheidung liegt in der Lemma- oder der Wortformenzählung, wobei nur die erste Art den Wortschatz mißt.

Zählen wir unter den obigen Kriterien die types und die tokens in Goethes "Erlkönig", dann bekommen wir die Resultate wie in Tabelle 2.16 Spalte T dargestellt. Man sieht, daß der type-Verlauf schon bei L = 15 absinkt und dieser Trend immer stärker wird. Es handelt sich nun darum, wie man diesen Trend modellieren kann.

Viele Autoren haben die Zählungen nicht per Einzeltoken sondern per 100 oder mehr tokens durchgeführt; die meisten haben Wortformen gezählt, und es ging ihnen um eine Prognose des Wortschatzes des Autors, zu der wir unten noch zurückkehren werden.

Man kann die Ansätze in zwei grobe Gruppen aufteilen:

(I) Betrachtung des Textes als stochastischer Prozess, die sicherlich eine große Zukunft hat. Die Ansätze wurden jedoch nur für die Prognose des Wortschatzes benutzt - eine etwas illusorische Auffassung der Aussagekraft des Textes (vgl. Brainerd 1972; McNeil 1973; Gani 1975). Andere Autoren haben mit stochastischen Prozessen die Verteilung der Worthäufigkeitsklassen mit besseren Resultaten modelliert (vgl. Simon 1955; Haight, Jones 1974; Lánský, Radil-Weiss 1980).

(II) Ableitung einer Kurve aus Überlegungen über den Informationsfluß in Texten oder Ausprobieren einer Kurve zum Erreichen einer guten Anpassung (vgl. Herdan 1966; Müller 1971; Maas 1972; Nešitoj 1975; Ratkowsky, Halstead, Hantrais 1980; Tuldava 1980; Orlov, Boroda, Nadarejšvili 1982). Diese Ansätze haben bessere Resultate erbracht, daher werden wir bei diesen Verfahren bleiben.

Einen im Entstehen befindlichen Text betrachten wir als ein System, das sich in einem mehrdimensionalen Eigenschaftsraum

$$T = \langle E_{11}, E_{12}, \ldots, E_{21}, E_{22}, \ldots, E_{n1}, E_{n2}, \ldots \rangle$$

entfaltet. Diese Dimensionen (E_{it}) sind die zur Zeit t erreichten Ausprägungen der Eigenschaften i, wobei einige davon die Textlängen darstellen. Da man Textlänge in Zahl der Kapitel, Absätze, Sätze, Teilsätze, Wörter, Szenen, Auftritte, Pausen usw. messen kann, gibt es gleichzeitig mehrere Textlängen, und jede von ihnen stellt einen Ordnungsparameter dar, der seine eigenen Subsysteme "versklavt" (vgl. Haken 1978). Mißt man z.B. die Textlänge in Sätzen, so muß sich diese Längenart auf irgendeine Eigenschaft der Sätze auswirken. Bei Sätzen ist natürlich ein type-token-Verhältnis sinnlos, da jeder Satz neu ist. Bei einer Komödie kann sich die Textlänge, gemessen als Anzahl der Auftritte, auf die steigende Komik der Auftritte auswirken usw. Die Aufgabe der Textanalyse besteht auch darin, diese "Versklavung" zu erforschen.

Bei Textlänge, gemessen als Anzahl der Worttokens regelt diese Größe den Zuwachs neuer Wörter (und möglicherweise noch andere Worteigenschaften). Um die Art dieser Regelung abzuleiten, greifen wir zu einer Beziehung, die zahlreiche linguistische Konstrukt-Komponente-Abhängigkeiten regelt (vgl. Altmann, Schwibbe, Kaumanns, Köhler, Wilde 1988) und im Köhlerschen selbstregulierenden System (Köhler 1986) eine grundlegende Rolle spielt, nämlich den Ansatz (vgl. § 2.4)

$$D = a/x, \qquad (2.5.27)$$

d.h. hier

$$\frac{dT}{T} = \frac{adL}{L}, \qquad (2.5.28)$$

wo T die Anzahl der types im Text der Länge L ist. Der Koeffizient a zeigt in der Lösung der Differentialgleichung

$$T = cL^a \qquad (2.5.29)$$

den Anstieg der Kurve, der von der Art der types abhängt. Bei der type-token Beziehung der Wörter ist immer $0 < a < 1$, und zwar bei den Lemma-types kleiner als bei den Wortformen-types. Die Konstante c hängt von der Zählungsart der tokens ab. Zählt man die Wörter einzeln,

dann müßte c = 1 sein, zählt man sie in Zehnergruppen, Hundertergruppen usw., dann wird es immer größer.

Gleichung (2.5.28) sagt, daß die relative Zunahmerate der types der relativen Zunahmerate der Textlänge proportional ist. Zu diesem Resultat ist bereits Herdan gekommen (1966:76), und es hat den Vorteil, daß man ähnliche Überlegungen auch in der Biologie und der Ethologie kennt (vgl. Fagen, Goldman 1977). Die analoge Strukturierung der Beziehung System-/Subsystemgröße ist eine starke Unterstützung gerade für diese Sicht.

In einer System/Subsystem-Beziehung besteht immer eine Dominanz des Systems, das seine Subsysteme integrieren will. Der Autonomiedrang der Subsysteme (hier z.B. die Eigenart des Stils, der Textsorte u.ä.) kann eine Abweichung verursachen, die man durch Adaptation der Grundformel (2.5.27) bzw. (2.5.28) abfangen kann (vgl. die Formeln von Tuldava 1980).

Die Anpassung von (2.5.29) an den "Erlkönig" ergibt das Resultat in Tabelle 2.16. Die Kurve lautet

$$T = L^{0.8672} \quad ,$$

und der F-test ergibt nach Logarithmierung F(1,222) = 6410.43, was eine so gut wie perfekte Anpassung bedeutet.

Die Untersuchung des type-token-Verhältnisses ist linguistisch bedeutsam, weil es zu einem Sprachproduktionsgesetz führt. Sprachtheoretisch läßt es sich in der obigen Form gut systematisieren und zeigt einen synergetischen Aspekt der Texte. In der Literaturwissenschaft kann es z.B. der Diskrimination der Texte (Textarten, Stile u.a.) dienen.

Das type-token-Verhältnis wurde des öfteren zur Schätzung des Vokabulars eines Autors aufgrund eines Textes herangezogen, was wir als weniger sinnvoll betrachten. Es läßt sich nämlich zeigen, daß jeder Text eines Autors zu sehr unterschiedlichen Schätzungen führt, die darüber hinaus auch unrealistisch sind. Auch die ad-hoc-Hypothese, daß es sich dabei nur um diejenigen Wörter handele, die der Autor für die Verfassung des gegebenen Textes "zur Verfügung" gestellt habe, ist äußerst problematisch. Jeder erwachsene Sprecher einer Sprache kennt nämlich (mit einer unbedeutender Toleranz) etwa die gleiche Anzahl der Wörter seiner Sprache, auch wenn es Spezialisierungsunterschiede gibt. Welches relevante Resultat kann also eine Schätzung des Gesamtwortschatzes eines Autors erbringen? Sinnvoll wäre eine derartige Schätzung bei der Entwicklung des Kinderwortschatzes, der erst bei einem bestimmten Alter gegen den Erwachsenenwortschatz zu konvergieren anfängt.

Tabelle 2.16

Anpassung der Potenzkurve an die Daten von "Erlkönig"

L	T	T'	L	T	T'	L	T	T'	L	T	T'	L	T	T'
1	1	1.00	48	35	28.69	95	56	51.87	142	74	73.49	189	92	94.17
2	2	1.82	49	35	29.21	96	56	52.34	143	74	73.94	190	92	94.60
3	3	2.59	50	36	29.73	97	57	52.81	144	74	74.39	191	92	95.04
4	4	3.33	51	36	30.24	98	58	53.28	145	75	74.84	192	92	95.47
5	5	4.04	52	37	30.76	99	58	53.75	146	75	75.29	193	92	95.90
6	6	4.73	53	37	31.27	100	59	54.23	147	75	75.73	194	93	96.33
7	7	5.40	54	37	31.78	101	60	54.69	148	76	76.18	195	94	96.76
8	8	6.07	55	37	32.29	102	60	55.16	149	77	76.62	196	94	97.19
9	9	6.72	56	37	32.80	103	60	55.63	150	77	77.07	197	94	97.62
10	10	7.36	57	38	33.31	104	60	56.10	151	78	77.52	198	95	98.05
11	11	8.00	58	39	33.81	105	60	56.57	152	78	77.96	199	95	98.48
12	12	8.62	59	39	34.32	106	61	57.04	153	78	78.41	200	95	98.91
13	13	9.24	60	40	34.82	107	62	57.50	154	78	78.85	201	95	99.33
14	14	9.86	61	40	35.32	108	63	57.97	155	78	79.29	202	96	99.76
15	14	10.47	62	41	35.82	109	63	58.43	156	78	79.74	203	96	100.19
16	15	11.07	63	42	36.33	110	63	58.90	157	78	80.18	204	96	100.62
17	15	11.67	64	42	36.82	111	64	59.36	158	79	80.62	205	96	101.05
18	16	12.26	65	43	37.32	112	65	59.82	159	79	81.06	206	96	101.47
19	17	12.85	66	44	37.82	113	65	60.29	160	80	81.51	207	96	101.90
20	18	13.43	67	45	38.32	114	65	60.75	161	80	81.95	208	97	102.33
21	18	14.01	68	46	38.81	115	65	61.21	162	81	82.39	209	97	102.75
22	19	14.59	69	47	39.31	116	65	61.67	163	82	82.83	210	98	103.18
23	19	15.16	70	47	39.80	117	65	62.13	164	82	83.27	211	98	103.61
24	20	15.73	71	47	40.29	118	65	62.59	165	83	83.71	212	99	104.03
25	20	16.30	72	47	40.78	119	66	63.05	166	84	84.15	213	99	104.46
26	21	16.86	73	48	41.28	120	67	63.51	167	85	84.59	214	100	104.88
27	21	17.42	74	49	41.76	121	67	63.97	168	85	85.03	215	100	105.31
28	22	17.98	75	50	42.25	122	68	64.43	169	85	85.47	216	101	105.73
29	22	18.54	76	50	42.74	123	68	64.89	170	86	85.91	217	101	106.16
30	23	19.09	77	51	43.23	124	68	65.34	171	86	86.34	218	101	106.58
31	24	19.64	78	51	43.72	125	68	65.80	172	86	86.78	219	101	107.00
32	25	20.19	79	52	44.20	126	69	66.26	173	87	87.22	220	101	107.43
33	26	20.73	80	52	44.69	127	69	66.71	174	87	87.66	221	101	107.85
34	27	21.28	81	53	45.17	128	70	67.17	175	87	88.09	222	101	108.27
35	28	21.82	82	53	45.65	129	71	67.62	176	87	88.53	223	102	108.70
36	29	22.36	83	53	46.14	130	71	68.08	177	87	88.96			
37	29	22.90	84	54	46.62	131	72	68.53	178	88	89.40	a = 0.8672		
38	30	23.43	85	55	47.10	132	72	68.98	179	88	89.84			
39	31	23.97	86	55	47.58	133	73	69.44	180	89	90.27	F(1,222)=6410.43		
40	32	24.50	87	55	48.06	134	73	69.89	181	89	90.71			
41	33	25.03	88	55	48.54	135	74	70.34	182	90	91.14			
42	33	25.56	89	55	49.01	136	74	70.79	183	90	91.57			
43	33	26.08	90	55	49.49	137	74	71.24	184	90	92.01			
44	33	26.61	91	56	49.97	138	74	71.70	185	90	92.44			
45	34	27.13	92	56	50.44	139	74	72.15	186	90	92.87			
46	35	27.66	93	56	50.92	140	74	72.60	187	91	93.31			
47	35	28.18	94	56	51.39	141	74	73.04	188	92	93.74			

Nichtsdestoweniger kann man den Parameter a als ein Charakteristikum des Textes benutzen: Je weniger sich einmal benutzte Wörter wiederholen, d.h., je schneller der Zuwachs neuer Wörter, desto größer wird a. Dieser Parameter bewegt sich im offenen Intervall (0,1). Den Wert 0 kann er nicht erreichen, da in dem Fall der Text nur aus der ständigen Wiederholung eines einzigen Wortes bestünde (c wäre gleich 1). Er kann nicht größer als 1 werden, da es nicht mehr types als tokens geben kann. Aus systemtheoretischen Gründen muß er notgedrungen kleiner als 1 sein. Da Text ein input für den Hörer ist, mit dem bestimmte Inhalte übertragen werden, die der Hörer verarbeiten muß und sich mindestens bis zu Ende des Textes merken soll, "zwingt" er den Sprecher, den Informationsfluß so zu steuern, daß seine Informationsverarbeitungs- und Gedächtniskapazität nicht überfordert werden. Das gleiche Geschehen haben wir beim Lernen der Sprachen: Werden das Kind oder der Lerner der Sprache nur mit immer neuen Wörtern überschüttet, so werden sie sich kaum etwas merken. Das Kind, der Lerner brauchen Wiederholung, damit sich die Assoziationen zwischen Laut und Bedeutung verfestigen. Der Hörer/Leser braucht Wiederholung, damit sich inhaltliche Einheiten verfestigen. Der Text als input besteht nicht nur aus neuer Information, sondern auch aus Wiederholung der alten Information. Die Analogie zu anderen Systemen ist hier evident: Lebende Systeme brauchen sowohl input zur Aufrechterhaltung ihrer Existenz (maintenance input) als auch Informationsinput (signal input). Der Hörer braucht maintenance input (= Wiederholung), damit die Inhalte des Textes nicht erlischen, und signal input, damit die Kommunikation überhaupt im Gange bleibt. Zwischen den beiden input-Arten muß es aber eine invariante Beziehung geben, denn nach Berrien (1968: 80) "...an optimum balance must be struck between maintenance and signal input, for without an adequate supply of the former, the latter can not be proceeded." Das obige Gesetz drückt gerade diese Ausgewogenheit aus.

Es wäre lohnenswert, zu untersuchen, welche Unterschiede es zwischen Texten gibt, ob es Unterschiede auch zwischen Sprachen gibt, ob es eine Veränderung des Parameters a mit steigendem Alter eines Verfassers gibt und ob sich a von einfachen Formen bis zu modernen Gedichten irgendwie ändert.

Der weitere Verlauf der Kurve kann bei langen Texten leicht und gut vorausgesagt, aber schlecht verfolgt werden. Daher empfiehlt es sich, aus langen Texten mehrere Stichproben zu nehmen. Es kann sich eventuell herausstellen, daß Unterschiede auch kapitelweise vorhanden sind.

2.5.5. Ausblick

In dem Augenblick, da man die Ebene der morphologischen und der syntaktischen Regeln verläßt, die in einem Text sozusagen notgedrungen befolgt werden, stößt man auf Erscheinungen, die nach Gesetzen verlaufen. Man sollte sich darunter keine deterministischen Richtlinien vorstellen, die den Autor in ein Prokrustesbett so hineinzwingen, daß er keine Freiheit hätte. Vielmehr verläuft die Textschöpfung so, daß der Autor, um seinem Text Originalität zu verleihen, Extrema einiger Texteigenschaften zu erreichen sucht. Füllt er aber den Text mit Entitäten einer Art, so kann er wenige von einer anderen Art hineinbringen. Wenn eine spezielle Charakteristik des Textes einen hohen Wert erreicht, dann erreicht eine andere einen niedrigeren. Und dies ist genau der Zustand, den man aus systemtheoretischer Sicht erwartet: Die Eigenschaften der kollateralen Subsysteme konkurrieren und kooperieren miteinander, und das System (der Texte) bleibt im Gleichgewicht. Die Aufgabe des Textwissenschaftlers besteht – auf der theoretischen Ebene – darin, die Selbstregulation aller im Text verlaufenden Prozesse zu ermitteln. Dazu muß er natürlich zuerst die Prozesse bzw. die aus den Prozessen resultierenden Verteilungen oder Kurvenverläufe kennen. Es wäre daher nicht verfehlt, alle in § 2 aufgeführten Charakteristika an vielen verschiedenen Texten zu ermitteln, damit man einen Einblick in die Gesamtdynamik der Texte bekommt.

3. POSITIONALE WIEDERHOLUNG

Tendenzielle Wiederholung von Texteinheiten ist möglich an bestimmten Stellen des Textes, und zwar in bestimmten Positionen in größeren Texteinheiten, z.B. am Anfang, in der Mitte, am Ende oder vor bzw. hinter Texteinheiten der gleichen Klasse.

Von der ersten Art ist etwa das Vorkommen eines Wortes am Ende des Verses oder die Positionierung der Wortarten im Satz (vgl. Průcha 1967). In der Linguistik bezeichnet man dies als eine funktionale Relation. Von der zweiten Art ist z. B. das Vorkommen eines Nomens vor einem Verb. Dies bezeichnet man als eine distributionale Relation.

Es werden hier keine poetischen Figuren untersucht, weil die Poetik gerade an deren Einmaligkeit oder an deren Stereotypie (z.B. Reim) interessiert ist, vielmehr solche, die man nur als Tendenzen erkennen kann. Sollte eine Figur tendenziell erscheinen, dann kann man sie mit den in diesem Band aufgeführten Methoden aufspüren.

3.1. Reimendung im Erlkönig

Betrachten wir die phonetische Transkription des "Erlkönigs" von Goethe, wie sie von Grotjahn (1979) durchgeführt wurde: Man sieht sofort, daß der letzte Laut des Verses am häufigsten ein [t] ist. Handelt es sich hier um eine Tendenz, oder ist dies in Übereinstimmung mit der Gestaltung und Anwendung deutscher Wörter?

Wir testen also die Hypothese, daß am Versende im "Erlkönig" keine Tendenz zum häufigen Vorkommen von [t] vorliegt, gegen die Hypothese, daß [t] hier häufiger als erwartet vorkommt. Vergleicht man die Häufigkeit (f_i) der Laute am Versende im "Erlkönig", so findet man folgende Zahlen:

Laut	f_i
[f]	2
[m]	2
[n]	6
[o]	2
[r]	2
[t]	18

Eine Tendenz würde nicht vorliegen, wenn wir zeigen könnten, daß auch an anderen Stellen des Gedichts [t] sehr häufig am Wortende vorkommt. Das bedeutet, daß wir die erwartete Häufigkeit von [t] aus dem Rest des Gedichts berechnen müssen. Im "Erlkönig" gibt es 255 Wörter, davon 32 Reimwörter. Im Rest, d.h. unter 193 Wörtern, gibt es folgende Endlaute:

Laut	f_i	Laut	f_i
[i:]	1	[s]	12
[e]	1	[r]	29
[e:]	2	[l]	2
[ə]	14	[n]	50
[o:]	4	[t]	43
[u:]	7	[ç]	19
[f]	1	[x]	1
[m]	6	[ŋ]	1

Die Proportion der auf [t] endenden Wörter ist im Restgedicht also

$$p_t = 43/193 = 0.2228.$$

Unter 32 Versen erwarten wir also

$$Np_t = 32(0.2228) = 7.1296$$

Verse, die auf ein [t] enden. Wir haben aber 18 beobachtet. Ist der Unterschied zwischen Np_t und f_t signifikant, d.h., kann man von einer Tendenz sprechen?

Das Problem läßt sich mit einem Binomialtest lösen. Wir formulieren unsere Hypothese um und fragen, wie groß die Wahrscheinlichkeit ist, daß unter 32 Endlauten 18 oder mehr [t] sind, wenn die Vorkommenswahrscheinlichkeit von [t] p_t ist.

Die Lösung läßt sich folgendermaßen formulieren: Die Wahrscheinlichkeit, daß unter N Endlauten genau x [t], und die restlichen N - x Laute nicht [t] sind, ist

$$P_x = P(X=x) = \binom{N}{x} p^x q^{N-x}, \qquad (3.1)$$

wo q = 1-p und $\binom{N}{x}$ die Binominalkoeffizienten sind. Die gesuchte Wahrscheinlichkeit ergibt sich dann als

$$P(X \geq x_c) = \sum_{x=x_c}^{N} \binom{N}{x} p^x q^{N-x}$$

$$= 1 - \sum_{x=0}^{x_c-1} \binom{N}{x} p^x q^{N-x} \quad . \qquad (3.2)$$

Ist nun $P(X \geq x_c) \leq 0.05$, dann betrachten wir die Tendenz nach [t]-artigen Endungen als reell.

In unserem Fall haben wir zu berechnen

$$P(X \geq 18) = \sum_{x=18}^{32} \binom{32}{x} (0.2228)^x (0.7772)^{32-x} ,$$

was mit

$$P(X \geq 18) = 1 - \sum_{x=0}^{17} \binom{32}{x} 0.2228^x (0.7772)^{32-x}$$

identisch ist. Es ist üblich, die zweite Formel zu benutzen. Zuerst berechnen wir den ersten Summanden, nämlich

$$P_0 = P(X=0) = \binom{32}{0} 0.2228^0 (0.7772)^{32-0} = 0.7772^{32}$$

$$= 0.000314 .$$

Die restlichen Summanden bekommen wir rekursiv mit Hilfe der Formel

$$P_x = \frac{(N-x+1)}{x} \frac{p}{q} P_{x-1} . \qquad (3.3)$$

So wird

$$P_1 = \frac{(N-1+1)}{1} \frac{p}{q} P_0$$

in unserem Fall

$$P_1 = 32 \frac{0.2228}{0.7772} 0.000314 = 0.002881$$

$$P_2 = \frac{31}{2} \frac{0.2228}{0.7772} 0.002881 = 0.012803$$

usw. So erhalten wir

$$P(X \geq 18) = 1-(0.000314 + 0.002881 + 0.012803 + \ldots + 0.000106)$$

$$= 1 - 0.999968$$

$$= 0.000032 .$$

(Die Rechnungen wurden auf 10 Dezimalstellen durchgeführt und gerundet.)

Die Wahrscheinlichkeit ist viel kleiner als unsere kritische Grenze $\alpha = 0.05$, so daß wir im "Erlkönig" mit ziemlich großer Sicherheit von einer "t"-Tendenz sprechen können.

Diese Art der Berechnung ist zwar exakt, aber im Falle von großem N oft langwierig. Bei großem N empfehlen sich folgende Approximationen:

(a) Wenn $p \approx 0.5$, dann benutze man die Normalverteilung und berechne

$$\frac{x - Np}{(Npq)^{1/2}} = z . \qquad (3.4)$$

Bei $N = 100$, $x_c = 70$ und $p = 0.49$ erhalten wir laut (3.2)

$$P(X \geq 70) = 0.00001679,$$

während (3.4)

$$\frac{70 - 1006(0.49)}{[100(0.49)0.51]^{1/2}} = 4.2008$$

ergibt.

Die entsprechende Wahrscheinlichkeit, die man in den Tabellen findet, ist P = 0.000013307, eine einigermaßen akzeptable Annäherung. Für unseren konkreten Fall ist diese Approximation schlecht, da die Bedingungen nicht erfüllt sind.

(b) Ist p sehr klein, dann empfiehlt sich die Approximation durch die Poissonverteilung. Anstelle von (3.2) berechnet man

$$P(X \geq x_c) = 1 - \sum_{x=0}^{x_c-1} \frac{e^{-Np}(Np)^x}{x!} , \qquad (3.5)$$

und zwar die einzelnen Summanden rekursiv als

$$P_0 = e^{-NP}$$

$$P_x = \frac{Np}{x} P_{x-1} . \qquad (3.6)$$

3.2. Offene Reime

Reimwörter sind Träger zahlreicher Funktionen, und ihre Gestaltung ist ein dankbares Untersuchungsobjekt. Sie haben spezielle phonische, metrische, grammatische und semantische Eigenschaften, deren Erscheinen in erhöhtem Maße am Versende der Poesie eine besondere Eigenart verleiht. Diese Eigenschaften, die sich in Wiederholungstendenzen manifestieren, können bei einem Dichter, in einer Epoche, in einer "Schule" konstant sein, verändern aber mit der Zeit ihre anfänglich markante Ausprägung und gehen womöglich zu einem anderen Extrem über. Hat sich die gegebene Eigenschaft als ganze abgenutzt, dann wird sie irrelevant und weist keine Wiederholungstendenzen mehr auf.

Eine derartige "Regularität" ist die Verwendung von "offenen Reimen", d.h. Reimwörtern, die auf Vokal enden, in der slovakischen Poesie (vgl. Štukovský, Altmann 1964, 1965, 1966, aus denen wir die Daten übernehmen). Die Untersuchungsmethode ist die gleiche wie in § 3.1.

Aus dem Werk von S. Chalupka (Spevy Sama Chalupka. Turč. Sv. Martin, 1921) wurden zufällig 206 Reime erhoben; davon endeten 162 auf einen Vokal, 44 auf einen Konsonanten. Das Übergewicht offener Reime

ist zwar offensichtlich, aber man muß den Vergleich mit einer Stichprobe von "nicht-Reimwörtern" durchführen, um festzustellen, ob das Slovakische als ganzes eventuell die gleiche Tendenz (d.h. die größere Proportion der Wörter mit einem Vokal zu enden) aufweist. Die geeignete Stichprobe ist wieder der Rest der erhobenen Verse (Restwörter). So erhält man für Chalupka:

	vokalisch endende	konsonantisch endende	
Reimwörter	162	44	206
Restwörter	638	418	1048

Die Proportion der vokalisch endenden Reimwörter ist 0.7864, die der vokalisch endenden Restwörter 0.6088. Ob dieser Unterschied nun signifikant ist, muß mit einem Test überprüft werden. Bezeichnen wir mit

n_{vr} = Zahl der vokalisch endenden Reimwörter
n_{vg} = Zahl der vokalisch endenden Restwörter
n_r = Zahl der Reimwörter
n_g = Zahl der Restwörter
N = Zahl aller Wörter im Gedicht ($N = n_r + n_g$).

Weiter sei

$$p_v = \frac{n_{vr} + n_{vg}}{n_r + n_g}$$

$$p_{vr} = n_{vr}/n_r$$

$$p_{vg} = n_{vg}/n_g .$$

Dann prüfen wir den Unterschied $p_{vr} - p_{vg}$ mit dem Kriterium

$$t = \frac{p_{vr} - p_{vg}}{\left[p_v(1-p_v)\left(\frac{1}{n_r} + \frac{1}{n_g}\right)\right]^{1/2}} \qquad (3.7)$$

Wegen der großen Zahl der Freiheitsgrade ist die Größe t normalverteilt. Setzen wir die Zahlen in (3.7) ein, dann bekommen wir

$$p_v = \frac{162 + 638}{206 + 1048} = 0.6380$$

und

$$t = \frac{0.7864 - 0.6088}{[0.6380(0.3620)(1/206 + 1/1048)]^{1/2}} = 4.85 .$$

Hypothesen dieser Art sind immer als einseitig zu betrachten, da wir eine Tendenz nur dann testen, wenn bei der Beobachtung $p_{vr} > p_{vs}$. Ein so großer oder ein noch extremerer t- oder z- Wert ergibt sich mit $P \approx 6x10^{-7}$, was eine deutliche Tendenz zu offenem Reim signalisiert.

Ähnliche Zählungen wurden an Texten von 12 slovakischen Dichtern durchgeführt, (vgl. Štukovský, Altmann 1964), nämlich an

1. S. Chalupka, Spevy Sama Chalupku,
2. J. Král', Básne,
3. A. Sládkovič, Spisy básnické II,
4. J. Botto, Spevy Jána Bottu,
5. P.O. Hviezdoslav, Krvavé sonety,
6. Š. Krčméry, Ked sa sloboda rodila,
7. I. Krasko, Dielo,
8. J. Kostra, L'úbostné verše,
9. V. Turčány, Jarky v kraji,
10. Š. Žáry, Aká to vôňa,
11. A. Plávka, Sláva života,
12. J. Stacho, Svadobná cesta.

Die Daten und Resultate sind in Tabelle 3.1 zu finden. Außer Kostra, Plávka und Stacho zeigen alle Autoren eine Tendenz, das Reimwort mit einem Vokal zu beenden. Es wurde aber gezeigt, daß diese Tendenz zwischen 1840 bis 1960 eine Entwicklung in dem Sinne durchgemacht hat, daß sie linear abnahm und gegen 1960, als sich die Proportionen etwa ausgeglichen hatten, zeigte sich der Trend, reimlose Poesie zu schreiben (vgl. Štukovský, Altmann 1965/1966).

Tabelle 3.1

Häufigkeiten von offenen und von geschlossenen Reim- und Restwörtern bei 12 slovakischen Dichtern
(nach Štukovský, Altmann 1964)

Autor	Zahl der Reimwörter		Zahl der Restwörter			
	Vokal	Kons.	Vokal	Kons.	t	p
1	162	44	638	414	4.85	$6x10^{-7}$
2	128	28	570	312	4.27	$9.8x10^{-6}$
3	348	80	564	404	8.35	$3.5x10^{-17}$
4	175	32	603	379	6.35	10^{-10}
5	287	159	1202	1078	4.51	$3.2x10^{-6}$
6	179	76	544	325	2.23	0.0129
7	187	32	680	469	7.38	$8x10^{-14}$
8	132	62	656	321	0.24	0.4052
9	152	38	515	292	4.27	$9.8x10^{-6}$
10	187	79	566	410	3.64	0.0001
11	135	78	566	409	1.43	0.0764
12	91	75	573	375	-1.36	(0.0869)

3.3. Die graduelle Klimax

Eine andere Art positionaler Wiederholung ist die sogenannte Klimax (vgl. Groot 1946), bei der eine quantitativ ausdrückbare Eigenschaft einer Texteinheit im Rahmen einer höheren Einheit statistisch nachweisbar anwächst. Ist diese Eigenschaft (Variable) z.B. die Wortlänge und die höhere Einheit der Vers, dann äußert sich die graduelle Klimax dadurch, daß die Wortlänge in jeder nächsten Position des Verses zunimmt.

An dieser Stelle werden wir die lineare, die reduzierte und die exponentielle Klimax untersuchen.

3.3.1. Die lineare Klimax

Die lineare Klimax werden wir anhand der malayischen Volksquatrine, Pantuns genannt, untersuchen. Hier sei ein Beispiel eines Pantuns gegeben:

Anak běruk dikayu rěndang,
turun mandi didalam paya.
Hodoh buruk dimata orang,
Cantik manis dimata sahaya.

Mißt man die Wortlänge in Silbenzahl, so wird man in diesem Beispiel keine graduelle Klimax erkennen. Nimmt man jedoch eine zufällige Stichprobe von 250 (vgl. Altmann, Štukovský 1965) Versen aus einer Pantunsammlung (Wilkinson, Winstedt 1914), dann erhält man die Resultate, die in Tabelle 3.2 dargestellt sind.

Tabelle 3.2

Häufigkeiten (n_{ij}) von Wortlängen im malayischen Pantun

	Position im Vers x_i			
Länge des Wortes in Silben y_j	1	2	3	4
1	6	-	-	-
2	181	163	148	131
3	62	86	97	118
4	1	1	5	1
Durchschnittliche Länge $\bar{y}_i$	2.232	2.352	2.428	2.480

Wie man sieht, wächst die durchschnittliche Wortlänge mit der Position, die Frage ist nur, ob dieser Anstieg nicht zufällig ist. Um dies zu testen, berechnen wir die lineare Regression der Wortlänge, d.h., wir prüfen, ob der gefundene Trend einer Geraden

$$y = a + bx$$

folgt. In den folgenden Formeln ist x_i die Position ($i = 1,2,3,4$), $\bar{y}_i$ ist die durchschnittliche Wortlänge in Position i. Die Koeffizienten a und b schätzen wir aus den Daten wie folgt

$$b^* = \frac{\Sigma n_i (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\Sigma n_i (x_i - \bar{x})^2},$$

wegen $n_i = 250$ für $i = 1,\ldots,K$ ($K = 4$) bekommen wir jedoch

$$b^* = \frac{\Sigma (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$= \frac{\Sigma x_i \bar{y}_i - \Sigma x_i \Sigma \bar{y}_i / K}{\Sigma x_i^2 - (\Sigma x_i)^2 / K}$$

$$= \frac{K \Sigma x_i \bar{y}_i - \Sigma x_i \Sigma \bar{y}_i}{K \Sigma x_i^2 - (\Sigma x_i)^2} \qquad (3.8)$$

und

$$a^* = \bar{y} - b^* \bar{x}. \qquad (3.9)$$

In unserem Fall ergibt sich

$$\Sigma x_i = 1 + 2 + 3 + 4 = 10$$

$$\Sigma \bar{y}_i = 2.232 + \ldots + 2.480 = 9.492$$

$$\Sigma x_i \bar{y}_i = 1(2.232) + 2(2.352) + \ldots + 4(2.480) = 24.14$$

$$\Sigma x^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$$

$$K = 4$$

$$\bar{y} = \Sigma \bar{y}_i / K = 9.492/4 = 2.373$$

$$\bar{x} = \Sigma x_i / K = 10/4 = 2.5$$

$$b^* = \frac{4(24.14) - 10(9.492)}{4(30) - 10^2} = 0.082$$

$$a^* = 2.373 - 0.082(2.5) = 2.168.$$

Der Trend folgt also der Geraden $y^* = 2.168 + 0.082x$.

Uns interessiert lediglich die Frage, ob der Anstieg b signifikant größer als 0 ist. Zu diesem Zweck berechnen wir das Kriterium

$$t = \frac{b^*}{s_b}, \qquad (3.10)$$

wo t eine Studentsche Variable mit $\Sigma n_i - 2$ Freiheitsgraden ist, in unserem Fall also mit 4(250) - 2 = 998 Freiheitsgraden, so daß wir sie als Normalvariable betrachten können. Die Varianz von b schätzen wir folgendermaßen:

Sei

$$s_i^2 = \frac{1}{N-1} \sum_{j=1}^{K} (y_{ij} - \bar{y}_i)^2 n_{ij}$$

$$= \frac{1}{N-1} \left[\sum_{j=1}^{K} y_{ij}^2 n_{ij} - \frac{(\Sigma y_{ij} n_{ij})^2}{n_i} \right], \qquad (3.11)$$

wo n_{ij} die Häufigkeiten in Tabelle 3.2 sind. Weiter sei wegen $n_i = 250$ für i = 1,...,K

$$s^2 = \frac{\sum_{i=1}^{K} (n_i - 1) s_i^2}{\sum_i (n_i - 1)} = \frac{1}{4} \sum_{i=1}^{4} s_i^2 \qquad (3.12)$$

und

$$s_b^2 = \frac{s^2}{\sum_i n_i (x_i - \bar{x})^2}. \qquad (3.13)$$

Für i = 1 erhalten wir aus (3.11)

$$\sum_{j=1}^{4} y_{1j}^2 n_{1j} = 1^2(6) + 2^2(181) + 3^2(62) + 4^2 = 1304$$

$$[\Sigma_j y_{1j} n_{1j}]^2/250 = [1(6)+2(181)+3(62)+4(1)]^2/250 = 1245.456$$

$$s_1^2 = (1304 - 1245.456)/249 = 0.2351$$

usw. Daraus wird in unserem Fall

$$s^2 = \frac{s_1^2 + s_2^2 + s_3^2 + s_4^2}{4} = 0.2542 \quad \text{nach (3.12)}$$

und

$$s_b^2 = \frac{0.2542}{250(5)} = 0.000203 \quad \text{nach (3.13)}$$

$$s_b = 0.01426$$

$$t = \frac{0.082}{0.01426} = 5.75. \quad \text{nach (3.10)}$$

Dieser Wert ist hoch signifikant ($P < 5 \times 10^{-7}$) und zeigt, daß es im Pantun einen deutlichen Wortverlängerungstrend gibt.

3.3.2. Die reduzierte Klimax

Die unterschiedlichen Ausprägungen einer Variablen können jedoch auch so verteilt sein, daß sie sich nicht unbedingt im Rahmen des Wortes manifestieren, sondern in größeren Einheiten, deren Anzahl für die Berechnung der Regression nicht ausreicht, zum Beispiel in einem Halbvers, in einem Teilsatz usw. Auch in dem Fall ist es aber möglich, zu prüfen, ob ein Unterschied in der Ausprägung der Variablen besteht. Diesen Fall muß man schon aus dem Grunde in Betracht ziehen, weil im malayischen Pantun nicht alle Verse aus 4 Wörtern bestehen, so daß man für jede Verslänge eine separate Regression berechnen müßte.

Betrachten wir wieder eine zufällige Stichprobe von 25 Pantunzeilen, wie sie von Altmann und Štukovský (1965) ermittelt wurden (vgl. Tabelle 3.3, erste und zweite Spalte).

Beim ersten Verfahren testen wir in verbundenen Stichproben, beim paarweisen Vergleich die Abweichung der mittleren Differenz von Null (vgl. Sachs 1972: 242). Bezeichnen wir

$$d_i = x_{i2} - x_{i1}, \qquad (3.14)$$

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_i, \qquad (3.15)$$

Tabelle 3.3

Länge der Halbverse in Pantun

Zahl der Silben		Differenz		Typ der Differenz
erster Halbvers	zweiter Halbvers	d_i	d_i^2	
4	5	1	1	A
4	5	1	1	A
4	5	1	1	A
4	4	0	0	B
4	4	0	0	B
5	4	-1	1	D
4	5	1	1	A
4	5	1	1	A
5	5	0	0	B
4	5	1	1	A
4	5	1	1	A
4	5	1	1	A
4	5	1	1	A
4	5	1	1	A
4	4	0	0	B
4	5	1	1	A
6	5	-1	1	D
4	4	0	0	B
6	4	-2	4	D
3	5	2	4	A
4	5	1	1	A
5	5	0	0	B
4	5	1	1	A
4	5	1	1	A
4	5	1	1	A
	Σ	13	25	

$$s_d^2 = \frac{1}{N-1} \sum_i (d_i - \bar{d})^2$$

$$= \frac{1}{N-1}\left[\sum_i d_i^2 - \frac{(\Sigma d_i)^2}{N}\right], \qquad (3.16)$$

dann können wir einen t-Test durchführen, und zwar laut

$$t = \frac{\bar{d}}{s_d/\sqrt{N}}, \qquad (3.17)$$

wo t eine Studentsche Variable mit N − 1 Freiheitsgraden ist. Einfachheitshalber haben wir $d_i = x_{i2} - x_{i1}$ bezeichnet, man kann jedoch auch $x_{i1} - x_{i2} = d_i$ setzen. Die für die Berechnungen notwendigen Werte sind in der dritten und der vierten Spalte von Tabelle 3.3 aufgeführt. So erhalten wir

$$\bar{d} = \frac{13}{25} = 0.52 \qquad \text{laut (3.15)}$$

$$s_d^2 = \frac{1}{24}(25 - 13^2/25) = 0.76 \qquad \text{laut (3.16)}$$

$$t = \frac{0.52}{0.8718/\sqrt{25}} = 2.9823 .$$

Beim zweiseitigen Test mit 24 Freiheitsgraden entspricht dieses Resultat einem P = 0.006, was uns zu der Entscheidung führen kann, daß der zweite Halbvers tatsächlich länger ist als der erste.

Eine andere Methode ist McNemars Test für die Signifikanz der Veränderungen (vgl. Siegel 1956: 63–67). Hier betrachten wir nur die Richtung einer Differenz, nicht aber ihre Größe, und wählen folgende Bezeichnungen:

A wenn $x_{i2} > x_{i1}$ (d.h. alle positiven Zahlen in der dritten Spalte von Tab. 3.3)

D wenn $x_{i2} < x_{i1}$ (d.h. alle negativen Zahlen in der dritten Spalte von Tab. 3.3)

B wenn $x_{i2} = x_{i1}$ (d.h. alle Nullen in der dritten Spalte von Tab. 3.3).

Die entsprechenden Symbole sind in der vierten Spalte von Tabelle 3.3 angegeben. Den Test für die Signifikanz der Veränderung der Länge im zweiten Halbvers im Vergleich zum ersten führen wir nach der Formel

$$X^2 = \frac{(A - D - 1)^2}{A + D} \qquad (3.18)$$

durch. X^2 ist verteilt wie eine Chiquadrat-Variable mit 1 Freiheitsgrad. Für unsere Daten erhalten wir

$$A = 16$$

$$D = 3$$

$$B = 6$$

und daraus

$$X^2 = \frac{(16 - 3 - 1)^2}{16 + 3} = 7.58.$$

Dieses Resultat entspricht einem P = 0.0059, was mit den vorigen Resultaten fast identisch ist. In der Formel (3.18) haben wir im Zähler −1 als Korrektur für Kontinuität eingeführt. Läßt man diese Korrektur aus, dann bekommt man in unserem Beispiel bei beiden Tests identische Resultate.

3.3.3. Die exponentielle Klimax

Nicht nur Volkspoesie scheint eine Wortlängenklimax zu haben, sie kann auch in der künstlerischen Dichtung vorkommen, ja sogar in einer viel stärkeren Form. Die Untersuchung einiger slovakischer Gedichte zeigt, daß hier eine nichtlineare Regression vorhanden ist, wenn man hinreichend viele Verse durchzählt.

Betrachten wir die Wortlängen im Gedicht "Samota" von A. Sládković (1820–1872), in dem ein Vers höchstens 8 Wörter enthält. Wir bezeichnen die Positionen "rechtsbündig", d.h., in einem Vers wie

"Šľachetnosť pevne poobjíma dušu"

steht das letzte Wort in Position 8, das erste in Position 5, usw. Auf diese Weise erhalten wir die positionsbedingten durchschnittlichen Wortlängen, wie in der vierten Zeile von Tabelle 3.4 dargestellt.

Die beste Anpassung liefert die Exponentialkurve

$$y^* = ae^{bx} = 1.0807e^{0.0984x},$$

wo x die Position und y die durchschnittliche Wortlänge ist. Es ist nicht nötig, alle einzelnen Wortlängen in Betracht zu ziehen, die durchschnittlichen Werte reichen aus. Die Koeffizienten a und b kann man mit den Formeln der linearen Regression erhalten, indem man die Transformation

$$\ln y = \ln a + bx$$

$$Y = A + Bx$$

Tabelle 3.4

Durchschnittliche Wortlängen in Sládkovićs Gedicht "Samota"
(in Silben)

1 Position x	1	2	3	4	5	6	7	8
2 Zahl der Wörter in Position x	4	25	59	89	94	94	94	94
3 Gesamtlänge der Wörter	5	35	88	141	167	155	195	244
4 Durchschnittliche Länge der Wörter y	1.25	1.40	1.49	1.58	1.78	1.65	2.08	2.60
5 Berechnete Länge y*	1.19	1.32	1.45	1.60	1.77	1.95	2.15	2.37

durchführt und A und B wie oben in (3.8) und (3.9) berechnet, wobei man statt Y immer ln y setzt.

Einen Test führt man jetzt folgendermaßen durch: Man berechnet

$$SSR = \sum_i (\bar{Y} - Y_i^*)^2 , \qquad (3.19)$$

d.h. die Summe der quadrierten Abweichungen der berechneten Werte von dem Mittelwert der beobachteten Werte, die die "erklärte" Variabilität darstellt; weiter

$$SSE = \Sigma (Y_i - Y_i^*)^2 , \qquad (3.20)$$

d.h. die Summe der quadrierten Abweichungen der beobachteten Werte von den berechneten Werten, die die "nicht erklärte" Variabilität darstellt, und setzt beides in die Formel

$$F_{1,n-2} = \frac{SSR}{SSE/(n-2)} \qquad (3.21)$$

ein. Diese eine Größe ist wie eine F-Variable mit 1 und n - 2 Freiheitsgraden verteilt, wobei n die Zahl de Beobachtungen ist. Die Berechnung illustrieren wir an dem obigen Beispiel. Die notwendigen Zahlen findet man in Tabelle 3.5. Daraus ergibt sich

Tabelle 3.5

Test für die exponentielle Regression

x	y_i	y_i^*	$\ln y_i = Y_i$	$\ln y_i^* = Y_i^*$	$(\bar{Y}-Y_i^*)^2$	$(Y_i-Y_i^*)^2$
1	1.25	1.19	0.223146	0.173953	0.121160	0.002420
2	1.40	1.32	0.336472	0.277632	0.059732	0.003462
3	1.49	1.45	0.398776	0.371564	0.022641	0.000741
4	1.58	1.60	0.457425	0.470004	0.002707	0.000158
5	1.78	1.77	0.576613	0.570980	0.002396	0.000032
6	1.65	1.95	0.500775	0.667829	0.021256	0.027907
7	2.07	2.15	0.727549	0.765468	0.059261	0.001438
8	2.60	2.37	0.955511	0.862890	0.116183	0.008579
$\bar{Y}$ = 0.522033					0.405336	0.042316

$$F_{1,6} = \frac{0.405336}{0.042316/6} = 57.48$$

Ein so hoher F-Wert (P = 0.0003) deutet auf einen realen exponentiellen Trend. Es ist zu bemerken, daß die obigen Koeffizienten a und b nicht mit der obigen Methode, sondern mit Optimierung errechnet wurden, die eine zusätzliche Verbesserung der Anpassung gewährleistet.

Möglicherweise ist dieser Trend (Klimax) eine spezielle Art eines sekundären Rhythmus, dessen Träger ganze Wörter sind. Die Gerade oder die exponentielle Kurve sind nur die ersten Approximationen, da anzunehmen ist, daß dieser sekundäre Rhythmus sehr kompliziert sein wird. Möglicherweise wird er durch die rechtsbündige Zählung nur verdeckt.

Betrachten wir die Analyse des slowakischen Gedichts "Morho" von S. Chalupka, wo wir die Verse mit unterschiedlicher Zahl der Wörter separat berücksichtigt haben. Für die Verslängen 6, 7, 8 (Wörter) ergeben sich die durchschnittlichen Wortlängen wie in Tabelle 3.6 dargestellt. Andere Verslängen sind nicht häufig genug vertreten, um repräsentativ zu sein.

Tabelle 3.6

Durchschnittliche Wortlängen in einzelnen Verspositionen in Chalupkas Gedicht "Morho"

Vers-länge	Position 1	2	3	4	5	6	7	8	Zahl der Verse
6	1.75	1.97	2.18	1.83	2.59	2.66	-	-	29
7	1.58	1.64	1.77	1.84	1.81	1.92	2.42	-	64
8	1.29	1.37	1.37	1.92	1.61	1.63	1.13	2.35	51

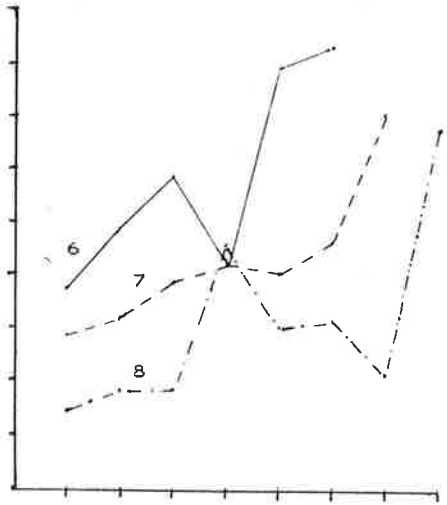

Abb. 3.1.

Man kann nun an die Daten eine Gerade oder eine Exponentialfunktion anlegen, aber dies würde die Tatsachen verschleiern. Wir sehen mehrere "Knicke" in den Daten. Bei den Verslängen 6 und 7 scheinen hier zwei Klimaxe zu sein (1-3 und 4-6 bzw. 1-4 und 5-7), bei der Länge 8 sind es 4, die irgendwie symmetrisch sind (1-2, 3-4, 5-6,7-8), wobei die

erste und die dritte kleiner, die zweite und die vierte steiler sind. Verbindet man sie jeweils mit einer Kurve, wie in Abb. 3.1, dann werden sie deutlicher. Die Bewegung ist sehr regelmäßig und deutet eine explodierende Schwingung an, deren Untersuchung noch nicht fortgeschritten ist.

3.4. Andere positionale Wiederholungen

Eine Textart kann den Autor zwingen, spezielle Verteilungen von (qualitativen) Texteinheiten auf bestimmte Positionen vorzunehmen. Dies ist leichter von der Grammatik her zu erklären, wenn die betreffenden Texteinheiten grammatische Kategorien darstellen. Besonders in nicht-poetischen Texten übt die Syntax einen starken Druck auf die Distribution der syntaktischen Einheiten im Satz aus, während man in poetischen Texten mit weniger starken grammatisch bedingten Tendenzen rechnen darf.

Die Untersuchung der Positionierung der Wortarten auf die i-te Stelle im Satz wurde von Průcha (1967) in die Wege geleitet. Seine Daten können hier jedoch nicht verwendet werden, da er nur die Proportionen angibt. Stattdessen werden wir zur Illustration die Positionierung der Nomina im "Erlkönig" betrachten.

Die Zeilen des "Erlkönig" enthalten 5 bis 9 Wörter (= Positionen). Die Verteilung der Nomina auf die einzelnen Positionen ist in Tabelle 3.7 dargestellt.

Tabelle 3.7

Häufigkeit der Nomina in einzelnen Positionen im "Erlkönig"

Verslänge	Position								
	1	2	3	4	5	6	7	8	9
5	1	2	–	–	1	–	–	–	–
6	1	7	1	1	2	6	–	–	–
7	–	–	6	2	3	–	2	–	–
8	–	2	–	3	–	1	–	3	–
9	–	3	–	2	–	–	–	–	2

Die Suche nach einem Trend stößt hier an zwei Schwierigkeiten: (a) Die Häufigkeiten sind zu klein, um daraus zuverlässige Schlüsse ziehen zu können, (b) die Verse sind unterschiedlich lang, so daß Position 5

eines fünfwortigen Verses nicht dasselbe ist wie die Position 5 in einem neunwortigen Vers. Um diese beiden Schwierigkeiten zu überwinden, verfahren wir folgendermaßen: Wir bilden relative Intervalle, für jede Verslänge separat. Die oberen Grenzen der Intervalle bestimmen wir als

$$\frac{\text{Position im Vers}}{\text{Zahl der Positionen im Vers}}$$

So erhalten wir für Verse der Länge 5 die oberen Grenzen als

1/5, 2/5, 3/5, 4/5, 5/5
oder 0.2, 0.4, 0.6, 0.8, 1.0,

für einen sechswortigen Vers als

1/6 2/6 3/6 4/6 5/6 6/6
oder 0.1667, 0.3333, 0.5000, 0.6667, 0.8333, 1.0000

usw. Die Intervalle des kürzesten Verses nehmen wir als Norm und erhalten dann

$\langle 0, 0.2\rangle$, $(0.2, 0.4\rangle$, $(0.4, 0.6\rangle$, $(0.6, 0.8\rangle$ $(0.8, 1.0\rangle$.

Diesen Intervallen ordnen wir auch die Häufigkeiten in anderen Verslängen zu.

So gehören z.B. die Häufigkeiten der Nomina, die in sechswortigen Versen auf die fünfte und die sechste Position entfallen, in das Intervall (0.8, 1.0). Dadurch erhalten wir eine neue Verteilung der Nomina auf 5 Positionen als

Relative Position	1	2	3	4	5	N
Häufigkeit	2	14	14	5	16	51

Wie man sieht, sind die Nomina nicht gleichmäßig verteilt. Wir überprüfen dies mit Hilfe der Informationsstatistik

$$2I = 2 \sum_{i=1}^{5} n_i \ln \frac{n_i}{E_i} \qquad (3.22)$$

was in Übereinstimmung mit der Struktur des Deutschen ist, wo vor dem Nomen meistens ein Artikel steht. Die Häufigkeit $n_1 = 3$ wäre nicht mehr signifikant, denn $\sum_{x=0}^{3} P_x = 0.09$.

Für die anderen Positionen können wir $N = 51$ nehmen. So ergibt sich mit $p = 0.2$ die untere Grenze als

$$x_u = 5 \qquad \text{mit} \quad \sum_{x=0}^{5} P_x = 0.046$$

und die obere

$$x_o = 16 \qquad \text{mit} \quad \sum_{x=16}^{51} P_x = 0.036$$

d.h. Positionen, in denen 5 oder weniger Nomina stehen sind "antinominal"; diejenigen, wo 16 oder mehr Nomina stehen, sind "nominal". Wie man sieht, gibt es drei Positionen im Erlkönig, die eine derartige "Wortarttendenz" zeigen, nämlich die erste, die vierte und die fünfte.

4. ASSOZIATIVE WIEDERHOLUNG

An dieser Stelle werden uns nicht die Assoziationen, wie sie in der Psychologie verstanden und an den Reaktionen von Versuchspersonen untersucht werden (vgl. Cramer 1968), interessieren, sondern diejenigen, die durch überdurchschnittlich häufiges gemeinsames Vorkommen zweier Wörter in einem Text entstehen.

Gemeinsames Vorkommen oder ***Koinzidenz*** bedeutet das Auftreten der beiden Wörter in einem textuellen ***Rahmen***, der unterschiedliche Maße haben kann. Der kleinste Rahmen kann der Satz (Teilsatz) oder Vers sein, größere sind Absatz, Strophe, Kapitel, Text, Genre, u.ä.

Überdurchschnittlich häufig bedeutet häufiger als erwartet, wobei diese Erwartung theoretisch berechnet werden muß, wie unten gezeigt werden soll.

Im Laufe der Zeit haben sich zahlreiche Varianten, Aspekte und Methoden der Erforschung der assoziativen Wiederholung herauskristallisiert, die hier nur dokumentarisch erfaßt werden können. Bereits Osgood (1959) bringt eine ganze Zahl von Forschungsproblemen, Methoden und Darstellungsarten, die später noch verfeinert wurden. Man hat nicht nur die assoziative Wiederholung zweier Wörter, sondern die ganzer Begriffssysteme untersucht und sprach in der Psychologie von der Assoziationsstruktur, während in der Linguistik von semantischen Feldern die Rede war.

Einige Autoren analysieren nur einzelne Texte (z.B. Berry-Rogghe 1973; Geffroy, Lafon, Seidel, Tournier 1973), andere verwenden gleichzeitig viele Texte (Rieger 1971, 1972, 1974; Dannhauer, Wickmann 1972). Der Rahmen, in dem die Koinzidenz vorkommt, erstreckt sich von einem Minimalrahmen (= Wortpaar) bis hin zu ganzen Texten; es wird außerdem auch noch linksseitige vs. rechtsseitige Koinzidenz unterschieden (vgl. z.B. Dolphin 1977).

So findet beispielsweise Rieger (1971, 1974) aufgrund von Gedichten aus den Jahren 1820-1840 ein semantisches Umgebungsfeld für das Wort "Blüte" mit einem von ihm vorgeschlagenen Abstand, wie in Tabelle 4.1 dargestellt.

Ähnliche Umgebungen ermitteln auch Geffroy, Lafon, Seidel und Tournier (1973) und stellen sie mit Hilfe verschiedener Graphen dar (vgl. Abb. 4.1).

Tabelle 4.1

Umgebungsfeld von "Blüte"
nach Rieger (1971, 1974)

Semantische Umgebung U(i;s) von i = Blüte, s = 4.50					
Frühling	2.768	Duft	3.412	Baum	3.339
Rose	3.435	Schön/ht	3.641	Lenz	3.598
Garten	3.788	Wiese/Aue	3.971	Vogel	3.859
Hold	3.983	Zärt/lich	3.995	Zweig/Ast	3.987
Berg,Gebirg	4.006	Gras/halm	4.030	Traum	4.028
Nachtigall	4.031	Wunder	4.050	Blume	4.042
Neu	4.084	Lust	4.119	Sonne	4.098
Blatt	4.120	Pracht	4.148	Winter	4.137
Liedweise	4.164	Wonne	4.306	Treu/e	4.290
Hügel	4.341	Feld/Gefild	4.365	Herz/en	4.362
Anmut	4.370	Märchen	4.398	Zeit	4.392
Quelle/n	4.398	Laub	4.410	Mai	4.399
Eiche	4.428	Hoffnung	4.439	Bach	4.432
Liebe/n	4.439	Silber/n	4.444	Leise	4.440
Land	4.457	Grün/en	4.484	Früh/e	4.482

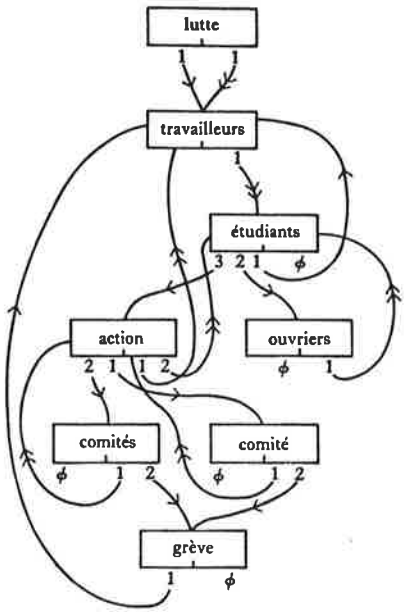

Abb. 4.1. Ein beschränkter Umgebungsgraph
nach Geffroy, Lafon, Seidel, Tournier (1973)

Dolphin (1977) stellt mit Hilfe eines von ihr entwickelten Maßes ein "Lexikogramm" für "yeux" in einem französischen Text (vgl. Abb. 4.2) auf.

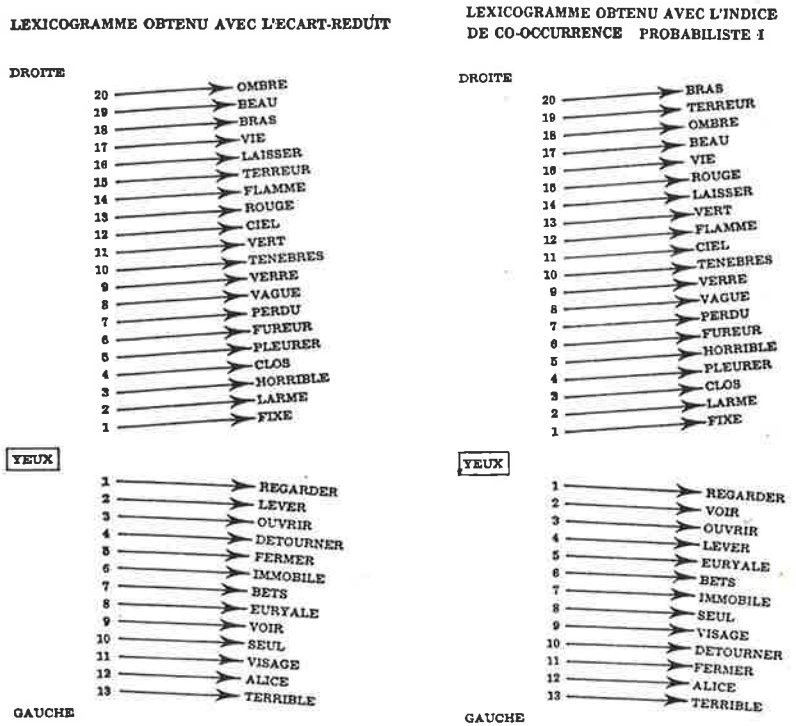

Abb. 4.2. Umgebungsfeld von "yeux"
nach Dolphin (1977)

4.1. Assoziative Wiederholung zweier Wörter

Wir beschränken uns hier auf eine elementare Überlegung, aus der sich eine leicht erweiterbare Methode ergibt.

Nehmen wir an, daß wir die Assoziation zweier Nomina A und B in einem Text T untersuchen. Unser Text sei Goethes Erlkönig, A sei "Vater", B sei "Erlkönig". Stellt man sich die Frage, ob "Vater" mit dem

"Erlkönig" assoziiert ist, dann muß erst der ***Rahmen*** bestimmt werden, in dem man die Assoziation mißt. Sei dieser Rahmen zunächst der Vers.

Kleine Texte

In Goethes "Erlkönig" können wir folgendermaßen verfahren: Wir stellen eine Tabelle der Vorkommen von "Vater" und "Erlkönig" in den 32 Versen auf; vgl. Tabelle 4.2, in der mit einem "+"-Zeichen das Vorkommen dieser zwei Wörter symbolisiert wird.

Tabelle 4.2

Vorkommen von "Vater" und "Erlkönig" in einzelnen Versen von Goethes "Erlkönig".

Vers Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Vater		+				+							+				
Erlkönig					+	+								+			

Vers Nr.	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Vater				+						+		+			
Erlkönig					+						+				

Wir finden nur eine einzige Koinzidenz im 6. Vers. Falls nur eine einzige Koinzidenz vorliegt, kann man eine Assoziation ruhigen Gewissens ausschließen. Intuitiv wissen wir aber, daß irgendeine Assoziation vorhanden sein sollte. Daher vergrößern wir den Rahmen auf jeweils zwei Verse (Halbstrophe). Jetzt bekommen wir ein Resultat, wie in Tabelle 4.3 dargestellt.

In 16 Halbstrophen treffen sich "Vater" und "Erlkönig" 4-mal, wobei "Vater" hier insgesamt 6-mal und "Erlkönig" 5-mal vorkommen (falls ein Wort zwei- oder mehrmals in einem Rahmen vorkommt, wird es nur einmal registriert). Hier ist es schon sinnvoll, nach einer Assoziation zu fragen.

Tabelle 4.3

Vorkommen von "Vater" und "Erlkönig" in Halbstrophen in Goethes Erlkönig.

Halbstrophe Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Vater	+		+			+					+			+		+
Erlkönig		+		+		+					+			+		

Wir suchen jetzt die Wahrscheinlichkeit, daß unter den gegebenen Bedingungen :

N = 16 Halbstrophen

M = 6 Vorkommen von "Vater"

n = 5 Vorkommen von "Erlkönig"

eine "Halbstrophenkoinzidenz" von "Vater" und "Erlkönig" vorkommt.

Die Wahrscheinlichkeit P(X = x) von x Koinzidenzen läßt sich folgendermaßen berechnen: Die Anzahl aller Möglichkeiten, n-Exemplare von "Erlkönig" und M-Exemplare von "Vater" jeweils auf N Halbstrophen zu verteilen, ist

$$\binom{N}{n}\binom{N}{M}.$$

Die Zahl der "günstigen" Fälle ergibt sich wie folgt:

Die x koinzidierenden Vorkommen kann man auf N Stellen auf $\binom{N}{x}$ Weisen verteilen; die verbleibenden n-x Vorkommen von "Erlkönig" kann man auf die N-x freien Stellen auf $\binom{N-x}{n-x}$ Weisen verteilen, und die verbleibenden M-x Vorkommen von "Vater" kann man auf die restlichen N-n Stellen auf $\binom{N-n}{M-x}$ Weisen verteilen. Die Zahl der günstigen Fälle ist

$$\binom{N}{x}\binom{N-x}{n-x}\binom{N-n}{M-x},$$

woraus sich die gesuchte Wahrscheinlichkeit als

$$P(X = x) = \frac{\binom{N}{x}\binom{N-x}{n-x}\binom{N-n}{M-x}}{\binom{N}{n}\binom{N}{M}} \qquad (4.1)$$

ergibt. Ordnet man die Fakultäten etwas um, dann bekommt man leicht

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = 0,1,\ldots,\min[n,M], \qquad (4.2)$$

worin man die ***hypergeometrische Verteilung*** erkennt. Da wir aber das Ereignis in (4.2) oder ein noch extremeres Ereignis suchen, bekommen wir

$$P(X \geq x_c) = \sum_{x=x_c}^{\min[n,M]} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} . \qquad (4.3)$$

In unserem Beispiel gab es

$$N = 16,\ M = 6,\ n = 5,\ x_c = 4,$$

so daß man erhält

$$P(X \geq 4) = \frac{\binom{6}{4}\binom{16-6}{5-4}}{\binom{16}{5}} + \frac{\binom{6}{5}\binom{16-6}{5-5}}{\binom{16}{5}}$$

$$= \frac{\binom{6}{4}\binom{10}{1}}{\binom{16}{5}} + \frac{\binom{6}{5}\binom{10}{0}}{\binom{16}{5}}$$

$$= \frac{15(10)}{4368} + \frac{6(1)}{4368}$$

$$= 0.0357.$$

Diese Wahrscheinlichkeit ist kleiner als 0.05, daher könnten wir eine Tendenz zur Assoziation annehmen. Jedoch ist diese Assoziation nicht "erster Ordnung", weil die Koinzidenzen nicht im Minimalrahmen (Vers) vorkommen, sondern etwa "zweiter Ordnung" (Halbstrophe).

Die Stärke der Assoziation sollte nicht mit der Rahmengröße, die eine andere Dimension darstellt, vermischt werden, obwohl sie in einer einzigen Charakteristik kombiniert werden können. Eine zweidimensionale Darstellung wäre geeigneter.

Die Berechnung von (4.3) läßt sich manchmal etwas vereinfachen, indem man es als

$$P(X \geq x_c) = 1 - P(X < x_c)$$

$$= 1 - \sum_{x=0}^{x_c-1} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (4.4)$$

darstellt. Der erste Wert der Summe rechts ergibt

$$P(X = 0) = \frac{\binom{M}{0}\binom{N-M}{n-0}}{\binom{N}{n}}$$

$$= \frac{(N-M)(N-M-1)\ldots(N-M-n+1)}{N(N-1)\ldots(N-n+1)}, \qquad (4.5)$$

die anderen lassen sich mit der Rekursionsformel

$$P(X = x) = \frac{(M-x+1)(n-x+1)}{x(N-M-n+x)} P(X = x-1) \qquad (4.6)$$

ermitteln.

Große Texte

Ist der Text (d.h. seine Länge N) groß, so ist das Rechnen mit der hypergeometrischen Verteilung umständlich und langwierig, auch wenn man sich mit der Rekursionsformel behelfen kann. In solchen Fällen nutzt man die Tatsache, daß die hypergeometrische Verteilung unter bestimmten Bedingungen gegen die Poisson-Verteilung konvergiert, und verfährt folgendermaßen:

Sei p_A die Wahrscheinlichkeit des Vorkommens des Wortes A in einer Gesamtheit; man kann sie durch die relative Häufigkeit $p_A^* = n_A/N$ schätzen.

Sei p_B bzw. p_B^* die Wahrscheinlichkeit bzw. die relative Häufigkeit des Wortes B. Diese relativen Häufigkeiten kann man auch aus dem gegebenen (langen) Text ermitteln; wir nehmen an, daß sie sehr klein sind.

Sei weiter N die Anzahl der Rahmen, in denen die Koinzidenz von A und B untersucht wird. Dann ist unter der Hypothese der Unabhängigkeit von A und B die Wahrscheinlichkeit ihrer Koinzidenz gleich $p_A p_B$, und die erwartete Anzahl der Rahmen, in denen A und B gemeinsam vorkommen, ist

$$Np_A p_B = a. \tag{4.7}$$

Die beobachtete Zahl der Koinzidenzen (x_K) betrachten wir nun als Poisson-verteilt und ziehen folgende Schlüsse:

(1) Wenn $x_K > a$ und

$$P(X \geq x_K) = \sum_{x=x_K}^{\infty} \frac{e^{-a} a^x}{x!} \leq 0.05, \tag{4.8}$$

dann betrachten wir die Assoziation als signifikant, d.h. die Koinzidenz ist ***assoziativ***.

(2) Wenn $x_K < a$ und

$$P(X \leq x_K) = \sum_{x=0}^{x_K} \frac{e^{-a} a^x}{x!} \leq 0.05, \tag{4.9}$$

dann bezeichnen wir die Koinzidenz als ***dissoziativ***.

(3) In allen anderen Fällen betrachten wir die Koinzidenz als ***neutral***.

Beispiel. Betrachten wir als Beispiel fünf Gedichte an Laura von Schiller (Phantasie an Laura; Laura am Klavier; Entzückung an Laura; Das Geheimnis der Reminiszenz; Melancholie).

In

$$N = 117$$

Sätzen kommt "Tod" in

$$n_{Tod} = 7$$

Sätzen vor, "Leben" in

$$n_{Leben} = 9$$

Sätzen, und gemeinsam treten sie in

$$x_K = 2$$

Sätzen auf. Wir schätzen

$$p^*_{Tod} = \frac{7}{117} = 0.0598$$

$$p^*_{Leben} = \frac{9}{117} = 0.0769,$$

und die erwartete Anzahl der Koinzidenzen ergibt sich als

$$a = Np_{Tod}p_{Leben} = 117(0.0598)0.0769 = 0.5385.$$

Da $x_K > a$, benutzen wir die Formel (4.8) und berechnen

$$P(X \geq 2) = \sum_{x=2}^{\infty} \frac{e^{-a} a^x}{x!}$$

$$= 1 - \sum_{x=0}^{1} \frac{e^{-a} a^x}{x!}$$

$$= 1 - e^{-a}\left[\frac{a^0}{0!} + \frac{a^1}{1!}\right]$$

$$= 1 - e^{-0.5385}(1 + 0.5385)$$

$$= 0.1021.$$

Da dieser Wert größer als 0.05 ist, handelt es sich um eine neutrale Koinzidenz.

Für die Wörter "Wange" und "Blut" haben wir

$$n_{Wange} = 6$$

$$n_{Blut} = 3$$

$$x_{\cap} = 2.$$

Daher ist

$$a = 6(3)/117 = 0.1538$$

und

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - P_0 - P_1$$

$$= 1 - e^{-0.1538}(1 + 0.1538)$$

$$= 0.0107.$$

Da $P(X \geq 2) < 0.05$, schließen wir, daß zwischen "Wange" und "Blut" eine positive Assoziation besteht. (Die exakte Wahrscheinlichkeit ist in diesem Falle 0.0065).■■■

Bei der assoziativen Analyse ergeben sich einige qualitative Probleme, die wir hier zumindest andeuten wollen.

(1) Lemmatisierung

Arbeitet man mechanisch (d.h. mit dem Computer), dann muß man die Texte erst lemmatisieren, sonst sind z.B. "Erlkönig", "Erlenkönig" und "Erlkönigs" drei unterschiedliche Wörter. Es gibt bereits fertige Lemmatisierungsprogramme fürs Deutsche bzw. lemmatisierte maschinenlesbare Texte.

(2) Komposition

Man muß entscheiden, ob man ein "Wort" auch in einem Kompositum erkennen will oder nicht, denn je nachdem, wie man verfährt, können sich die Verhältnisse etwas verändern. In der Kurzgeschichte "Die ruhelose Kugel" von K. Kusenberg, die wir unten analysieren werden, gibt es folgende Wörter: "Kugel", "Kugelschütze", "Kugelhascher", "Höllenkugel", "Schütze", "Schützenverein".

Soll man hier "Kugelschütze" als distinktes Wort nehmen, wenn hier "Kugel" und "Schütze" am stärksten assoziiert sind?

(3) Synonyme

Wenn es nicht direkt um Wörter geht, sondern um Begriffe, soll man dann z.B. "schießen" und "feuern" als separate Wörter oder als identische Begriffe betrachten? Ebenso bezeichnen im genannten Text "Kugel" und "Geschoß" dasselbe.

(4) Versteckte Begriffe

Manche Lexeme fallen aus der Untersuchung völlig heraus, da sie nur in Komposita vorkommen. So gibt es etwa im oben genannten Text "Hexenblut" und "Hexenkugel", aber keine "Hexe". Man muß entscheiden, ob man sie in Betracht zieht oder nicht.

(5) Schlüsselwörter

Aus der Analyse kann man alle Wortarten außer Nomina, Verben und Adjektiven ausschließen, auch Modalverben oder Verben, die in Phrasen keine besondere Bedeutung haben (z.B. "zum Fall bringen").

(6) Homonyme

Wie soll man Homonyme wie "Lauf der Pistole" und "Lauf der Kugel" bewerten?

4.2. Darstellung

Nachdem man die einzelnen Assoziationen ermittelt hat, hat man zahlreiche Wege, um ein Gesamtbild der Assoziationsstrukturen graphisch

darzustellen. Hier werden wir nur eine Möglichkeit zeigen, nämlich den Minimalgraph.

Da die Assoziationsstärken durch Wahrscheinlichkeiten angegeben sind, die im Intervall <0,α>, d. h. <maximale Assoziation, minimale Assoziation> liegen, ist es empfehlenswert, sie so zu transformieren, daß sie zwischen 0 für minimale und 1 für maximale Assoziation liegen. Dies erreichen wir z.B. einfach durch

$$A_s(W_1, W_2) = 1 - \frac{P_{berechnet}}{\alpha} \quad . \qquad (4.10)$$

So ergibt sich aus der Wahrscheinlichkeit P = 0.04, wenn wir α = 0.05 wählen,

$$A_s = 1 - \frac{0.04}{0.05} = 0.20,$$

aus P = 0.004 wird

$$A_s = 1 - \frac{0.004}{0.05} = 0.92,$$

aus P = 0.0004 wird

$$A_s = 1 - \frac{0.0004}{0.05} = 0.99.$$

Wählt man eine andere Signifikanzgrenze als 0.05, dann muß man natürlich diese in den Nenner von (4.10) setzen. Die graphische Darstellung wird dann dementsprechend eine andere Gestalt annehmen, wie in Abb. 4.3 und 4.4 ersichtlich. Es gibt natürlich zahlreiche andere Möglichkeiten der Normierung.

Nachdem man die Assoziationswerte auf diese oder eine andere Weise ermittelt hat, ist es empfehlenswert, sie in eine symmetrische Matrix einzutragen, mit der man auch manuell arbeiten kann.

4.3. Der Minimalgraph

Bei der Darstellung des Assoziationsnetzes als Minimalgraph werden nur die stärksten Assoziationen der Wörter als Graphenkanten dargestellt. Man fängt bei einem beliebigen Wort A an und sucht dasjenige Wort B auf, mit dem A die größte Assoziation hat (es können auch mehrere sein). Man verbindet beide mit einem Pfeil (gerichtete Kanten) von A nach B. Dies ist der Anfangsbaum. Diesem Baum fügt man mit einer Kante das nächste meistassoziierte Wort an und wiederholt diese Prozedur solange, bis alle Wörter in einem Graph verbunden sind.

In der folgenden Analyse, wo es nur um eine Demonstration geht, werden wir folgendermaßen verfahren:

(1) Der Text wird lemmatisiert:

(2) Schlüsselwörter wie "Kugel", "Schütze", "Hexe" werden auch in den Komposita erkannt.

(3) Eine Begriffsidentifikation von Synonymen wird nicht durchgeführt.

(4) Komposita werden nur bei den Schlüsselwörtern zerlegt.

(5) Nur Nomina, Verben und Adjektive werden in Betracht gezogen; modale Verben werden ausgeschlossen.

(6) Homonyme werden mechanisch als jeweils ein Lemma betrachtet.

Wir nehmen an, daß diese Regelung zu keiner Verzerrung führen wird.

In dem Text "Die ruhelose Kugel" von K. Kusenberg, der N = 48 Sätze enthält, werden folgende Autosemantika, die mindestens 2-mal vorkommen, auf Assoziationen untersucht (die Zahl hinter dem Wort zeigt, in wie vielen Sätzen das Wort vorkommt):

Kugel	*18*	*Lauf*	*2*
Geschoß	*11*	*Welt*	*2*
Mann	*5*	*Kraft*	*2*
Schütze	*5*	*Stadt*	*2*
Schuß	*5*	*Spiel*	*2*
Bahn	*4*	*Zufall*	*2*
Garten	*3*	*Ehepaar*	*2*
Pistole	*3*	*Bildnis*	*2*
Mensch	*3*	*Hindernis*	*2*
Ziel	*2*	*Postkarte*	*2*
Zeit	*2*	*Leuchtturmwärter*	*2*

geraten	6	*handeln*	2	*groß*	5
bringen	6	*abfeuern*	2	*seltsam*	2
wissen	3	*befinden*	2	*hoch*	2
sitzen	2	*schicken*	2	*vereinzelt*	2
stehen	2	*anrichten*	2	*schwer*	2
halten	2	*geschehen*	2		
fliegen	2	*ausbleiben*	2		

Die Assoziationswahrscheinlichkeiten wurden mit Hilfe der hypergeometrischen Verteilung (Formeln 4.4–4.6) berechnet. Als assoziiert galten die Wörter, die höchstens mit P = 0.09 zusammen vorkamen, d.h., α = 0.09. Die ermittelten Assoziationen nach (4.10) sind in Tabelle 4.4 aufgeführt. Der Text war zu kurz, so daß mehrere Assoziationen gleich ausgefallen sind. In der Tabelle sind die Wörter wie folgt angeordnet:

1. vereinzelt	*14. handeln*	*27. Stadt*
2. halten	*15. ausbleiben*	*28. bringen*
3. hoch	*16. befinden*	*29. Kraft*
4. Spiel	*17. groß*	*30. Mensch*
5. Postkarte	*18. sitzen*	*31. Mann*
6. Bildnis	*19. Lauf*	*32. geraten*
7. Welt	*20. seltsam*	*33. Hexe*
8. Hindernis	*21. schießen*	*34. stehen*
9. wissen	*22. Ziel*	*35. Bahn*
10. schwer	*23. Zeit*	*36. Geschoß*
11. Zufall	*24. fliegen*	*37. abfeuern*
12. Schuß	*25. Schütze*	
13. Leuchtturmwärter	*26. Pistole*	

Der Minimalgraph zeichnet sich dadurch aus, daß es in ihm von jeder Ecke (Wort) zu jeder anderen Ecke höchstens einen einzigen Weg gibt. In dem Falle, daß es viele gleiche Assoziationen gibt, kann der Graph verschiedene Formen annehmen. In solchen Fällen ist es vielleicht besser, nicht den Minimalgraphen, sondern alle Kanten zur Darstellung zu benutzen oder das Assoziationskriterium strenger zu fassen. In Abbildung 4.3 findet man den Minimalgraphen des Textes mit α = 0.09, in Abbildung 4.4 mit α = 0.05.

Tabelle 4.4

Assoziationsmaße für Wörter aus Kusenberg

2	3	4	5	6	7	8	9	10	11	13	16	19	20	21	23	24	25	26	28	29	30	31	32	33	34	
.08	.08	.08	.08	.08																						1
			.08		.08	.08	.08																			2
		.08																								3
								.08	.08																	4
				.08						.08																5
					.08						.08															9
								.08				.08														11
		.99							.08																	12
			.08	.08																						14
						.08									.08	.08										15
									.90			.90					.14	.72								17
									.08			.08		.08												18
										.08	.08								.85							20
										.08			.08													21
										.08			.08	.08												22
															.08					.08						23
											.08										.08			.08	.08	27
															.08											30
																.14										31
																	.58									32
																								.08		34
																			.21							35
																				.46			.79			36
																						.90				37

4.4. Ausblick

In den Texten wird man zwei Arten von Assoziationen finden:

(a) ***Allgemeine***, wie man sie auch in der täglichen Sprache findet. Sie sind zum Assoziationsrepertoire der Sprachgemeinschaft geworden, bilden mehr oder weniger feste "Umgebungen" im Sinne von Rieger, und Telle von ihnen lassen sich auch in den von Psychologen verfertigten Assoziationsbüchern finden (vgl. z.B. Palermo, Jenkins 1964).

(b) ***Spezielle***, die dem Text eigen sind. Die Unterscheidung ist nicht kategorisch, sondern graduell, und es wäre mühsam zu untersuchen, wie sich eine Assoziation vom Einzeltext zum Gemeingut durchsetzt.

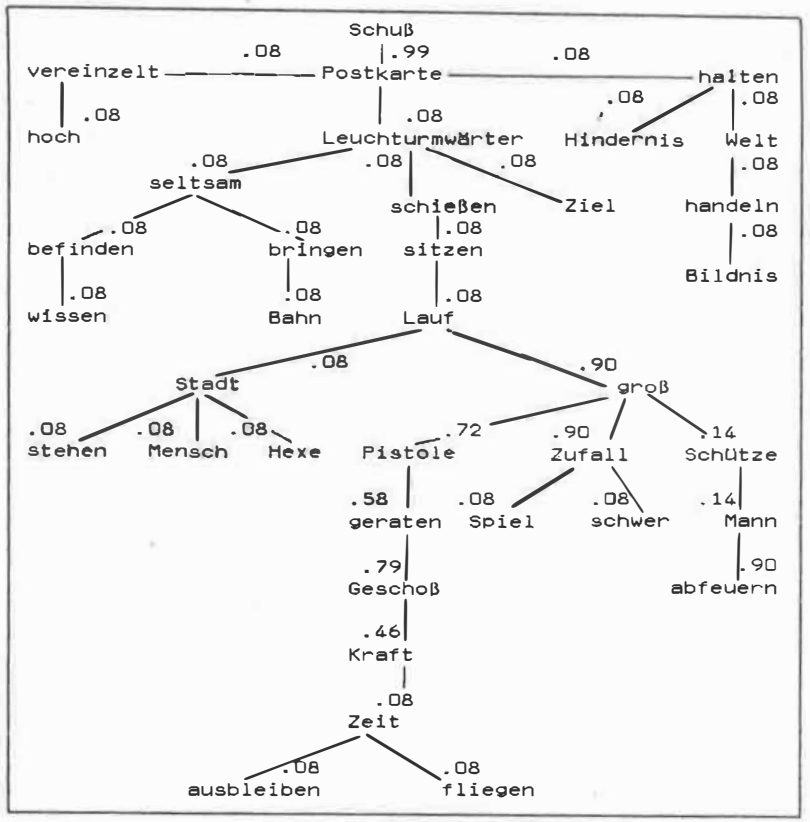

Abbildung 4.3. Minimalgraph mit α = 0.09

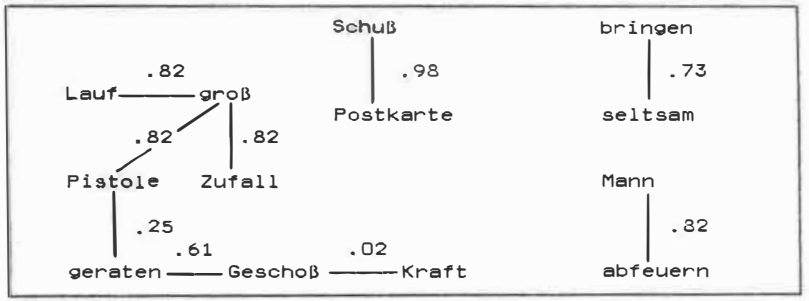

Abbildung 4.4. Minimalgraph mit α = 0.05

Die allgemeinen Assoziationen sind keineswegs Bildungen ad hoc, da sie eine signifikante Stärke aufweisen müssen. Sie entstehen durch Prozesse, deren Untersuchung zur Entdeckung von Gesetzen führen kann, denn nicht alles assoziiert sich mit allem in der Sprache. Wir vermuten, daß auch für den geschichtlichen Übergang einer speziellen Assoziation zu einer allgemeinen das Piotrowski-Gesetz gilt (vgl. Best, Kohlhase 1983), jedoch wäre diese Untersuchung besonders mühsam.

In Texten finden wir einen Teil der allgemeinen Assoziationen, die restlichen signifikanten Funde sind textbedingt (Thema, Textsorte), haben bestimmte Stärken, erscheinen in bestimmten Reihenfolgen, bilden ganz bestimmte Netze, die sich irgendwie regulär erweitern usw. Bisher gibt es nicht einmal Hypothesen über diese Erscheinungen, geschweige denn Gesetze.

Große Textmengen müssen untersucht werden, damit wir überhaupt an die Schwelle dieses Phänomens gelangen können. Wahrscheinlich wird eine Kooperation von Psychologen und Linguisten in verschiedenen Sprachen nötig sein, um von der heutigen deskriptiven Stufe aus ein wenig tiefer zu gelangen.

5. ITERATIVE WIEDERHOLUNG

Unter einer Iteration verstehen wir eine ununterbrochene Folge gleicher Elemente in einer Reihe von unterschiedlichen Elementen. So enthält die folgende Reihe von Buchstaben

AA BBB A BB

4 Iterationen. In Texten sind Iterationen vor allem im formalen Bereich möglich. So gibt es Folgen von Wörtern, die die gleiche Länge haben können, oder Folgen von Sätzen mit der gleichen Struktur, aber Folgen von Elementen, die die gleiche Bedeutung haben, sind eher verboten.

Iterationen können in solchen Sequenzen entstehen, in denen Elemente von mindestens zwei Arten vorkommen. Bei mehr als zwei Arten von Elementen werden die Formeln sehr umfangreich und die Arbeit mit ihnen umständlich.

In der Poetik wurde die Theorie der Iterationen wohl zum ersten Mal von Woronczak (1961) angewendet, später benutzte sie Fucks (1970); ihre Anwendungsmöglichkeiten in der Textanalyse wurden ausführlich von Grotjahn (1979, 1980) behandelt.

Untersucht man einen Text, dessen Elemente man dichotomisiert, d.h. in zwei Klassen unterteilt hat, z.B. "hypotaktische Sätze" und "alle anderen Sätze", dann kann man nicht nur fragen, wie oft diese zwei Klassen vorkommen, sondern auch, ob ihre Reihenfolge hintereinander zufällig ist. So ist eine Sequenz

AAABBB

anders als die Sequenz

ABABAB

oder

BAABBA ,

obwohl alle jeweils drei A und drei B enthalten, und der dritten Sequenz würden wir eher eine Zufälligkeit beimessen als den ersten beiden.

Es ist aber plausibel, zu fragen, welche von diesen Sequenzen eine zufällige Reihenfolge von Buchstaben aufweist bzw. in welcher ein Trend verborgen ist. Unter diesem Gesichtspunkt untersuchte Fucks (1968, 1970, 1971) Sequenzen von langen und von kurzen Sätzen, Grotjahn (1980) Sequenzen von betonten und von unbetonten Silben im Gedicht, Sequenzen von Verslängen, gemessen in Silbenzahl (vgl. auch Woronczak 1961), Sequenzen von langen und von kurzen Silben in Prosa und Poesie, Sequenzen von Vokalen und Konsonanten sowie Sequenzen von rhythmischen Mustern im Hexameter. Es ergibt sich hier ein breites Untersuchungsfeld, aber mit der Iterationstheorie kann nur ein Teil der Probleme der sequenziellen Wiederholungen beantwortet werden.

5.1. Binäre Sequenzen

Jede Folge von sprachlichen Einheiten läßt sich dichotomisch darstellen, wenn man die untersuchte Einheit (A) gegen alle anderen ($\bar{A}$) setzt. Handelt es sich um qualitative Variablen, so ist dies einfach, auch wenn es Grenzfälle gibt, wie z.B. in der Dichotomie Vokal:Konsonant, wo man Halbvokale bzw. Gleitlaute durch Entscheidung in eine dieser Klassen einordnet. Bei quantitativen Variablen nimmt man als Trennpunkt den Mittelwert oder den Median.

Betrachten wir die Zahl der Silben in einzelnen Versen des "Erlkönigs", wie sie von Grotjahn (1979:144) ausgezählt wurden, (s.Tab. 5.1). Die durchschnittliche Silbenzahl ist $\bar{x} = 9.625$. Betrachten wir

Tabelle 5.1

Zahl der Silben in Versen im "Erlkönig"

Silbenzahl x	8	9	10	11	12
Häufigkeit f_x	2	16	7	6	1

also jeden Vers, der 9 oder weniger Silben enthält, als kurz (K) und jeden, der 10 oder mehr Silben enthält, als lang (L), dann ergibt sich diese spezielle Struktur des "Erlkönig" als Folge

KKKK L KKKKKK LLL KKK LLLL KK LLLL K LL KK (5.1)

Nun wählen wir folgende Bezeichnungen:

n_1 = Zahl der Elemente der ersten Art, hier Zahl von K

n_2 = Zahl der Elemente der zweiten Art, hier Zahl von L

r_1 = Zahl der Iterationen von K

r_2 = Zahl der Iterationen von L

$n = n_1 + n_2$

$r = r_1 + r_2$

Die in (5.1) voneinander abgetrennten Folgen stellen die einzelnen Iterationen dar. So ist

$n_1 = 18$

$n_2 = 14$ $\qquad n = 32$

$r_1 = 6$

$r_2 = 5$ $\qquad r = 11$

Unsere Frage lautet: Gibt es im Gedicht irgendeine Tendenz, die Folgen von kurzen und langen Versen zu gestalten, oder ist die beobachtete Reihenfolge als zufällig zu betrachten?

Die Frage läßt sich noch etwas spezifizieren, wenn wir fragen, ob es eine Tendenz nach zu vielen oder nach zu wenigen Iterationen gibt. Je nachdem, wie wir die Frage stellen, bekommen wir im spezifizierten Fall eine einseitige, im nicht-spezifizierten Fall eine zweiseitige Hypothese.

Stellt man also a priori die Hypothese auf, daß der Verfasser einem kurzen Vers lieber einen kurzen folgen läßt und einem langen einen langen, dann bedeutet dies, daß man wenige Iterationen erwartet, denn die kurzen Verse klumpen sich miteinander und die langen ebenfalls. In diesem Falle fragt man, wie groß die Wahrscheinlichkeit ist, daß man die beobachtete oder eine noch kleinere Anzahl der Iterationen findet, d.h., man berechnet

$$P(R \leq r) \quad , \tag{5.2}$$

wo R die Variable "Iterationsanzahl" bedeutet.

Nimmt man aber a priori an, daß es eine recht reguläre Abwechslung der kurzen und der langen Verse gibt, dann fragt man, wie groß die Wahrscheinlichkeit ist, daß man die beobachtete oder eine noch größere Anzahl findet, d. h.

$$P(R \geq r). \tag{5.3}$$

Man erspart sich etwas Arbeit, wenn man sich diese Fragen "nicht ganz a priori" stellt, denn ist die beobachtete Anzahl der Iterationen kleiner als ihre mathematische Erwartung, dann erhält man für (5.3) eine große Wahrscheinlichkeit, ist hingegen ihre Zahl größer als ihre mathematische Erwartung, dann erhält man für (5.2) auch eine große Wahrscheinlichkeit. Daher pflegt man erst die mathematische Erwartung zu berechnen, und zwar nach der Formel

$$E(R) = 1 + \frac{2n_1n_2}{n_1+n_2} = \frac{2n_1n_2 + n}{n} . \tag{5.4}$$

In unserem Beispiel ist

$$E(R) = \frac{2(18)14 + 32}{32} = 16.75$$

Da $r = 11 < E(R)$, ist es sinnvoll, nur danach zu fragen, ob die Folge überhaupt zufällig ist oder ob $P(R \leq r)$ kleiner ist als eine vorgegebene Zahl, z.B. 0.05, die man als das Signifikanzniveau betrachtet.

Betrachten wir erst den zweiten Fall und berechnen (5.2). Die Wahrscheinlichkeitsfunktion von R ergibt sich als (vgl. Mood 1940; Gibbons 1971; Grotjahn 1980)

$$P(R = r) = \begin{cases} \dfrac{2\binom{n_1 - 1}{r/2 - 1}\binom{n_2 - 1}{r/2 - 1}}{\binom{n}{n_1}}, & \text{für gerade } r \quad (5.5) \\[2ex] \dfrac{\binom{n_1 - 1}{(r-1)/2}\binom{n_2 - 1}{(r-3)/2} + \binom{n_1 - 1}{(r-3)/2}\binom{n_2 - 1}{(r-1)/2}}{\binom{n}{n_1}} & \text{für ungerade } r \quad (5.6) \end{cases}$$

$$P(R=2)=\frac{2n_1!n_2!}{n!}$$

$$P(R=r)=\frac{n-r+1}{r-1}P_{r-1} \quad \text{ungerades r}$$

$$P(r=r)=\frac{2(n_1-\frac{r}{2}+1)(n_2-\frac{r}{2}+1)}{(\frac{r}{2}-1)(n-r+2)}P_{r-1} \quad \text{gerades r}$$

was erleichtern (vgl. Grotjahn [illegible]).
Iterationen als

$$r = 2k$$

und rechnen

$$P(R = 2k+1) = \frac{n-2k}{2k} P(R = 2k) \qquad \text{für ungerade r} \qquad (5.7)$$

bzw.

$$P(R = 2k) = \frac{2(n_1-k+1)(n_2-k+1)}{(k-1)(n-2k+2)} P(R=2k-1) \qquad \text{für gerade r} \qquad (5.8)$$

Für r = 3 ergibt sich nach (5.7) mit k = 1

$$P(R = 3) = \frac{32-2}{2} 0.0000000042 = 0.000000063.$$

Für r = 4 folgt nach (5.8) mit k = 2

$$P(R = 4) = \frac{2(18-2+1)(14-2+1)}{1(18-4+2)} 0.000000063$$

$$= 0.00000174.$$

Für r = 5 nach (5.7) mit k = 2

$$P(R = 5) = \frac{32-4}{4} 0.00000174 = 0.00000609$$

usw. für die anderen Wahrscheinlichkeiten

$$P(R = 6) = 0.00004177$$

$$P(R = 7) = 0.000181$$

$$P(R = 8) = 0.000765$$

$$P(R = 9) = 0.002297$$

$$P(R = 10) = 0.006700$$

$$P(R = 11) = 0.014741.$$

Die Summe dieser Wahrscheinlichkeiten ergibt

$$P(R \leq 11) = \sum_{r=2}^{11} P(R = r)$$

$$= 0.0000000042 + 0.000000063 +...+ 0.0067 + 0.014741$$

$$= 0.0247.$$

Da $P(R \leq 11) < 0.05$, schließen wir, daß im "Erlkönig" eine Tendenz besteht, gleichen Verslängen gleiche folgen zu lassen.

Wäre $r > E(R)$, dann könnte man die Hypothese von zu vielen Iterationen prüfen und müßte

$$P(R \geq r) = \sum_{x=r}^{2n_1+1} P(R = x), \qquad (n_1 < n_2) \qquad (5.9)$$

berechnen, wobei man die Rekursionsformeln (5.7) und (5.8) umgewandelt benutzen könnte, nämlich

$$P(R = 2k) = \frac{2k}{n-2k} P(R = 2k+1), \text{ für gerade r} \qquad (5.10)$$

$$P(R = 2k-1) = \frac{(k-1)(n-2k+1)}{2(n_1-k+1)(n_2-k+1)} P(R = 2k),$$

für ungerade r. (5.11)

Zur Illustration berechnen wir einen einfacheren Fall, in dem

$$n_1 = 5, \quad n_2 = 7, \quad n = 12$$

ist, für verschiedene r. Es kann hier höchstens $2n_1 + 1 = 11$ Iterationen geben, also berechnen wir P(R = 11). Da r ungerade ist, haben wir (5.6) zu rechnen, wo der erste Teil ausfällt, wegen

$$\binom{n_1 - 1}{(r-1)/2} = \binom{4}{5} = 0,$$

da Binomialkoeffizienten $\binom{n}{m} = 0$ gesetzt werden, wenn $m > n$. Daher ist

$$P(R = 11) = \frac{\binom{4}{4}\binom{6}{5}}{\binom{12}{5}} = \frac{6}{792} = 0.0076.$$

Gäbe es also r = 11 Iterationen unter den obigen Bedingungen, dann könnten wir sagen, daß es signifikant viele sind, weil P(R = 11) < 0.05. Für P(R = 10) erhalten wir mit k = 5 nach (5.10)

$$P(R = 10) = \frac{2(5)}{12-10} 0.0076 = 0.0380.$$

Die Summe

$$P(R \geq 10) = 0.0076 + 0.0380 = 0.0456$$

ist noch immer kleiner als 0.05, daher noch immer signifikant. Für r = 9 mit k = 5 nach (5.11) ist

$$P(R = 9) = \frac{4(4)}{2(1)3} 0.0380 = 0.1013$$

und

$$P(R \geq 9) = 0.1013 + 0.0456 = 0.1469,$$

was größer ist als 0.05, und daher ist diese Zahl der Iterationen als zufällig zu betrachten.

Üblicherweise testet man aber eine zweiseitige Hypothese (ohne Richtung), wobei man für $n_1 \leq n_2 \leq 20$ auf Tabellen zurückgreifen kann, in denen die kritischen Werte von r angegeben sind (vgl. Swed, Eisenhart 1943; Siegel 1956; Bradley 1968). In diesen Tabellen sind die kritischen Werte für zweiseitige Hypothesen auf $\alpha = 0.05$, für einseitige hingegen auf $\alpha = 0.025$ angegeben, so daß wir nach diesen Tabellen r = 11 noch als signifikant, aber r = 10 nicht mehr als signifikant betrachten würden, denn $P(R \geq 10) = 0.0456 > 0.025$.

5.2. Große Stichproben

Die Stichproben in der Linguistik sind üblicherweise so groß, daß man in den meisten Fällen eine Approximation mit der Normalverteilung benutzen kann. Um zu testen, ob die beobachtete Zahl von Iterationen zufällig ist, transformiert man r auf die Normalvariable als

$$z = \frac{n(r-1) - 2n_1n_2}{\left\{\frac{2n_1n_2(2n_1n_2 - n)}{n-1}\right\}^{1/2}}, \qquad (5.12)$$

bzw. im absoluten Wert für eine zweiseitige Hypothese. Ist z größer als der kritische Wert $z_{1-\alpha}$, den man in den Tabellen findet, dann kann man die Zahl der Iterationen als signifikant groß betrachten; ist $z < z_\alpha$, dann ist die Zahl der Iterationen signifikant niedrig; ist $|z| > z_{1-\alpha/2}$, dann ist die Zahl der Iterationen bei zweiseitiger Hypothese nicht zufällig.

Illustrieren wir die Rechnung an unseren zwei Beispielen. Für $n_1 = 18$, $n_2 = 14$, n = 32, r = 11 erhalten wir zweiseitig

$$\frac{|32(11-1) - 2(18)14|}{\left\{\frac{2(18)14\ 2(18)14-32}{32-1}\right\}^{1/2}} = 2.10.$$

Da dieser Wert größer als 1.96 (= $z_{0.975}$) ist, betrachten wir die Zahl der Iterationen als nicht zufällig, was mit dem obigen Resultat übereinstimmt.

Dieser Test funktioniert auch in solchen Fällen recht gut, wo n_1, n_2 kleiner sind. In unserem zweiten Beispiel, wo $n_1 = 5$, $n_2 = 7$, $n = 12$ war, erhalten wir für

$$r = 10,\ z = 2.85,$$

was noch, wie oben, signifikant ist, aber für

$$r = 9 \text{ ist } z = 1.95,$$

was schon kleiner als 1.96 ist und daher (zweiseitig) nicht signifikant.

Man kann diesen Test recht zuverlässig für n_1, $n_2 \geq 10$ benutzen.

5.3. Vergleich der Iterationszahl in zwei Texten

Kennt man die mathematische Erwartung und die Varianz einer Zufallsvariablen, dann kann man sie bei großen Stichproben auf die Normalvariable transformieren. Diese Tatsache kann man auch dazu verwenden, die Iterationsstruktur zweier Texte auf Gleichheit zu testen.

Betrachten wir zwei Texte, A und B, in denen wir die Größen

$$r_A,\ n_{1A},\ n_{2A},\ n_A$$

$$r_B,\ n_{1B},\ n_{2B},\ n_B$$

empirisch feststellen können. Die Erwartungen und die Varianzen ergeben sich aus der Verteilung der Iterationen als

$$E(R) = \frac{2n_1n_2 + n}{n} \tag{5.4}$$

$$V(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}, \tag{5.13}$$

so daß die Größe

$$\frac{r_A - r_B - [E(R_A) - E(R_B)]}{[V(R_A) + V(R_B)]^{1/2}} = z \tag{5.14}$$

asymptotisch normalverteilt ist mit N(0,1). Man pflegt oft auch eine Korrektur für Stetigkeit einzubauen, so daß man

$$\frac{r_A - r_B - 1 - [E(R_A) - E(R_B)]}{[V(R_A) + V(R_B)]^{1/2}} = z \tag{5.15}$$

erhält ($r_A > r_B$). Mit diesem Kriterium kann man testen, ob die Zahl der Iterationen im Text A signifikant größer bzw. unterschiedlich ist als im Text B.

Betrachten wir ein von Grotjahn (1980) vorgelegtes Beispiel, in dem dieser die Iterationen von betonten und von unbetonten Silben in Goethes Balladen "Erlkönig" und "Totentanz" verglichen hat. Seien

A = Erlkönig, B = Totentanz

und

$n_{1A} = 128$ (Zahl der betonten Silben im "Erlkönig")

$n_{2A} = 180$ (Zahl der unbetonten Silben im "Erlkönig")

$n_A = 308$

$r_A = 252$ (Zahl der Iterationen im "Erlkönig")

$n_{1B} = 175$

$n_{2B} = 320$

$n_B = 495$

$r_B = 351$

Daten aus "Totentanz"

Um in (5.14) oder (5.15) einsetzen zu können, brauchen wir

$$E(R_A) = \frac{2(128)180 + 308}{308} = 150.61$$

$$V(R_A) = \frac{2(128)180[2(128)180 - 308]}{308^2(307)} = 72.4223$$

$$E(R_B) = \frac{2(175)320 + 495}{495} = 227.26$$

$$V(R_B) = \frac{2(175)320[2(175)320 - 495]}{495^2(494)} = 103.1751.$$

Setzen wir nun diese Zahlen in (5.14) ein, so erhalten wir

$$z = \frac{|252-351-(150.61-227.26)|}{[72.4223 + 103.1751]^{1/2}} = 1.6866,$$

während sich mit der Stetigkeitskorrektur z = 1.61 ergibt. Bei einem zweiseitigen Test (einseitig kann man hier nicht testen) ist dieser z-Wert bei $\alpha = 0.05$ nicht signifikant, da der kritische Wert $z_{0.975} = 1.96$ ist.

Andere Beispiele findet man in Grotjahn (1980).

5.4. Iterationen von mehr als zwei Arten von Elementen

Die Wahrscheinlichkeitsverteilungen der Iterationszahlen werden umso komplizierter und anwendungsfeindlicher, je mehr Arten von Elementen im Text vorhanden sind. Auch die Rechnungen werden langwieriger und umständlicher. Zum Glück ist es möglich, immer eine Transformation der Iterationszahl auf die Normalvariable durchzuführen, und zwar wie üblich als

$$z = \frac{r - E(R)}{[V(R)]^{1/2}} . \qquad (5.16)$$

Sei k die Zahl der Arten der Elemente und n_i (i=1,2,...,k) die Anzahl der Elemente i-ter Art im Text, $n = \Sigma n_i$, dann bekommt man die notwendigen Größen als

$$E(R) = n + 1 - \frac{\sum_{i=1}^{k} n_i^2}{n} \qquad (5.17)$$

$$V(R) = \frac{\sum_{i=1}^{k} n_i^2 \sum_{i=1}^{k} n_i^2 + n(n+1) - 2n \sum_{i=1}^{k} n_i^3 - n^3}{n^2(n-1)} \qquad (5.18)$$

Setzt man sie in (5.16) ein und ordnet, so erhält man

$$z = \frac{\Sigma n_i^2 - n(n-r)}{\{\{\Sigma n_i^2[\Sigma n_i^2 + n(n-1)] - 2n\Sigma n_i^3 - n^3\}/(n-1)\}^{1/2}} \qquad (5.19)$$

Das Verfahren illustrieren wir wieder an einem Beispiel von Grotjahn (1980), der die Iteration von Daktylen, Spondeen und Trocheen in den ersten 30 Zeilen der "Aeneis" gezählt hat und

$n_{Daktyle} = 89$

$n_{Spondeen} = 81 \qquad n = 180$

$n_{Trocheen} = 10$

$r = 127$

bekommen hat.

Daraus ergibt sich

$$E(R) = 180 + 1 - (89^2 + 81^2 + 10^2)/180 = 99.99$$

$$\Sigma n_i^2 = 89^2 + 81^2 + 10^2 = 14582$$

$$\Sigma n_i^3 = 89^3 + 81^3 + 10^3 = 1237410$$

$$V(R) = \frac{14582[14582+180(181)]-2(180)1237410-180^3}{180^2(179)} = 40.7643.$$

Setzen wir nun diese Zahlen in (5.16) ein, so erhalten wir

$$z = \frac{127 - 99.99}{(40.7643)^{1/2}} = 4.23.$$

Die Zahl der Iterationen ist größer als erwartet. Testet man einseitig, so kann man schließen, daß es im Text *signifikant mehr Iterationen* gibt als man durch Zufall erwarten würde, d.h., der Autor trachtete nach rhythmischer Abwechslung. Testet man zweiseitig, dann kann man schließen, daß die Zahl der Iterationen *nicht zufällig* ist.

6. AGGREGATIVE WIEDERHOLUNG

Ununterbrochene Sequenzen, die Iterationen bilden, sind ein sehr spezieller Fall von "Klumpung" identischer Einheiten, und ihr unter- oder überzufälliges Vorkommen ist ein Zeichen starker Tendenzen. In vielen Fällen kann aber eine Einheit, z.B. ein konkretes Wort, eine konkrete Silbe u.ä., nicht in ununterbrochener Folge hintereinander stehen, jedoch kann man ihr häufiges Vorkommen an einigen Stellen des Textes beobachten, und zwar in kleineren Abständen als man aufgrund der Häufigkeit dieser Einheit erwarten würde. Man sagt, daß es "Klumpungen" oder "Aggregationen" gibt, die sich in Form von vielen kleinen und wenigen großen Abständen zwischen einzelnen Vorkommen der Einheit manifestieren. Zur Feststellung derartiger Tendenzen ist die Iterationstheorie nicht geeignet, man muß zu anderen Mitteln greifen.

Die Untersuchung von Distanzen zwischen identischen Einheiten wurde von Zipf (1949) in Gang gesetzt. Sie ist heute ein recht gut entwickeltes linguistisches Gebiet, und ihre Resultate sind befriedigend. Die meisten Autoren arbeiteten mit binären Einheiten und kamen zu der geometrischen Verteilung der Distanzen (Spang-Hanssen 1956; Yngve 1956; Epstein 1953; Uhlířová 1967), die von Brainerd (1976) aus einer Markov-Kette abgeleitet wurde. Herdan (1966) und Králík (1977) erhielten aufgrund anderer Grundannahmen die Exponentialverteilung, Strauss, Sappok, Diller und Altmann (1984) zeigten das Modell der tendenzlosen Verteilung, und durch Annahme einer "Klumpungstendenz" erhielten sie die negative Binomialverteilung. Eine Verallgemeinerung der Verteilung der Distanzen von mehreren Einheiten wurde schließlich von Zörnig (1984a, b, 1986) entwickelt.

6.1. Zufällige Distanzen: Binäre Daten

Das Problem der Verteilung der Distanzen ist zweifach: Erstens ist zu fragen, ob die identischen Einheiten in rein zufällige Distanzen voneinander plaziert werden. Dies widerspricht jedoch der Hypothese von Skinner, nach der das Aussprechen einer Entität ihre Vorkommenswahrscheinlichkeit in geringer Distanz danach erhöht (Skinner 1939, 1941). Daher

ist es notwendig, das Modell dieser tendenzlosen, rein zufälligen Verteilung abzuleiten.

Zweitens geht es uns um folgendes Problem: Falls die Distanzen nicht rein zufällig sind, sondern einem stochastischen Gesetz folgen, wie lautet dann dieses Gesetz? Auf diese Frage gibt es sicherlich mehrere Antworten; denn die Art der Verteilung kann sowohl von irgendwelchen psycholinguistischen, kommunikationstheoretischen oder subjektiven Faktoren abhängen, als auch von der (Eigen)art der betreffenden Einheit, ihrer Ebene in der Hierarchie der Spracheinheiten, von der Textsorte u.ä.. Es ist daher zu erwarten, daß sich hier ein breites Untersuchungsgebiet entwickeln wird.

Betrachten wir eine Einheit A in einem fertigen Text. Die Distanzen zwischen den Vorkommen von A messen wir als Anzahl aller anderen Einheiten (A) des gleichen Typs, d.h.. wenn A ein Wort ist, dann sind A auch Wörter, wenn A ein Buchstabe ist, dann ist A auch ein Buchstabe. Steht kein A zwischen zwei As, dann ist die Distanz gleich 0, steht ein A dazwischen, dann ist die Distanz gleich 1 usw. Man kann sich aber jeweils zwei benachbarte A als eine Urne vorstellen, in die man Kugeln (A) wirft, und das ganze Problem auf ein Urnenmodell überführen.

Gibt es im Text k Vorkommen von A, dann gibt es zwischen ihnen k-1 Zwischenräume, d.h. k-1 "Urnen". Einfachheitshalber bezeichnen wir k-1 = n. In diese Urnen plazieren wir zufällig r Kugeln, die den r Vorkommen von A entsprechen (den Text vor dem ersten und nach dem letzten Erscheinen von A läßt man aus).

Unsere Frage lautet: Wie ist die Wahrscheinlichkeit, daß beim zufälligen Plazieren von r Kugeln in n Urnen genau n_0 Urnen leer sind, n_1 Urnen jeweils 1 Kugel enthalten, n_2 Urnen jeweils 2 Kugeln enthalten usw. Die Summe aller Urnen muß n sein, d.h.

$$n_0 + n_1 + \ldots + n_r = n$$

und die Zahl der Kugeln r, d.h.

$$n_1 + 2n_2 + 3n_3 + \ldots + rn_r = r.$$

Die Zahl der Möglichkeiten, r Kugeln in n Urnen zu plazieren, ist n^r; die Zahl der Möglichkeiten, n Urnen in Gruppen von $n_0, n_1, \ldots, n_r$ aufzuteilen, ist

$$\frac{n!}{n_0!\, n_1! \ldots n_r!}, \tag{6.1}$$

und die Zahl der Möglichkeiten, r Kugeln so zu verteilen, daß in alle n_i Urnen genau i Kugeln geraten, ist

$$\frac{r!}{(0!)^{n_0} (1!)^{n_1} \ldots (i!)^{n_i} \ldots (r!)^{n_r}}. \tag{6.2}$$

Multipliziert man die "günstigen" Anzahlen (6.1) und (6.2) miteinander und dividiert das Ergebnis durch n^r, dann ergibt sich die gesuchte Wahrscheinlichkeit als

$$P(n_0, n_1, \ldots, n_r) = \frac{n!\, r!}{n^r \prod_{i=1}^{r} n! \prod_{i=2}^{r} (i!)^{n_i}}. \tag{6.3}$$

Linguistisch interpretiert, ergibt (6.3) die Wahrschenlichkeit, daß zwischen n_0 Einheiten A eine Distanz 0 besteht, gleichzeitig zwischen n_1 Einheiten A eine Distanz 1 usw.

Die erwartete Anzahl der Urnen n_i ergibt sich aus (6.3) als (vgl. David 1950; Strauss, Sappok, Diller, Altmann 1984)

$$E(n_i) = n\binom{r}{i}\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{r-i}, \tag{6.4}$$

woraus man dann die einzelnen Häufigkeiten schrittweise als

$$\begin{aligned} E(n_0) &= n\left(1 - \frac{1}{n}\right)^r \\ E(n_1) &= r\left(1 - \frac{1}{n}\right)^{r-1} \\ E(n_2) &= \frac{r(r-1)}{2}\left(\frac{1}{n}\right)^1 \left(1 - \frac{1}{n}\right)^{r-2} \end{aligned} \tag{6.5}$$

usw. oder besser mit Hilfe der Rekursionsformel

$$E(n_{i+1}) = \frac{r-i}{i+1} \cdot \frac{1}{n-1} E(n_i) \qquad (6.6)$$

berechnen kann.

Illustrieren wir das Verfahren an einem Beispiel von Strauss, Sappok, Diller, Altmann (1984). In dem in Hexametern geschriebenen Gedicht von Bridges "Poems in Classical Prosody, Epistle II: To a Socialist in London" wurden die rhythmischen Muster von 300 Versen ausgezählt und zwar so, daß ein Daktylus als D, ein Spondeus als S bezeichnet wurde und die letzten zwei Versfüße, die immer gleich sind, ausgelassen wurden. Die ersten Verse ergaben

1. DSSS	*11. SDSS*	*21. DDDS*
2. SDSS	*12. DSDS*	*22. DSDD*
3. SDSS	*13. SDSS*	*23. DDSS*
4. DDSS	*14. SSSS*	*24. SDSS*
5. SDSS	*15. DSSD*	*25. DSDS*
6. DDSS	*16. SDDS*	*26. DSDS*
7. DSDD	*17. SDSS*	*27. SSDS*
8. SDSS	*18. DSSS*	*28. DDSS*
9. DSSS	*19. DSSD*	*29. DSSS*
10. SSSS	*20. DDDS*	*30. DSSS*

Die Abstände für das Muster DSSS sind

7, 8, 10, 0.

Die Auszählung aller Distanzen für DSSS ist in Tabelle 6.1 aufgeführt. Es gab k = 65 DSSS-Muster, daher ist die Zahl der "Urnen" n = k-1 = 64, und die Zahl der "Kugeln" ist r = 300-65 = 235. Die erwarteten Zahlen von einzelnen Abständen ergeben sich aus (6.5) und (6.6) als

$$E(n_0) = 64\left(1 - \frac{1}{64}\right)^{235} = 1.58$$

$$E(n_1) = \frac{235-0}{0+1}\left(\frac{1}{63}\right)1.58 = 5.90$$

$$E(n_2) = \frac{235-1}{1+1}\left(\frac{1}{63}\right)5.90 = 10.95$$

usw. (s. Tab. 6.1).

Tabelle 6.1

Verteilung der Abstände zwischen den Wiederholungen des Musters DSSS in Bridges Gedicht

Abstand i	Beobachtet n_i	Erwartet $E(n_i)$
0	17	1.58
1	13	5.90
2	4	10.95
3	4	13.50
4	6	12.43
5	3	9.12
6	6	5.55
7	2	2.88
8	2	1.30
9	1	0.52
10	2	0.19
11	2	0.06
13	1	0.02
33	1	
Σ	64	

Der Unterschied zwischen dem Modell und der Beobachtung ist bereits optisch so groß, daß man die Hypothese der Zufälligkeit ablehnen kann. Die Berechnung des Chiquadrats, wobei die theoretischen Werte in eckigen Klammern zusammengefaßt werden, ergab $X^2 = 109.48$ mit 5 Freiheitsgraden, einen extrem hohen Wert, der zeigt, daß in den Daten irgendein "Klumpungstrend" vorhanden ist. Die größte Klumpung muß selbstverständlich bei n_0 bestehen, so daß es im Grunde reichen muß, n_0 mit $E(n_0)$ zu vergleichen, um über Klumpung oder Zufälligkeit zu entscheiden.

Zu diesem Zweck empfiehlt es sich, n_0 auf eine Normalvariable zu transformieren, nach der schon bekannten Formel

$$\frac{n_0 - E(n_0)}{[V(n_0)]^{1/2}} = z \,, \qquad (6.7)$$

wobei

$$E(n_0) = n\left(1 - \frac{1}{n}\right)^r$$

$$V(n_0) = n(n-1)(1-\frac{2}{n})^r + n(1-\frac{1}{n})^r - n^2(1-\frac{1}{n})^{2r} \quad (6.8)$$

(vgl. David 1950). Für unsere Daten haben wir

$$\frac{17 - 1.58}{[64(63)(1-\frac{2}{64})^{235}+64(1-\frac{1}{64})^{235}-64^2(1-\frac{1}{64})^{2(235)}]^{1/2}} = 13.03$$

Dieser z-Wert ist so groß, daß er Zufälligkeit ausschließt und die Skinnersche Hypothese unterstützt.

6.2 Klumpungstrendmodelle

Wenn nun die Verteilung der Distanzen nicht rein zufällig geschieht, wie wird sie dann gesteuert? Auf diese Frage gibt es schon mehrere Antworten, die alle recht plausibel sind. Drei Lösungen werden wir nur kurz erwähnen, das Brainerd-Modell werden wir ausführlicher behandeln.

(a) Herdan (1966:127-130) und Králik (1977) gehen davon aus, daß das Vorkommen einer Einheit A im Text von einem Poisson-Prozess gesteuert wird, der zu den Gleichungen

$$P'_x(t) = aP_{x-1}(t) - aP_x(t), \quad x = 1,2,... \quad (6.9)$$

führt. Die Zwischenzeiten zwischen zwei Poisson-Ereignissen sind dann exponentialverteilt, mit der Wahrscheinlichkeitsfunktion

$$f(x) = ae^{-ax} \quad , \quad (6.10)$$

die sich als ein geeignetes Modell erwiesen hat. Der Parameter a, den man zwar aus den Daten schätzen kann, ist frei wählbar und kann als ein "Klumpungsparameter" interpretiert werden. Je größer a, desto stärker die Klumpung. Die Tatsache, daß die Zwischenzeiten der Poisson-Ereignisse stetig sind, ist nicht weiter von Belang, da es allgemein üblich ist, diskrete Daten mit stetigen Modellen zu approximieren.

(b) Epstein (1953), Spang-Hanssen (1956) und Yngve (1956) gehen davon aus, daß die Einheit A mit der Wahrscheinlichkeit p im Text vorkommt; das erste Vorkommen, von dem an man zu zählen anfängt, hat natürlich die Wahrscheinlichkeit 1; dahinter stehen x Einheiten $\bar{A}$, die alle mit der Wahrscheinlichkeit q = 1 - p vorkommen, dann wieder ein A usw. Aufgrund der Unabhängigkeit ergibt sich

$$P_x = 1 \cdot q^x p = pq^x \quad , \qquad x = 0,1,... \quad (6.11)$$

d.h. die geometrische Verteilung. Man sieht leicht, daß die Resultate (6.10) und (6.11) bis auf die Normierungskonstante und Stetigkeit identisch sind, denn a ist die Normierungskonstante bei der Integration der stetigen Funktion e^{-ax} von 0 bis ∞, während p die Normierungskonstante bei der Summierung der diskreten Wahrscheinlichkeitsfunktion q^x von 0 bis ∞ ist, und e^{-a} liegt genauso wie q immer zwischen 0 und 1. Hier könnte man p als den Klumpungsparameter bezeichnen, denn je größer p, desto steiler die Kurve.

Man sieht jedoch, daß beide Funktionen monoton fallend sind, so daß sie nur dort geeignet sein können, wo reale Aggregation stattfindet, jedoch nicht für alle möglichen Verteilungen von Distanzen. Bei einigen Spracheinheiten, z.B. Präpositionen, ist es sogar regelwidrig, zwei oder mehrere identische hintereinander zu stellen. Eine gute Anpassung kann man nur dann erreichen, wenn man die Variablenwerte so in Intervalle zusammenfaßt, daß das erste Intervall die höchste Frequenz hat. So hat Herdan (1966:127) die Distanzen zwischen den Vorkommen der Präposition "k" (zu) in Puschkins "Hauptmannstochter" in Intervalle 1-20, 21-40,... zusammengefaßt. Dadurch entsteht natürlich nur ein vorgetäuschter Eindruck der Aggregation, die es bei "k" gar nicht geben kann.

(c) Strauss, Sappok, Diller und Altmann (1984) betrachteten nicht das Erscheinen der Einheit A, sondern das Plazieren der Einheiten $\bar{A}$ in die Zwischenräume zwischen zwei As als einen Poisson-Prozess. Man kann sich die Zwischenräume wieder als Urnen vorstellen, in die man "Kugeln" einwirft.

Die Urnen können sich zu diesem Verfahren neutral verhalten, d.h., in den Urnen nimmt die Zahl der Kugeln ($\bar{A}$-Einheiten) konstant und regulär zu. In dem Falle erhält man die Formeln (6.9), und die Verteilung der Distanzen führt zur Poisson-Verteilung.

Die Zwischenräume können jedoch auf das Verfahren auch einen Einfluß ausüben:

(i) Wenn ein Zwischenraum die neuen Kugeln um so stärker abstößt, je mehr Kugeln er bereits enthält, dann ersetzen wir a in (6.9) durch eine Funktion $f_x(t) = c-bx$, d.h. $P'_x(t) = f_x(t)P_x(t) + f_{x-1}(t)P_{x-1}(t)$, und die Lösung ergibt die Binomialverteilung;

(ii) Wenn ein Zwischenraum die Kugeln um so mehr anzieht, je mehr Kugeln er bereits enthält, dann ersetzen wir a durch $f_x(t) = c+bx$, was zu der negativen Binomialverteilung führt. Gerade dies ist unser Fall, denn durch ein derartiges Verhalten ergeben sich wenige Zwischenräume mit vielen Kugeln, d.h. wenige große Distanzen, und viele Zwischenräume mit wenigen Kugeln, d.h. viele kleine Distanzen, die eine starke Klumpung bedeuten. Die Formel lautet

$$P_x = \binom{k+x-1}{x} p^k q^x , \qquad x = 0,1,\ldots , \tag{6.12}$$

wobei p und k die Klumpungsparameter sind. Man sieht sofort, daß (6.11) ein Spezialfall von (6.12) ist, wenn k = 1. Während die Ansätze (a) und (b) oft nur mit Klassenzusammensetzung oder nur für spezielle Einheiten geeignet sind, ist Formel (6.12) sehr breit für das Testen von Aggregationstendenzen anwendbar, da sie nicht nur monoton fallend, sondern auch konkav, mit einem Maximum auf $x \neq 0$ verlaufen kann. Unterschiedliche Arten von Spracheinheiten werden unterschiedliche Paare von Parametern k und p haben, wodurch man eine Charakterisierung des "Distanzverhaltens" der Spracheinheiten gewinnen kann.

Die Anpassung von (6.11) bzw. (6.12) erfolgt folgendermaßen. Da wir Klumpung untersuchen, ist es am einfachsten, die ersten Häufigkeitsklassen zur Schätzung der Parameter heranzuziehen. Für die geometrische Verteilung nimmt man

$$p^* = f_0/N; \qquad q^* = 1 - p^*, \tag{6.13}$$

wobei f_0 die Häufigkeit der nullten Klasse und N die Anzahl aller Distanzen ist, $N = \Sigma f_i$. Für die negative Binomialverteilung erhält man am einfachsten

$$q^* = \frac{2f_2}{f_1} - \frac{f_1}{f_0}$$

$$p^* = 1 - q^* \tag{6.14}$$

$$k^* = \frac{f_1}{q^* f_0}$$

oder

$$k^* = \frac{\bar{x}^2}{s^2 - \bar{x}} \tag{6.15}$$

$$p^* = \frac{\bar{x}}{s^2} .$$

Selbstverständlich kann man bessere Schätzungen erhalten (s. Kap. 2), aber da es heute üblich ist, die Anpassung mit Optimierungsmethoden iterativ zu verbessern, kann man (6.13) und (6.14) als Anfangswerte nehmen.

Illustrieren wir die Anpassung anhand der Daten aus Tabelle 6.1 (s. Tabelle 6.2). Die Schätzung von p für (6.11) ergibt

$$p^* = 17/64 = 0.2656$$

$$q^* = 1 - p^* = 0.7344.$$

Die Anpassung mit diesem Parameter findet man in der dritten Spalte von Tabelle 6.2, die optimierte Anpassung in der vierten Spalte.

Für die negative Binomialverteilung kann man in unserem Fall – wegen der etwas "pathologischen" Häufigkeit $f_2 = 4$ – die Schätzung nach (6.14) nicht verwenden, daher benutzen wir (6.15). Wir erhalten

$$\bar{x} = 3.3281$$

$$s^2 = 12.1892,$$

woraus sich

$$k^* = \frac{3.3281^2}{12.1892 - 3.3281} = 1.2500$$

$$p^* = \frac{3.3281}{12.1892} = 0.2730$$

ergibt.

Die berechneten Werte sind in der fünften Spalte von Tabelle 6.2, die optimierte Anpassung in der sechsten Spalte enthalten. Wie man sieht, reicht die geometrische Verteilung völlig aus. Die optimierte negative Binomialverteilung liefert zwar das kleinste Chiquadrat, aber die Zahl der Freiheitsgrade ist kleiner als bei der optimierten geometrischen Verteilung, so daß man ein kleineres P erhält. Dies muß natürlich nicht in jedem Fall so sein.

Durch geschicktes Zusammenfassen von Häufigkeitsklassen liessen sich die Resultate noch etwas verbessern, aber dies ist hier nicht von Belang.

Tabelle 6.2

Anpassung der geometrischen und der negativen Binomialverteilung an die Daten von Tabelle 6.1.

Abstand x	Beobachtete Häufigkeit f_x	Geometrische Verteilung NP_x	Optimierte geometrische Verteilung NP_x	Negative Binomialverteilung NP_x	Optimierte negative Binomialverteilung NP_x
0	17	17.00	14.61	12.63	15.24
1	13	12.48	11.27	11.48	10.75
2	4	9.17	8.70	9.39	8.12
3	4	6.73	6.71	7.39	6.27
4	6	4.94	5.18	5.71	4.90
5	3	3.63	4.00	4.36	3.85
6	6	2.67	3.09	3.30	3.04
7	2	1.96	2.38	2.49	2.40
8	2	1.44	1.84	1.86	1.91
9	1	1.06	1.42	1.39	1.52
10	2	0.78	1.09	1.04	1.21
11	2	0.57	0.84	0.77	0.96
≥12	2	1.57	2.86	2.19	3.83
	64	p =0.2656 X^2=11.04 FG=8 P =0.20	p =0.2283 X^2=9.78 FG=10 P =0.46	k =1.2500 p =0.2730 X^2=11.89 FG=9 P =0.22	k =0.8751 p =0.1941 X^2=9.58 FG=9 P =0.39

6.3. Brainerds Markov-Ketten-Modell

Die Theorie der Markov-Ketten ist ein mächtiges Instrument der Sprach- und Textforschung, da sie es ermöglicht, in linear geordneten Daten Abhängigkeiten aufzuspüren. Für die Textanalyse ist dieser Umstand von besonderer Wichtigkeit, da sich ein Text immer als eine Kette von Elementen auffassen läßt; wenn die Skinnersche Hypothese gilt, dann muß es zwischen den Erscheinungen gleicher Einheiten eine bestimmte Abhängigkeit geben, deren Charakter man gerade mit Hilfe dieser Theorie erfassen kann. Auch wenn das Verfahren keine direkte Antwort auf die Frage nach der Klumpung gibt, eröffnet sie ein breites Forschungsfeld.

Die Theorie der Markov-Ketten läßt sich auch auf Folgen von nichtbinären Daten anwenden, hier bleiben wir jedoch bei den binären, weil es uns um die Abstände geht.

Betrachten wir den Text als eine Kette von Elementen A und $\bar{A}$, die als Zustände dieser Kette bezeichnet werden. Man pflegt den Zustand A als 1, den Zustand $\bar{A}$ als 0 zu bezeichnen. So ergeben die ersten 30 Verse in dem oben angegebenen Gedicht von Bridges, wo DSSS als 1 und $\overline{\text{DSSS}}$ als 0 symbolisiert werden, die Kette

10000000100000000100000000011.

Jeder Text kann auf diese Weise kodiert werden, so daß man für alle Arten von Einheiten ihre sequentiellen Eigenschaften erforschen kann. Wenn es irgendwelche Abhängigkeiten gibt, dann muß es möglich sein, die Wahrscheinlichkeit des Erscheinens einer Einheit in einer Position aufgrund der Kenntnis der Vorgänger zu berechnen. Die bedingte Wahrscheinlichkeit, daß in der Position n die Einheit E erscheint (d.h. das Ereignis E eintritt), wenn in den ersten n-1 Positionen bekannte Einheiten (E_1 bis E_{n-1}) stehen, ist

$$P(E_n \mid E_1 E_2 \ldots E_{n-1}). \qquad (6.3.1)$$

Wir haben es hier nur mit zwei Einheiten (Ereignissen) zu tun, nämlich 1 und 0, daher bedeutet z.B. $E_n = 1$ die Tatsache, daß die n-te Position des Textes den Zustand 1 annimmt. Im allgemeinen kann man die Zustände als x bezeichnen (x = 0,1), so daß man (6.3.1) expliziter als

$$P(E_n = x_n \mid E_1 = x_1, E_2 = x_2, \ldots, E_{n-1} = x_{n-1}) \qquad (6.3.2)$$

schreiben kann.

Wenn die n-te Einheit von den anderen völlig unabhängig ist, dann reduziert sich (6.3.2) auf

$$P(E_n = x_n), \tag{6.3.3}$$

und eine Kette dieser Art heißt ***Markov-Kette nullter Ordnung.***

Wenn das Erscheinen der n-ten Einheit lediglich von der unmittelbar vorausgehenden abhängt, dann wird aus (6.3.2)

$$P(E_n = x_n \,|\, E_{n-1} = x_{n-1}), \tag{6.3.4}$$

und diese Kette ist die eigentliche Markov-Kette oder ***Markov-Kette erster Ordnung.***

Je nachdem, wie viele Vorgänger das Erscheinen einer Einheit bedingen, definiert man Ketten höherer Ordnung, z.B. die Kette zweiter Ordnung

$$P(E_n = x_n \,|\, E_{n-2} = x_{n-2},\ E_{n-1} = x_{n-1}), \tag{6.3.5}$$

die Kette dritter Ordnung

$$P(E_n = x_n \,|\, E_{n-3} = x_{n-3}, E_{n-2} = x_{n-2}, E_{n-1} = x_{n-1}) \tag{6.3.6}$$

usw.

Sind die Einheiten binär, d.h., gibt es nur zwei Zustände, 0 und 1, so stellt die Folge von Nullen eigentlich die Distanz zwischen zwei Einsen dar. Die Größe dieser Distanz ist eine Zufallsvariable, die wir als X bezeichnen. Ihre Wahrscheinlichkeitsverteilung läßt sich aus der Markov-Kette ablesen.

(a) Markov-Kette nullter Ordnung

In einer Kette nullter Ordnung sind die Einheiten voneinander unabhängig, und die Wahrscheinlichkeit der Sequenz von Einheiten wird nach (6.3.1)

$$P(E_1)P(E_2)\ldots P(E_n). \tag{6.3.7}$$

Wenn es sich um binäre Daten mit zwei Zuständen (0,1) handelt, dann erhält man aus (6.3.7) die Wahrscheinlichkeitsverteilung des Abstandes, denn wenn die erste 1 mit der Wahrscheinlichkeit 1 vorkommt, dann erhalten wir

$$1 \cdot P(E_1 = 0)P(E_2 = 0)\ldots P(E_x = 0)P(E_{x+1} = 1)$$

oder einfacher

$$1 \cdot P(0)P(0)\ldots P(0)P(1) = P(0)^x P(1), \tag{6.3.8}$$

d.h. die Wahrscheinlichkeit, daß die Zufallsvariable X (Abstand) den Wert x annimmt, ist

$$P(X = x) = P(1)P(0)^x \qquad x = 0,1,\ldots \tag{6.3.9}$$

Hier ist P(1) einfach die Wahrscheinlichkeit des Vorkommens von 1, die wir oben als p bezeichnet haben, P(0) ist die Wahrscheinlichkeit von 0, die man wegen P(0) = 1 - P(1) = 1 - p als q bezeichnen kann, so daß wir wie oben die geometrische Verteilung

$$P_x = pq^x, \qquad x = 0,1,2,\ldots$$

erhalten.

(b) Markov-Kette erster Ordnung

Hier ist ein Abstand 0 beim Übergang von 1 zu 1, d.h.

$$P(X = 0) = P(X_n = 1 \,|\, X_{n-1} = 1) = P(1|1),$$

die anderen ergeben sich aus dem Übergang von 1 zu 0, aus x-1 Übergängen von 0 zu 0 und einem Übergang von 0 zu 1, d.h.

$$P(X = x) = P(0|1)P(0|0)^{x-1}P(1|0),$$

d.h. wir erhalten insgesamt

$$P(X = x) = P_x = \begin{cases} P(1|1) & \text{für } x=0 \\ P(0|1)P(1|0)P(0|0)^{x-1} & \text{für } x=1,2,\ldots . \end{cases} \tag{6.3.10}$$

Hier haben wir im Grunde nur zwei Parameter, da

$$P(0|0) = 1 - P(1|0)$$

und

$$P(0|1) = 1 - P(1|1). \tag{6.3.11}$$

Schreibt man

$$P(1|1) = \alpha$$

und

$$P(0|0) = q, \quad P(1|0) = 1 - q = p ,$$

dann kann man (6.3.10) als

$$P_x = \begin{cases} \alpha & x = 0 \\ (1-\alpha)pq^{x-1} & x = 1,2,\ldots \end{cases} \qquad (6.3.12)$$

schreiben. In dieser Form erkennt man die sogenannte erweiterte verschobene geometrische Verteilung.

Als Schätzer kann man das Paar

$$P^*(1|1) = \alpha^* = f_0/N$$

$$P^*(1|0) = p^* = \frac{1 - f_0/N}{\bar{x}} \qquad (6.3.13)$$

oder das Paar

$$P^*(1|1) = \alpha^* = f_0/N$$

$$P^*(1|0) = p^* = \frac{f_1/N}{1 - f_0/N} \qquad (6.3.14)$$

nehmen. Zur Illustration passen wir diese Verteilung an die Wiederholung des DSSS-Musters in Bridges Gedicht an, vgl. Tab. 6.3. Laut (6.3.13) erhalten wir

$$P^*(1|1) = f_0/N = 17/64 = 0.2656;$$

$$P^*(0|1) = 1 - 0.2656 = 0.7344$$

$$P^*(1|0) = \frac{1 - 0.2656}{3.6718} = 0.2;$$

$$P^*(0|0) = 1 - 0.2 = 0.8 ,$$

laut (6.3.14) ist

$$P^*(1|0) = \frac{13/64}{1 - 0.2656} = 0.2766;$$

$$P^*(0|0) = 1 - 0.2766 = 0.7234.$$

Tabelle 6.3

Anpassung der Markov-Kette erster Ordnung an die Daten von Bridges

x	f_x	Markov-Kette erster Ordnung	Optimierte Anpassung
0	17	17.00	15.88
1	13	9.40	10.01
2	4	7.52	7.93
3	4	6.02	6.28
4	6	4.81	4.97
5	3	3.85	3.94
6	6	3.08	3.12
7	2	2.46	2.47
8	2	1.97	1.96
9	1	1.58	1.55
10	2	1.26	1.23
11	2	1.01	0.97
≥12	2	4.04	3.70
		$P(1\|1) = 0.2656$	0.2481
		$P(1\|0) = 0.2000$	0.2081
		$X^2 = 7.91$	$X^2 = 7.71$
		FG = 9	FG = 9
		P = 0.54	P = 0.56

Die nach (6.3.13) berechneten Werte ergeben die Verteilung in der dritten Spalte von Tabelle 6.3. Die Anpassung mit den Schätzern (6.3.14) ist schlechter und wird hier nicht aufgeführt. Optimiert man noch die Anpassung, dann erhält man die Resultate in der vierten Spalte von Tabelle 6.3.

Wie man sieht, sind die beiden Anpassungen besser als alle vorherigen. Insbesondere bemerkt man eine Verbesserung gegenüber der Anpassung aufgrund der Markov-Kette nuller Ordnung. Würde man zur Kette zweiter, dritter usw. Ordnung übergehen, so würde man ständig eine etwas bessere Anpassung erhalten. Die Entscheidung, ob man der Kette erster Ordnung den Vorzug gibt, muß mit Hilfe eines Tests durchgeführt werden. Zu diesem Zweck hat Brainerd das Likelihood-ratio-Kriterium aufgestellt, das aus einem Quotienten der Likelihood-Funktionen der beiden Ketten besteht. Für die Markov-Kette nuller Ordnung ist die Likelihood-Funktion

$$L_0 = \prod_{x=0}^{n} [P(1)P(0)^x]^{f_x} \qquad (6.3.15)$$

und für die erster Ordnung

$$L_1 = P(1|1)^{f_0} \prod_{x=1}^{n} [P(1|0)P(0|0)^{x-1}]^{f_x} , \qquad (6.3.16)$$

so daß der Likelihood-Quotient sich als

$$\lambda_{10} = \frac{L_1}{L_0} = \left[\frac{P^*(1|1)}{P^*(1)}\right]^{f_0} \left[\frac{P^*(0|1)P^*(1|0)}{P^*(1)P^*(0|0)}\right]^{N-f_0} \left[\frac{P^*(0|0)}{P^*(0)}\right]^{N\bar{x}} \qquad (6.3.17)$$

darstellt. In diese Formel setzt man die Maximum-Likelihood-Schätzungen der einzelnen Parameter ein, die wir durch die optimierten Parameter ersetzen können, und erhält

$$\lambda_{10} = \left[\frac{0.2481}{0.2283}\right]^{17} \left[\frac{0.7519(0.2081)}{0.2283(0.7919)}\right]^{64-17} \left[\frac{0.7919}{0.7717}\right]^{64(3.6719)}$$

$$= 2.01.$$

Die Größe $2\ln\lambda$ ist ungefähr wie ein Chiquadrat mit 1 Freiheitsgrad verteilt, und da $2\ln(2.01) = 1.40$, können wir schließen, daß die Kette erster Ordnung keine signifikante Verbesserung der Anpassung bringt.

Die Distanzen sind natürlich nicht immer so monoton fallend verteilt, wie in obigem Beispiel, da Klumpungen nicht unbedingt eine 0-Distanz voraussetzen. Ermittelt man beispielsweise die Distanzen zwischen den Vorkommen von "Kind" bzw. allen seinen Referenzen (Sohn, Knabe, du, dir, dich, dein, mein, mir, ihn) in Goethes "Erlkönig", dann erhält man die Verteilung, die in Tabelle 6.4 dargestellt ist. Hier sieht man, daß die Zahl der 1-Abstände größer ist als die der Null-Abstände. Die Anpassung der geometrischen Verteilung (Kette nullter Ordnung) ergibt ein $P = 0.27$, während die Kette erster Ordnung ein $P = 0.49$ liefert. Die negative Binomialverteilung gibt $P = 0.40$.

Tabelle 6.4

Verteilung der Distanzen zwischen "Kind" und seinen Referenzen im "Erlkönig"

x	f_x	0. Ordnung	1. Ordnung
0	5	7.92	4.39
1	9	6.22	5.83
2	3	4.89	4.79
3	4	3.84	3.93
4	2	3.02	3.23
5	0	2.38	2.65
6	4	1.87	2.18
7	2	1.47	1.79
8	1	1.15	1.47
9	0	0.91	1.21
10	0	0.71	0.99
11	2	0.56	0.81
12	0	0.44	0.67
13	2	0.35	0.55
14	2	0.27	0.45
≥15	1	1.00	2.07
	37	$p = P(1) = 0.2139$	$P(1\|1) = 0.1186$
		$X^2 = 11.13$	$P(1\|0) = 0.1787$
		$FG = 9$	$X^2 = 7.40$
		$P = 0.27$	$FG = 8$
			$P = 0.49$

Das Likelihood-Verhältnis ergibt

$$\lambda_{10} = \left[\frac{0.1186}{0.2139}\right]^{5} \left[\frac{0.8814(0.1787)}{0.2139(0.8213)}\right]^{32} \left[\frac{0.8213}{0.7861}\right]^{37(4.6486)} = 2.9798,$$

woraus $2\ln\lambda = 2.18$. Auch wenn die zweite Anpassung eine nicht monoton fallende Verteilung liefert und adäquater erscheint, bringt sie keine signifikante Verbesserung.

(c) Markov-Ketten höherer Ordnung

Die Verteilung der Distanzen aufgrund einer Markov-Kette zweiter Ordnung ergibt sich als (vgl. Brainerd 1976)

$$P_x = \begin{cases} P(1|1) & \text{für } x = 0 \\ P(0|1)P(1|0) & \text{für } x = 1 \\ P(0|1)P(0|10)P(0|00)^{x-2}P(1|00) & \text{für } x = 2,3,\ldots \end{cases} \qquad (6.3.18)$$

und aufgrund der Markov-Kette dritter Ordnung als

$$P_x = \begin{cases} P(1|1) & \text{für } x = 0 \\ P(01|1) & \text{für } x = 1 \\ P(00|1)P(1|100) & \text{für } x = 2 \\ P(00|1)P(0|100)P(0|000)^{x-3}P(1|000) & \text{für } x = 3,4,\ldots . \end{cases} \qquad (6.3.19)$$

Die Maximum-Likelihood-Schätzungen ergeben sich als:

$$\begin{aligned} P^*(1|1) &= f_0/N \\ P^*(0|1) &= 1 - P^*(1|1) \\ P^*(1|10) &= f_1/(N-f_0) \\ P^*(0|10) &= 1 - P^*(1|10) \\ P^*(1|00) &= (N-f_0-f_1)/(N\bar{x}-N+f_0) \\ P^*(0|00) &= 1 - P^*(1|00) \end{aligned} \qquad (6.3.20)$$

für die zweite Ordnung und

$$\begin{aligned} P^*(1|1) &= f_0/N \\ P^*(00|1) &= (N - f_0 - f_1)/N \\ P^*(01|1) &= f_1/N \\ P^*(1|100) &= f_2/(N - f_0 - f_1) \\ P^*(0|100) &= 1 - P^*(1|100) \end{aligned} \qquad (6.3.21)$$

$$P^*(1|100) = \frac{(N - f_0 - f_1 - f_2)}{(N\bar{x} - 2N + 2f_0 + f_1)};$$

$$P^*(0|000) = 1 - P^*(1|000)$$

für die dritte Ordnung.

Die Likelihood-Quotienten für den Test einzelner Ordnungen sind wie folgt:

$$\lambda_{z1} = \left[\frac{P(1|10)}{P(1|0)}\right]^{f_1} \left[\frac{P(1|00)P(0|10)}{P(1|0)P(0|00)}\right]^{N-f_0-f_1} \left[\frac{P(0|00)}{P(0|0)}\right]^{N\bar{x}-N-f_0} \qquad (6.3.22)$$

$$\lambda_{z2} = \left[\frac{P(01|1)}{P(1|10)}\right]^{f_1} \left[\frac{P(1|100)}{P(1|000)}\right]^{f_2} \left[\frac{P(00|1)P(1|000)}{P(0|10)P(1|00)}\right]^{N-f_0-f_1}$$

$$\times \left[\frac{P(0|000)}{P(0|00)}\right]^{N\bar{x}-2N+2f_0+f_1} \left[\frac{P(0|100)}{P(0|000)}\right]^{N-f_0-f_1-f_2} \frac{1}{P(0|1)^{N-f_0}} . \qquad (6.3.23)$$

Brainerd untersuchte mit dieser Methode zahlreiche Daten und zeigte, welche Spracheinheiten in welcher Ordnung ihre Distanzen bilden (Artikel, Pronomina, lange Wörter). Sicherlich sind hinter dieser Erscheinung irgendwelche Sprachgesetze versteckt, deren Erforschung jedoch noch eine Aufgabe für die Zukunft ist. Die Theorie der Markov-Ketten gibt uns dazu ein mächtiges Instrument an die Hand.

Was die Klumpungen betrifft, ergibt sich hier ein ungünstiger Umstand für die Interpretation. Die Kette nullter Ordnung, die keine Abhängigkeiten voraussetzt, liefert ein Modell, in dem Klumpungen nach der Skinnerschen Hypothese eben vorhanden sind. Dort aber, wo die engen Klumpungen sich auflockern, muß man Ketten höherer Ordnung anwenden, in denen Abhängigkeiten vorausgesetzt werden. Dieser Umstand kann folgendermaßen erklärt werden: Klumpungen bedeuten keineswegs eine erhöhte Zahl von 0-Distanzen, die bei vielen Einheiten gar nicht erlaubt sind, sondern eher kleine nicht-0-Distanzen. Dadurch muß die zugrundeliegende geometrische Verteilung immer mehr modifiziert werden. Dies führt zwar zu einer verbesserten Anpassung, aber unglücklicherweise zu einer Vermehrung der Parameter, die sich aber schwerlich alle als

"Klumpungsparameter" interpretieren lassen. Auf jeden Fall kann man berechnen, welcher Abhängigkeitsgrad der gegebenen Distanzverteilung zugrundeliegt, und dies wäre sowohl für die Texttheorie als auch für eine Grammatiktheorie von großer Bedeutung. Eine andere Anwendung der Markov-Ketten findet man in Grotjahn (1979:212-219).

Die Wahl eines der Modelle in 6.1 bis 6.3 für einen konkreten Fall soll keineswegs aufgrund der besten Anpassung, sondern eher aufgrund der besten Interpretierbarkeit getroffen werden.

6.4. Nichtbinäre Daten: Zörnigs Modell

Die oberen Distanzmodelle bezogen sich immer nur auf die Distanzen zwischen den Vorkommen einer einzigen Einheit, alle anderen wurden als die komplementäre Einheit betrachtet. Man kann sich aber auch die Frage stellen, wie die Distanzen verteilt sind, wenn man m unterschiedliche Einheiten hat und die Distanzen zwischen jeweils gleichen zählt.

Die Lösung wurde von Zörnig (1984a,b) aufgrund kombinatorischer Überlegungen erzielt. Wir werden hier nur die zweite Variante in Betracht ziehen, bei der man lediglich die Distanzen zwischen benachbarten Einheiten wählt.

Zur Illustration nehmen wir eine kurze Sequenz von Buchstaben:

A B A C D B C A D D B.

Zwischen den A gibt es Distanzen : 1 und 4
Zwischen den B : 3 und 4
Zwischen den C : 2
Zwischen den D : 3 und 0.

Also gibt es hier eine Nulldistanz, eine Distanz 1, eine Distanz 2, zwei Distanzen 3, und zwei Distanzen 4.

Allgemein geht es also um Sequenzen von n Elementen, in denen es m ***unterschiedliche*** Einheiten gibt, von denen die erste k_1-mal vorkommt, die zweite k_2-mal, ..., die m-te k_m-mal.

Im obigen Beispiel war

$n = 11$

$m = 4$ (A, B, C, D)

$k_1 = k_A = 3$

$k_2 = k_B = 3$

$k_3 = k_C = 2$

$k_4 = k_D = 3.$

Zörnig zeigt, daß die Wahrscheinlichkeit einer Distanz x sich als

$$P_x = \frac{(n-x-1)!}{(n-m)n!} \sum_{i=1}^{m} k_i(k_i - 1)(n - k_i)_{(x)} \qquad (6.4.1)$$

ergibt, wobei

$$n_{(x)} = n(n-1)\ldots(n-x+1) \qquad (6.4.2)$$

ist. Da die Zahl aller Distanzen N = n - m ist, erhalten wir die theoretische Anzahl von Distanzen x als

$$NP_x = \frac{(n-x-1)!}{n!} \sum_{i=1}^{m} k_i(k_i - 1)(n - k_i)_{(x)} \qquad (6.4.3)$$

Dieses Modell gilt unter der Bedingung der Unabhängigkeit der Elemente, d.h. man kann Trends erkennen, indem man empirische Sequenzen mit diesem Modell vergleicht.

Betrachten wir als Beispiel die ersten zwei Verse des "Erlkönigs", transkribiert in Form der klassischen Wortarten:

Pron V Adv Adv Pr N K N

Pron V Art N Pr Pron N

Pron V Art N Adv Pr Art N

Pron V Pron Adv Pron V Pron Adv

Pron N Pron V Pron Adv Adv Pron N

V N Pron Art N Part

Art N Pr N K N

Pron N Pron V Art N.

In dieser Sequenz gibt es

m = 8 unterschiedliche Einheiten
(Adv, Art, K, N, Part, Pr, Pron, V)

mit den Häufigkeiten

$k_1 = k_{Adv} = 7$

$k_2 = k_{Art} = 6$

$k_3 = k_K = 2$

$k_4 = k_N = 15$

$k_5 = k_{Part} = 1$

$k_6 = k_{Pr} = 4$

$k_7 = k_{Pron} = 15$

$k_8 = k_V = 8$

Die Sequenz hat die Länge

$$n = \sum_{i=1}^{m} k_i = 58.$$

Die Verteilung berechnen wir also aus

$$NP_x = \frac{(58-x-1)!}{58!}\,[7(6)51_{(x)} + 6(5)52_{(x)} + 2(1)56_{(x)} + (2)15(14)43_{(x)} + 4(3)54_{(x)} + 8(7)50_{(x)}].$$

Da $n_{(0)} = 1$, erhalten wir

$$NP_0 = \frac{57!}{58!}\,[7(6)+6(5)+2(1)+2(15)14+4(3)+8(7)] = 9.69$$

$$NP_1 = \frac{56!}{58!}\,[7(6)51 + 6(5)52 + 2(1)56 + 2(15)(14)43 + 4(3)54 + 8(7)50] = 7.66$$

$$NP_2 = \frac{55!}{58!}\,[7(6)51(50) + 6(5)52(51) + 2(1)56(55) + 2(15)14(43)42 + 4(3)54(53) + 8(7)50(49)] = 6.07$$

usw. Alle berechneten Häufigkeiten sind in der dritten Spalte von Tabelle 6.5 aufgeführt. Die empirischen Distanzen ergeben sich durch direkte Zählung. So erhalten wir

für Pronomina die Distanzen

7, 4, 1, 7, 1, 1, 1, 1, 1, 1, 2, 3, 9, 1,

für Verben

7, 6, 7, 3, 5, 5, 14,

für Adverbien

0, 15, 6, 3, 5, 0,

für Präpositionen

7, 7, 27,

für Nomina

1, 3, 2, 3, 3, 9, 6, 1, 2, 2, 1, 1, 1, 3,

für Artikel

6, 3, 21, 2, 9,

für Konjunktionen

43

und keine Abstände für Partikel, da in der Sequenz nur eine vorhanden ist. Die Zahl der Distanzen für die Einheit i ist $k_i - 1$. Faßt man diese empirischen Distanzen zu Klassen zusammen, dann bekommt man die Verteilung wie in der zweiten Spalte von Tabelle 6.5.

Aufgrund der grammatischen Restriktionen werden 0-Distanzen zwischen gleichen Wortarten vermieden, und es ist zu erwarten, daß der Dichter die Distanzen zwischen gleichen Wortarten nach irgendeinem Muster, nach irgendeinem Rhythmus gestaltet hat. Es wäre interessant, zu

untersuchen, wie sich dieser Rhythmus in unterschiedlichen Textsorten und bei unterschiedlichen Autoren bildet.

Tabelle 6.5

Distanzen zwischen Wortarten in den ersten zwei Strophen im "Erlkönig"

Distanz x	Beobachtete Anzahl der Distanzen x f_x	Berechnete Anzahl der Distanzen x NP_x
0	2	9.69
1	13	7.66
2	5	6.07
3	8	4.82
4	0	3.84
5	3	3.07
6	4	2.47
7	6	2.00
8	0	1.62
9	3	1.33
10	0	1.10
11	0	0.91
12	0	0.76
13	0	0.64
14	1	0.55
15	1	0.46
21	1	
27	1	3.01
43	1	
N = n-m = 50		X^2 = 25.75 FG = 2 P = 0.0000026

Ein Chiquadrat-Test zeigt, daß die empirische Verteilung stark von dem Zufallsmodell abweicht.

Das Modell von Zörnig zeigt eine wichtige Tatsache, die auch für binäre Daten gilt: Ein monoton fallender Verlauf der Häufigkeiten von Distanzen kann sich auch ohne jegliche Tendenz ergeben. Dadurch werden der Skinnerschen Hypothese Einschränkungen auferlegt. Möglicherweise gilt sie nur im phonischen und im semantischen Bereich, aber nicht im grammatischen und im formalen. Es bedarf umfangreicher Untersuchungen, bis in diesem Bereich fundierte Resultate vorliegen werden.

6.5. Ähnlichkeitsaggregative Wiederholung

Wenn wir annehmen, daß im Text eine Selbststimulation für die Erscheinung gleicher Einheiten im Skinnerschen Sinne vorliegt, dann müssen wir auch zulassen, daß eine Einheit das Erscheinen einer ***ähnlichen*** Einheit stimuliert, gleichgültig, ob diese Ähnlichkeit nun formal oder inhaltlich ist.

Die Bestätigung dieser Hypothese wäre eine stärkere Unterstützung der Skinnerschen Hypothese als die Übereinstimmung der Klumpungsdaten mit den obigen Modellen, deren Teil, wie wir sahen, unter der Annahme der "Nichtaggregation" entstand.

Die Prüfung einer Ähnlichkeitsaggregation ist jedoch mit großen Problemen verbunden, von denen zwei besonders wichtig sind.

(1) ***Ähnlichkeit*** ist ein sehr heikler Begriff, der auf zahlreiche Weisen gemessen werden kann (vgl. z.B. Bock 1974), wobei man auf die Art der Daten, die Merkmale, das gegebene Problem u.a. Rücksicht nehmen muß.

(2) ***Spontaneität*** der Texterzeugung ist ein Faktor, der Ähnlichkeitsaggregation begünstigt. Schreibt der Autor spontan, dann haben die benutzten Einheiten noch die Fähigkeit, etwas zu stimulieren. Macht er aber große oder zahlreiche Pausen, dann erlischt der Stimulus und ruft keine ähnlichen Einheiten hervor. Auf der anderen Seite kann auch ein spontan erzeugter Text mit hoher Ähnlichkeitsaggregation nachträglich so korrigiert werden, daß jegliche Spur dieser Tendenz verschwindet. Daher ist anzunehmen, daß Ähnlichkeitsaggregation nur in wenigen Texten zu finden sein wird, vor allem in Folkloretexten. Den Schluß kann man aber schwerlich in die andere Richtung richten: Entdeckt man nämlich in einem Text beispielsweise keine lautliche Ähnlichkeitsaggregation, dann kann man daraus nicht schließen, daß der Text nicht spontan erzeugt wurde; es können nämlich zahlreiche andere, recht komplizierte Ähnlichkeiten vorhanden sein, die man nicht entdeckt hat. Man kann aber schließen, daß in der gegebenen Hinsicht keine Spontaneität vorhanden war.

An dieser Stelle werden wir nur die phonische Gestaltung eines malayischen Epos "Shair Cinta Berahi" (Djadjuli 1961) nach Altmann (1968) untersuchen. Epen dieser Art entstanden durch spontane Kreation eines Erzählers, der über ein bekanntes Thema improvisierte. Sollte das Gedicht in dem improvisierten Zustand aufgenommen worden sein, dann kann man davon ausgehen, daß einander näher liegende Verse phonisch ähnlicher

sind als weiter voneinander entfernte. Mit anderen Worten, die phonische Ähnlichkeit der Verse ist eine Funktion ihrer Distanz.

Um diese Funktion zu finden, ziehen wir in Betracht, daß das Lautinventar der Sprache beschränkt ist, so daß auch sehr weit entfernte Verse im Durchschnitt eine gewisse phonische Ähnlichkeit aufweisen, die nicht kleiner als 0 sein kann. Daher kann diese Funktion keine fallende Gerade sein. Das heißt, daß die relative Veränderung der Ähnlichkeit mit der Distanz nicht konstant, sondern invers proportional zu der Distanz ist und ständig kleiner wird. Formal ausgedrückt, ergibt dies (S = Ähnlichkeit, D = Distanz)

$$\frac{S'}{S} = -\frac{b}{D} \quad , \qquad (6.5.1)$$

woraus die Funktion

$$S = aD^{-b}$$

folgt. Um zu prüfen, ob diese Kurve die phonetische Ähnlichkeit erfaßt, müssen wir diese definieren und an den Daten messen.

Die phonetische Ähnlichkeit zweier Verse ist eine äußerst komplexe Angelegenheit, die wir etwas vereinfachen müssen. Wir betrachten die ersten zwei Verse des genannten Gedichts in phonetisch/phonologischer Transkription

1. děŋarkan tuan suatu cĕrita

2. dikaraŋ oleh dagaŋ yaŋ lata

und bilden aus den Lauten/Phonemen die Mengen A_1 und A_2.

$$A_1 = \{a_1, a_2, a_3, a_4, a_5, c, d, ė_1, ĕ_2, i, k, n_1, n_2, ŋ, r_1, r_2, s, t_1, t_2, t_3, u_1, u_2, u_3\},$$

wobei die einzelnen gleichen Phoneme durch einen Index unterschieden werden. Der zweite Vers ergibt

$$A_2 = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, d_1, d_2, e, g, h, i, k, l_1, l_2, ŋ_1, ŋ_2, ŋ_3, o, r, t, y\}.$$

Weiter bilden wir die Mengen von konsekutiven Phonempaaren B_1 und B_2

$$B_1 = \{an_1, an_2, ar, at, cĕ, dé, ėŋ, er, it, ka, ns,$$

$$nt, na, ri, rk, su, ta, tu_1, tu_2, ua_1, ua_2, uc\}$$

$$B_2 = \{ag, aŋ_1, aŋ_2, aŋ_3, ar, at, da, di, eh, ga, hd, ik, ka, la, le, ŋl, ŋo, ŋy, ol, ra, ta, ya\},$$

die als Approximation ausreichen. Wir bilden aus A_1 und A_2 die Schnittmenge der gleichen Phoneme

$$A_1 \cap A_2 = \{a_1, a_2, a_3, a_4, a_5, d, i, k, ŋ, r, t\}$$

und aus B_1 und B_2 die Schnittmenge gleicher Paare

$$B_1 \cap B_2 = \{ar, at, ka, ta\}.$$

Ein Ähnlichkeitsmaß zwischen den Versen i und j bilden wir wie folgt

$$S_{ij} = 100 \frac{|A_i \cap A_j|^2}{|A_i| \cdot |A_j|} + \frac{|B_i \cap B_j|^2}{|B_i| \cdot |B_j|} \quad , \qquad (6.5.3)$$

wo $|x|$ die Kardinalzahl der Menge x ist. Aus unseren Daten bekommen wir

$$|A_1| = 23 \qquad |B_1| = 22$$

$$|A_2| = 23 \qquad |B_2| = 22$$

$$|A_1 \cap A_2| = 11 \qquad |B_1 \cap B_2| = 4.$$

Setzt man diese Zahlen in (6.5.3) ein, so erhält man

$$S_{1,2} = 100\left(\frac{11^2}{23(23)} + \frac{4^2}{22(22)}\right) = 26.18.$$

Auf diese Weise berechnet man für jede Distanz d = j-i n Fälle und bildet die durchschnittliche Ähnlichkeit für die Distanz d als

$$\bar{S} = \frac{1}{n} \sum_{\substack{i,j \\ d=j-i}} S_{ij} \quad . \qquad (6.5.4)$$

Für unser Beispiel wurden für jede Distanz 150 Paare aus dem Epos zufällig ausgewählt und die durchschnittlichen phonischen Ähnlichkeiten der Verse für jede Distanz von 1 bis 100 separat berechnet. Da aber der Reim benachbarter Verse, der eine intentionale Ähnlichkeit darstellt (im malayischen shair hat der Reim die Form aaaa/bbbb/...), das Ähnlichkeitsmaß stark beeinflußt hätte, wurden die letzten vier Phoneme jedes Verses außer acht gelassen. Ihre Einbeziehung in die Rechnung hätte den Anfang der Kurve stark in die Höhe getrieben und die Resultate verfälscht. Die Resultate sind in Tabelle 6.6 dargestellt. Die Anpassung der Formel (6.5.2) erfolgte wie folgt. Die Anfangswerte für die Optimierung wurden ohne große Mühe aus den ersten zwei Werten berechnet, nämlich als

$$b = \frac{\log \bar{s}_1 - \log \bar{s}_2}{\log D_1 - \log D_2} \qquad (6.5.5)$$

$$\log a = \log S_1 + b \log D_1 .$$

Tabelle 6.6

Durchschnittliche Ähnlichkeiten von Versen in Distanz D_i

D_d	1	2	3	4	5	6	7	8	9	10
$\bar{s}_d$	36.06	34.60	34.44	34.77	33.78	33.84	34.01	33.76	33.21	32.91
$\bar{s}_t$	35.87	35.06	34.59	34.24	34.02	33.81	33.64	33.50	33.37	33.25

In unserem Fall ergab sich

$$b = \frac{\log 36.06 - \log 34.60}{\log 1 - \log 2} = \frac{0.0413}{-0.6931} = -0.0596$$

$$\log a = 3.5852,\ a = 36.06 ,$$

woraus nach Optimierung

$$a = 35.8672 \qquad b = 0.0329$$

folgte. Die mit diesen Parametern berechneten Werte findet man in der dritten Zeile Tabelle 6.6. Der F-test nach Linearisierung zeigte eine sehr gute Anpassung, so daß wir eine Ähnlichkeitsaggregation mit ziemlich großer Sicherheit annehmen können.

6.6. Ausblick

Mit der aggregativen Wiederholung ergeben sich zahlreiche Probleme, die erst durch Untersuchung vieler Texte entwickelt und gelöst werden können. Folgende Probleme sollen hier erwähnt werden:

(1) Welche Einheiten weisen überhaupt eine aggregative Wiederholung auf? Sind das nur phonische, oder auch metrische, formale, grammatische, semantische, metaphorische u.a. Einheiten?

(2) Gibt es eine Entwicklung der aggregativen Wiederholung in der Geschichte der Texte?

(3) Weisen bestimmte Textsorten mehr aggregative Wiederholungen als andere auf?

(4) Kann man von der aggregativen Wiederholung auf die Spontaneität der Texterzeugung schließen?

(5) Welches Ähnlichkeitsmaß soll man bei den einzelnen Einheiten verwenden?

7. BLOCKMÄSSIGE WIEDERHOLUNG

Teilt man einen Text in Blöcke (Passagen) von jeweils 50 Wörtern ein und untersucht das Vorkommen eines bestimmten Wortes A in einer derartigen Passage, dann wird man feststellen, daß man eine sehr reguläre empirische Verteilung bekommt, in der die Anzahl der Blöcke f_x, die genau x Vorkommen von A besitzen, mit Hilfe einer Wahrscheinlichkeitsverteilung modelliert werden könnte. Als Beispiel betrachten wir die Verteilung der Artikel des Englischen in einer Stichprobe aus *Cheevers Wapshot Chronicle*, die von Brainerd (1972) stammt und in Tabelle 7.1 präsentiert wird. Die Zahlen sind folgendermaßen zu lesen: Es gibt 8 Passagen von 50 Wörtern, in denen kein Artikel vorkommt; es gibt 14 Passagen in denen 4 Artikel vorkommen usw.

Tabelle 7.1

Verteilung der Artikel in einer Stichprobe aus *Cheevers Wapshot Chronicle*
nach BRAINERD (1972)

Zahl der Artikel in der Passage	Zahl solcher Passagen
0	8
1	8
2	12
3	11
4	14
5	4
6	8
7	8
8	6
9	2
10	2

Forscher haben seit langem vermutet, daß sich hinter diesen Verteilungen irgendein Gesetz verbirgt, das die Vorkommen von Wörtern in Passagen – wohl aus inhaltlichen, grammatischen, kommunikativen Gründen – auf eine bestimmte Weise steuert. Als erste hat Frumkina (1962) die blockmäßigen Wiederholungen von einigen russischen Wörtern unter-

sucht. Sie ging von der Annahme aus, daß einzelne Wörter niedrige Häufigkeiten hätten und daher die Poisson-Verteilung (= das Gesetz der kleinen Zahlen) das geeignete Modell sei. Man konnte aber zeigen (vgl. Altmann, Burdinski 1982), daß in 5 von 12 Fällen bei Frumkina die Poisson-Verteilung nicht geeignet ist. Brainerd (1972) gelangen jedoch zahlreiche sehr gute Anpassungen der Poisson-Verteilung an die Wiederholungen englischer Artikel in Blöcken. Die einzige mißlungene Anpassung hat er mit einer gemischten Poisson-Verteilung erfaßt. Piotrowski (1984:111–119) gab eine Übersicht über die Bemühungen sowjetischer Autoren (Maškina 1968; Bektaev, Lukjanenkov 1971; Paškovskij, Srebrjanskaja 1971), die die Poisson-, die Normal- und die Lognormalverteilung verwendet und ihre Schlüsse danach gezogen haben, welche von den Verteilungen für das gegebene Wort geeignet war. Mosteller und Wallace (1964) benutzten die Poisson- und die negative Binomialverteilung mit guten Resultaten (vgl. auch Francis 1966).

Ein Modell für die blockmäßige Wiederholung wurde von Altmann und Burdinski (1982) vorgelegt und soll hier dargestellt werden. Dieses Modell wurde als "Frumkina-Gesetz" bezeichnet.

7.1. Frumkina-Gesetz

Man betrachtet eine Spracheinheit A wie etwa ein Morphem, ein Wort, eine Phrase usw., die mit der Wahrscheinlichkeit p vorkommt. Ob dieses p für die Sprache als ganze oder nur für den gegebenen Text gilt, sei zunächst dahingestellt. Wenn man den Kontext nicht einbezieht, dann ist p in der gesamten Passage konstant. Wenn man aber die Kontextbedingtheit von Spracheinheiten in Betracht zieht, dann sieht man sofort, daß an vielen Stellen $p = 0$ sein muß, weil die gegebene Einheit A dort nicht vorkommen kann (z.B. zweimal hintereinander), während es Positionen gibt, wo A mit unterschiedlichen Wahrscheinlichkeiten vorkommen kann, d.h., p im Text nicht konstant ist.

Gehen wir aber zunächst davon aus, daß p konstant ist. Weiter nehmen wir an, daß in der gegebenen Passage ***höchstens*** n Stück A vorkommen können, wobei dieses n nicht a priori bekannt sein muß. Dann ist die Wahrscheinlichkeit, daß man in einer Passage mit höchstens n Einheiten A dieses A genau x-mal findet, durch die Binomialverteilung

$$P(X=x|p) = f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad ,x=0,1,\ldots,n \qquad (7.1)$$

gegeben. Die Wahrscheinlichkeit hängt natürlich von p ab, das wir jetzt als eine Variable betrachten, die ihre eigene stetige Wahrscheinlichkeitsverteilung f(p) hat, wobei $0 < p < 1$. Die gemeinsame Verteilung von x und p ergibt sich als

$$f(x,p) = f(x|p)f(p) \quad , \tag{7.2}$$

und daraus berechnet man die Verteilung von x als die Marginalverteilung von (7.2), nämlich

$$f(x) = \int_0^1 f(x|p)f(p)dp. \tag{7.3}$$

Die Frage nach der Verteilung von p ist sehr problematisch. Wie Orlov (Orlov, Boroda, Nadarejšvili 1982) gezeigt hat, gibt es keine Gesamtheit aller Texte einer Sprache, die nach festen Gesetzmäßigkeiten mit konstanten Parametern aufgebaut wäre. Es gibt lediglich Einzeltexte, für die wohl Gesetze gelten, deren Parameter sich aber von Text zu Text unterscheiden. Daher läßt sich die Verteilung von p theoretisch im allgemeinen nicht ableiten. Wir müssen uns vorläufig mit einem Ansatz abfinden. Wir nehmen an, daß p einer Beta-Verteilung der Form

$$f(p) = \frac{1}{B(M,K-M)} p^{M-1}(1-p)^{K-M-1}, \quad 0<p<1 \tag{7.4}$$

folgt. Hier bedeutet

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad ,$$

und Γ stellt die Gammafunktion dar, die beispielsweise für ganzzahlige Argumente $\Gamma(k) = (k-1)!$ ergibt.

Setzt man (7.1) und (7.4) in (7.2) ein, so erhält man

$$f(x) = \int_0^1 \binom{n}{x} p^x q^{n-x} \frac{1}{B(M,K-M)} p^{M-1} (1-p)^{K-M-1} dp$$

$$= \binom{n}{x} \frac{1}{B(M,K-M)} \int_0^1 p^{M+x-1} (1-p)^{K-M+n-x-1} dp$$

$$= \binom{n}{x} \frac{B(M-x,K-M+n-x)}{B(M,K-M)} \quad . \tag{7.5}$$

Schreibt man (7.6) in Form von Binomialkoeffizienten, dann bekommt man

$$f(x) = P(X=x) = \frac{\binom{-M}{x}\binom{-K+M}{n-x}}{\binom{-K}{n}}, \quad x=0,1,\ldots,n;\ n\varepsilon N,\ K>M>0 \tag{7.6}$$

Diese Verteilung wird als negativ hypergeometrisch (NH) oder als Betabinomial bezeichnet. Sie ist ein Spezialfall der verallgemeinerten hypergeometrischen Verteilung (Typ IIA von Kemp, Kemp 1956a; vgl. auch Johnson, Kotz 1969; Ord 1972). Sie ist sehr wendig und dürfte für unsere Zwecke hinreichend geeignet sein. Sie hat drei für linguistische Anwendungen wichtige Grenzfälle, nämlich

(i) die ***Binomialverteilung***, gegen die sie konvergiert, wenn $K \to \infty$, $M \to \infty$ und $M/K \to p$,

(ii) die ***Poisson-Verteilung***, gegen die sie konvergiert, wenn $K \to \infty$, $M \to \infty$, $n \to \infty$ und $Mn/K \to a$,

(iii) die ***negative Binomialverteilung***, gegen die sie konvergiert, wenn $K \to \infty$, $n \to \infty$ und $K/(K+n) \to p$.

Wir sind der Überzeugung, daß das Grenzverhalten und Spezialfälle von Wahrscheinlichkeitsverteilungen in der Linguistik von großer Bedeutung sind. Da man hier keine Grundgesamtheiten hat, ist es vernünftig, ein Grundmodell wie etwa die NH-Verteilung nur dann zu benutzen, wenn die Grenzfälle mit weniger Parametern ungeeignet sind. Es kann sich auch herausstellen, daß eine Einheit in einem Text nach dem Grundmodell verteilt ist, während sie in einem anderen ein bestimmtes Grenzverhalten, in einem dritten ein anderes Grenzverhalten aufweist. Dies kann von Stilunterschieden herrühren und braucht nicht als Falsifikation des Modells betrachtet zu werden. In allen Fällen soll man immer auf die einfachste Verteilung (d.h. die mit den wenigsten Parametern) zurückgreifen. Die obigen Angaben über das Streben eines Parameters gegen unendlich soll man etwas lockerer als "sehr groß" auffassen. Obwohl n die Größe der Passage nicht übersteigen darf, ist diese meistens so groß, daß sie praktisch als unendlich betrachtet werden kann.

7.2. Überprüfung des Frumkina-Gesetzes

Bei der Anpassung von Wahrscheinlichkeitsverteilungen verfahren wir folgendermaßen: Wir prüfen zunächst immer das Grundmodell (NH). Zu diesem Zweck berechnen wir irgendwelche Anfangswerte der Parameter mit einfachen Methoden, und dann verbessern wir die Anpassung iterativ mit Hilfe einer Optimierungsmethode so lange, bis wir die minimale Summe der Abweichungsquadrate oder ein minimales Chiquadrat erhalten. Als Methode benutzen wir den Algorithmus von Neider und Mead (1964) und den von Hook und Jeeves (1961).

Die Angangswerte der Verteilungen ermitteln wir wie folgt:

Für die ***Poisson-Verteilung***

$$a^* = \bar{x};$$

wo

$$\bar{x} = (1/N)\Sigma x f_x \qquad \text{(Stichprobenmittelwert)}, \qquad (7.7)$$

für die ***Binomialverteilung***

$$n^* = x_{max}$$

$$p^* = 1 - (f_0/N)^{1/n}, \qquad (7.8)$$

für die ***negative Binomialverteilung***

$$M^* = \frac{\bar{x}^2}{s^2 - \bar{x}}$$

$$p^* = \frac{\bar{x}}{s^2}$$

wo

$$s^2 = \frac{1}{N}\Sigma(x-\bar{x})^2 f_x \qquad \text{(Stichprobenvarianz)}, \qquad (7.9)$$

und für die ***negative hypergeometrische Verteilung*** (vgl. Kemp, Kemp 1956b)

$$n^* = x_{max}$$

$$K^* = \frac{n^*(n^*\bar{x} - \bar{x}^2 - s^2)}{n^*(s^2 - \bar{x}) + \bar{x}^2} \qquad (7.10)$$

$$M^* = \frac{K^*\bar{x}}{n^*}.$$

Die Rekursionsformeln für die Berechnung der einzelnen Wahrscheinlichkeiten lauten wie folgt:

Für die ***Poisson-Verteilung*** mit der Wahrscheinlichkeitsfunktion

$$P_x = \frac{e^{-a}a^x}{x!}, \qquad x=0,1,\ldots \qquad (7.11)$$

ist

$$P_0 = e^{-a}$$

$$P_x = \frac{a}{x}\cdot P_{x-1}. \qquad (7.12)$$

Für die ***Binomialverteilung*** ist

$$P_0 = q^n$$

$$P_x = \frac{n-x+1}{x}\cdot\frac{p}{q}\cdot P_{x-1}. \qquad (7.13)$$

Für die ***negative Binomialverteilung*** mit der Wahrscheinlichkeitsfunktion

$$P_x = \binom{M+x-1}{x} p^M q^x, \quad x = 0,1\ldots \qquad (7.14)$$

ist

$$P_0 = p^M$$

$$P_x = \frac{M+x-1}{x} \cdot qP_{x-1} \qquad (7.15)$$

Für die ***negative hypergeometrische Verteilung*** ist

$$P_0 = \frac{(K-M)(K-M+1)\ldots(K-M+n-1)}{K(K+1)\ldots(K+n-1)}$$

$$P_x = \frac{(M+x-1)(n-x+1)}{x(K-M+n-x)} P_{x-1} \qquad (7.16)$$

Die drei folgenden Grenzfälle kann man im voraus mit den folgenden Kriterien unterscheiden:

$$\text{wenn } \bar{x} > s^2 \text{ , dann Binomialverteilung,}$$

$$\text{wenn } \bar{x} = s^2 \text{ , dann Poisson-Verteilung,} \qquad (7.17)$$

$$\text{wenn } \bar{x} < s^2 \text{ , dann negative Binomialverteilung.}$$

Bei der Optimierung (oder der schrittweise durchgeführten Verbesserung) der Anpassung kann man sich auch nach den Kriterien des Grenzverhaltens der NH-Verteilung richten:

Wächst schnell

K und M --> Binomialverteilung

K und n --> negative Binomialverteilung

K und M und n --> Poisson-Verteilung (7.18)

Übersichtlich kann man die Kriterien wie folgt zusammenstellen:

K -> ∞	(M -> ∞) ∩ ($\bar{x}$ > s^2)	Binomialverteilung
	(n -> ∞) ∩ ($\bar{x}$ < s^2)	neg.Binomialverteilung
	(M -> ∞)∩(n -> ∞)∩($\bar{x}$ = s^2)	Poisson-Verteilung

Betrachten wir nun einige Beispiele aus verschiedenen Sprachen.

Frumkina (1962) hat in 110 Passagen von jeweils 1000 Wörtern aus Puschkins Texten die Verteilung des Wortes "bez" (ohne), wie in Tabelle 7.2 dargestellt, ermittelt.

Tabelle 7.2

Verteilung von "bez"
nach Frumkina (1962)

Vorkommen in der Passage	Zahl der Passagen mit x Vorkommen von "bez"	Poisson-Verteilung	NH-Verteilung
x	f_x	NP_x	NP_x
0	36	37.20	36.38
1	42	40.33	40.95
2	23	21.86	22.50
3	6	7.90	7.87
4	3	2.70	2.31
Σ	110	a = 1.0841	K = 96.1572
$\bar{x}$	1.0727	X^2 = 0.6565	M = 11.5778
s^2	0.9947	FG = 3	n = 9
		P = 0.8834	X^2 = 0.6955
			FG = 1
			P = 0.4043

Da hier $\bar{x} \approx s^2$, kann man versuchen, die Poisson-Verteilung anzupassen. Die berechneten Werte, NP_x, sind in der dritten Spalte von Tabelle 7.2 enthalten. Wie man sieht, ist die Anpassung sehr gut (P = 0.88).

Die NH-Verteilung ergibt in allen x-Punkten bis auf x = 4 eine bessere Anpassung, aber gerade diese Klasse liefert wegen des kleinen NP_x den größten Beitrag zum Chiquadrat, so daß das Resultat etwas ungünstig aussieht. Es läßt sich zeigen, daß sich mit dem Anwachsen des Parameters n von seinem Minimum (n = 4 für diesen Fall) die Anpassung der NH-Verteilung zunächst verbessert, dann wieder verschlechtert.

In anderen Fällen stellt man fest, daß die NH-Verteilung mit wachsenden Parametern zwar langsam aber ständig die Anpassung verbessert, ein Ende aber schwer auszumachen ist. In Tabelle 7.3 findet man die Anpassung der NH-Verteilung an das Vorkommen des Artikels "das" im Nominativ in Passagen von S. Lenz, "Deutschstunde".

Tabelle 7.4

Verteilung von "das" im Nominativ in Passagen von Lenz

Zahl von 'das' in der Passage x	Zahl der Passagen mit x 'das' f_x	negative hypergeometrische Verteilung NP_x					negative Binomial verteilung NP_x
0	95	96.22	96.14	95.93	95.96	95.87	95.81
1	58	51.04	52.23	53.88	54.48	54.62	54.86
2	28	27.33	27.40	27.15	26.96	26.93	26.91
3	11	14.19	13.64	12.91	12.60	12.55	12.46
4	3	6.77	6.35	5.86	5.68	5.66	5.60
5	2	2.88	2.72	2.55	2.49	2.50	2.47
6	1	1.04	1.05	1.06	1.07	1.08	1.08
7	1	0.29	0.35	0.42	0.45	0.46	0.46
≥8	1	0.05	0.12	0.24	0.31	0.33	0.35
$\bar{x}$ = 0.95	n	8	10	20	50	100	M = 1.40
s^2 = 1.68	K	7.02	9.91	24.36	67.63	139.2	p = 0.59
	M	0.87	0.97	1.18	1.31	1.35	-
	X^2	5.99	4.59	3.09	2.59	2.47	2.35
	FG	3	3	3	3	3	4
	P	0.11	0.20	0.38	0.46	0.48	0.67

Der Parameter n wird allmählich vergrößert, wodurch auch K schnell wächst, nicht aber M, was ein Zeichen dafür ist, daß die NH-Verteilung möglicherweise gegen die negative Binomialverteilung strebt, was auch durch das Kriterium (7.17) angedeutet wird. In der letzen Spalte von Tabelle 7.4 sieht man die beste Anpassung durch die negative Binomialverteilung.

Es gibt Fälle, wo (7.17) für die negative Binomialverteilung spricht, aber die Vergrößerung von n und K keine Verbesserung der Anpassung erbringt. In diesem Fall kann man nur die NH-Verteilung verwenden, wie an dem Beispiel von Brainerd gezeigt wird. (s. Tab. 7.5). Mit n = 10 erhalten wir ein X^2_7 = 6.04, mit n = 11 ein X^2_7 = 6.10, und mit wachsendem n und K wird die Anpassung immer schlechter. Die negative Binomialverteilung (NB) ergibt eine viel schlechtere Anpassung, obwohl die Abweichung auch hier nicht signifikant ist.

An einem weiteren Beispiel zeigen wir, daß die Anpassung auch ziemlich kompliziert sein kann. Piotrowski, Bektaev, Piotrovskaja

(1985:217) haben die Substantive in 400 Passagen von jeweils 25 Wortverwendungen in dem kazachisch geschriebenen Roman "Put' Abaja" von M. Auezov gezählt und erhielten die Verteilung in Tabelle 7.6. Die NH-Verteilung in Spalte 3 mit recht großen Parametern liefert ein sehr zufriedenstellendes Resultat (P = 0.35). Vergrößert man aber

Tabelle 7.5

Anpassung an die Verteilung des Artikels in *Cheevers Wapshot Chronicle*

(Daten aus BRAINERD 1972)

x	f_x	NH	NB
0	8	7.26	5.38
1	8	9.67	10.21
2	12	10.63	12.46
3	11	10.76	12.39
4	14	10.30	10.92
5	4	9.42	8.89
6	8	8.20	6.83
7	8	6.74	5.03
8	6	5.09	3.58
9	2	3.34	2.48
10	2	1.59	4.83
		n = 10	M = 3.4998
		K = 3.7090	p = 0.4575
		M = 1.4950	-
		X^2 = 6.04	10.82
		FG = 7	8
		P = 0.54	0.21

n auf 16, so erhält man ein X^2 = 10.63, also ein etwas schlechteres Resultat als mit n = 15, und ein weiteres Anwachsen ergibt noch schlechtere Anpassungen. Das Kriterium (7.17) zeigt, daß die Binomialverteilung angepaßt werden könnte. Die Autoren (P/B/P) haben die Anpassung in Spalte 4 erhalten, die etwas schlechter ist als die mit NH erzielte. Erhöht man das n der Binomialverteilung, so stellt man fest, daß sich die Anpassung noch etwas verbessern läßt. Geht man jedoch vom Kriterium (7.18) aus, dann müßte die Poisson-Verteilung eher geeignet sein. Man bekommt jedoch das beste Resultat mit a = 6.6757, X^2_{12} = 17.28, P = 0.14. Durch geschicktes Zusammenfassen einiger Häufigkeitsklassen (was durchaus legitim ist), z.B. der ersten drei, lassen sich die

Resultate beträchtlich verbessern, was jedoch bei der NH-Verteilung keineswegs nötig ist.

Zahlreiche andere Überprüfungen des Frumkina-Gesetzes für mehrere Sprachen findet man in Altmann, Burdinski (1982) oder Piotrowski, Bektaev, Piotrovskaja (1985).

Tabelle 7.6

Verteilung von Substantiven in Passagen aus Auezovs Roman "Put'Abaja"

nach Piotrowski, Bektaev, Piotrovskaja (1985).

x	fx	NH	Bin
1	3	0.56	0.63
2	5	3.37	2.96
3	7	10.58	9.71
4	24	23.15	22.88
5	40	39.20	41.21
6	52	54.28	58.87
7	64	63.32	68.48
8	66	63.24	66.03
9	47	54.44	53.44
10	48	40.37	36.64
11	24	25.57	21.44
12	14	13.59	10.72
13	3	5.87	4.60
14	2	1.95	1.68
15	1	0.50	0.69
	n = 15		n = 25
	K = 25.2950		p = 0.3
	M = 11.1791		-
	X^2 = 9.80		X^2 = 13.69
	FG = 9		FG = 10
	P = 0.35		P = 0.19

Daten über Verteilung von "kal" in Sätzen im Werk von Isokrates findet man bei Morton, Levison (1966). Wie man sieht, hat hier der Textblock (Passage) keine konstante Größe, denn die Satzlänge variiert bei Isokrates. Ebenso haben Altmann, Burdinski (1982) als Block eine gedruckte Seite genommen, was eine praktische Approximation ist, wenn man die Auszählung ohne Computer vornimmt. Alle Verteilungen bei Morton und Levison sind sehr gut und zeigen, daß hier die negative Binomialverteilung am einfachsten anzupassen ist.

7.3 Ausblick

Aus dem Verfahren in §7 kann man folgende Lehren ziehen:

(1) Das vorgestellte Modell zeigt, daß man zur Zeit bei der Aufstellung von textuellen Gesetzeshypothesen oft auf Annahmen angewiesen ist, die sich später als glückliche oder unglückliche Treffer erweisen können. Die Wahl der Beta-Verteilung als Wahrscheinlichkeitsfunktion von p ist nur der erste Ansatz. Ebenso könnte man auch für n eine diskrete Verteilung wählen, aber bisher hat sich das nicht als notwendig erwiesen.

(2) Weiter sieht man, daß eine Erscheinung mit mehreren Modellen erfaßt werden kann, wobei sich diese Modelle nicht widersprechen, sondern als Spezialfälle (Grenzfälle) eines allgemeineren Modells zu betrachten sind. Dies kann in der Texttheorie von großem Nutzen sein, denn es besteht die Möglichkeit, daß eine Texteinheit sich in unterschiedlichen Texten so heterogen verhält, daß das angewendete Modell nicht nur für die Einheit, sondern auch für den Text charakteristisch ist.

(3) Wahrscheinlichkeitsverteilungen mit mehreren Parametern haben sowohl Vorteile als auch Nachteile. Der Nachteil einer schweren Schätzung der Parameter und einer langwierigen Rechnung entfällt durch Anwendung moderner Optimierungstechniken und Computer, es bleibt aber der Nachteil der beschwerlichen Interpretation der Parameter. Die Parameter stehen sicherlich in Beziehung zu den Eigenschaften der Texteinheiten, und es bedarf mühsamer und langwieriger Arbeit an zahlreichen Texten in vielen Sprachen, bis man imstande sein wird, sie zu interpretieren. Der Vorteil liegt darin, daß eine Wahrscheinlichkeitsverteilung mit mehreren Parametern wendiger ist und sich relativ gut auch "pathologischen" Daten anpaßt. Gleichzeitig zwingt sie zur Suche nach Zusammenhängen, die den Nährboden einer künftigen Texttheorie bilden.

Piotrowski (1984) nennt folgende mögliche Anwendungen einer Theorie der blockmäßigen Wortwiederholungen:

(i) Mechanische Bestimmung der Zugehörigkeit eines Wortes zu einer Wortklasse.

(ii) Identifizierung von terminologisch oder semantisch dominanten Texteinheiten.

(iii) Feststellung und Messung der stilistischen Individualität des Textes.

(iv) Diagnostizierung der Schwerpunkte einiger psychischer Krankheiten (vgl. Paškovskij, Srebrjanskaja 1971).

(v) Konstruktion von lernenden Automaten.

Bei dem oben dargestellten komplexen Modell kann man folgenden Problemen nachgehen:

(i) Wie verändern sich die einzelnen Parameter, wenn die Blockgröße zunimmt?

(ii) Gibt es Wortarten oder Einzelwörter, die eine der vier Verteilungen in allen Texten bevorzugen?

(iii) Mit welchen anderen Faktoren hängen die Parameter der obigen Verteilungen zusammen? Als solche Faktoren kann man z.B. die Worthäufigkeit, den Bedeutungsreichtum des Wortes, die Textsorte usw. untersuchen.

(iv) Läßt sich die negative hypergeometrische Verteilung aus dem Ansatz in § 2.4 wohlinterpretiert ableiten?

(v) Kann man die Verteilung von p in (7.3) und (7.4) auch anders wählen? Wenn ja, wie?

Jedes neuformulierte Gesetz beleuchtet nicht nur neue Aspekte der Daten, sondern bringt auch zahlreiche neue Probleme mit sich, die sich lawinenartig vermehren. In diesem Fall wäre zunächst wichtig, zahlreiche Texte zu untersuchen und einfach empirische Resultate zu bringen.

8. PARALLELE WIEDERHOLUNG

Die bekannteste parallele Wiederholung ist der Reim. Man kann ihn auch unter die positionale Wiederholung einordnen, dort aber handelt es sich vielmehr um die Wiederholung ein und derselben Einheit in einer Position. Der Reim wird uns schon aus dem Grunde hier nicht interessieren, weil er bewußt quasi deterministisch positioniert wird. Beim Parallelismus handelt es sich eher um die Gleichgestaltung zweier Hypereinheiten, wobei die konstituierenden Einheiten nicht unbedingt identisch sein müssen. Es handelt sich eher um eine Spiegelung eines Gedankens, eines Bildes (Bildnisses), einer grammatischen Struktur, einer lautlichen Struktur, oft mehreres hintereinander kombiniert. Seine Funktion besteht in der Verstärkung einer Entität, ihrer Metaphorisierung, ihrer Übertragung in andere Bereiche. Man findet den Parallelismus wohl in jeder Volkspoesie (eine Übersicht vgl. z.B. bei Newman, Popper 1918), in magischen Beschwörungen, in Litaneien, in Sprichwörtern, sie sind auch ein stilistisches Mittel in modernen Texten.

Einzelne Parallelismen sind leicht zu ermitteln; schwieriger wird das Problem, wenn in einer Klasse von Texten nur eine ***Tendenz*** zum Parallelismus vorhanden ist, die nicht in jedem einzelnen Fall belegbar ist. Ein tendentieller Parallelismus ist nur statistisch zu ermitteln. Zu diesem Zweck werden wir hier drei einfache Methoden vorstellen. Diese werden anhand von malayischen Pantuns (Quatrine) illustriert, die laut Wilkinson (1907) so konstruiert sind, daß die zweite Halbstrophe eine parallele Assonanzstruktur mit der ersten Halbstrophe aufweist. Unter Assonanz verstehen wir hier die Identität der beiden Vokale des Stammes malayischer Wörter, die etwa zu 84% zweisilbig sind.

Betrachtet man den Pantun in § 3.3, so kann man ihn folgendermaßen kodieren:

1	2	3	4
Anak	*bĕruk*	*(di)kayu*	*rĕndang,*
5	6	7	8
Turun	*mandi*	*(di)dalam*	*paya*
9	10	11	12
Hodoh	*buruk*	*(di)mata*	*orang*
13	14	15	16
cantik	*manis*	*(di)mata*	*sahaya*

Die Zahlen bedeuten die einzelnen Positionen, in denen die Assonanz untersucht wird, die Klammern trennen Präfixe ab. Das Wort in der 16-ten Position ist zwar dreisilbig, weist aber trotzdem eine Assonanz mit dem Wort in der 8-en Position auf.

Man sieht, daß das Positionenpaar (i, i + 8) nicht immer eine Assonanz aufweist, wie z.B. (1, 9): *anak/hodoh* oder (2, 10): *beruk/buruk* usw., die nicht assonant sind, jedoch (1, 7, 8, 11) oder (13, 14) oder (5, 10), die assonant sind, was alles gegen die Hypothese von Wilkinson spricht. Für seine Hypothese sprechen die Paare (6, 14), (7, 15), (8, 16). Um eine Entscheidung darüber treffen zu können, ob es hier keine signifikanten Assonanzen oder nur parallele Assonanzen oder auch nichtparallele Assonanzen gibt, müssen wir statistisch verfahren.

8.1. Ein Vortest: Cochran's Q-Test

Man pflegt bei empirischen Untersuchungen, zuerst immer "über den Daumen peilend", mit einem schnellen und einfachen Test auszuloten, ob sich umfangreichere, langwierigere Auszählungen und Rechnungen lohnen würden. Sehr oft benutzt man dazu parameterfreie Tests, die man ohne große Mühe anwenden kann. Hier werden wir den Q-Test von Cochran (1950, vgl. auch Siegel 1956: 161-166) vorstellen.

Wir nehmen zufällig 20 Pantuns aus einer Sammlung malayischer Pantuns (Pantoen Melajoe. Weltevreden 1929) und untersuchen die Assonanz in den Positionen vor der Zäsur, d.h., die Worte in den Positionen 2, 6, 10 und 14. Wenn in einem Positionspaar eine Assonanz (= Identität der Stammvokale) vorkommt, dann bewerten wir sie mit 1, sonst mit 0. Die Resultate der Untersuchung tragen wir in Tabelle 8.1. ein. Die Spaltensummen bezeichnen wir mit G_j (j = 1,...,K), die Zeilensummen mit L_i (i = 1,...,n) und berechnen die Größe

$$Q = \frac{(k-1)\left[k \sum_{j=1}^{k} G_j^2 - \left(\Sigma G_j\right)^2\right]}{k \sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2} \qquad (8.1)$$

Tabelle 8.1

Vorkommen der Assonanz in vorzäsuralen Positionen.

Pantun	Positiospaare						L_i	L_i^2
	2-6	2-10	2-14	6-10	6-14	10-14		
1	0	1	0	0	1	0	2	4
2	0	1	0	0	1	0	2	4
3	0	0	0	0	1	0	1	1
4	0	0	0	0	1	0	1	1
5	0	1	0	0	1	0	2	4
6	0	1	0	1	1	1	4	16
7	1	1	0	1	0	0	3	9
8	0	1	0	0	0	0	1	1
9	0	1	0	0	1	0	2	4
10	0	1	0	1	0	0	2	4
11	0	0	0	0	0	1	1	1
12	0	1	1	0	1	1	4	16
13	0	1	1	0	0	1	3	9
14	0	1	0	0	1	0	2	4
15	0	0	1	0	0	0	1	1
16	0	1	0	0	1	0	2	4
17	0	0	0	1	1	0	2	4
18	0	1	0	0	1	0	2	4
19	1	1	1	0	1	0	4	16
20	0	1	0	0	1	0	2	4
	2	15	4	4	14	4	43	111
	G_1	G_2	G_3	G_4	G_5	G_6	ΣL_i	ΣL_i^2

wobei k die Zahl der Spalten (hier k = 6) und n die Zahl der Zeilen (hier n = 20) ist. Setzen wir die Zahlen aus Tabelle 8.1 in die Formel (8.1) ein, dann bekommen wir

$$Q = \frac{(6-1)\left[6(2^2+15^2+4^2+4^2+14^2+4^2)-43^2\right]}{6(43) - 111} = 33.64.$$

Die Größe Q ist ungefähr wie ein Chiquadrat mit k-1 Freiheitsgraden verteilt. Der kritische Wert des Chiquadrats mit 6-1 = 5 Freiheitsgraden auf $\alpha = 0.05$ ist 11.1. Da unser berechneter Wert größer als der theoretische Wert ist, schließen wir, daß die Assonanz nicht gleichmäßig unter

den untersuchten Positionen verteilt ist, sondern daß es eine signifikante Tendenz gibt. Man sieht schon den Ausprägungen der Spalten (G_j) an, wo die Tendenz liegt, aber diese Methode gibt nur den Anreiz zu weiteren Untersuchungen, eine Verallgemeinerung sollte man nur als Hypothese in Erwägung ziehen.

8.2. Varianzanalytische Untersuchung

Der folgende Ansatz gibt eine globale Antwort auf die Plazierung der Assonanz und ist rechnerisch relativ leicht durchzuführen.

Man schreibe sich die 32 Vokale des obigen Pantuns (nach Abtrennung der Affixe), die in den 16 Stämmen vorkommen (in "sahaya" wurden zwei "a" genommen), hintereinander auf zwei Papierstreifen. Legt man die Papierstreifen untereinander, so sind die beiden Vokalsequenzen gleich. Verschiebt man den unteren um einen Vokal (einen Schritt) nach rechts, dann bekommt man

```
a a ě u a u ě a u u a i a a a a
  |           |       | | |
  a a e u a u ě a u u a u a a a
```

```
o o u u a a o a a i a i a a a a
  |   |   |     |       | | |
  a o o u u a a o a a i a i a a a a .
```

Jetzt liegen 12 identische Vokale untereinander (verbunden mit einem Strich). Das Übereinstimmungsmaß beträgt 12/31 = 0.38709, wo 31 die Zahl der Vergleiche ist. Schiebt man einen Schritt weiter, so erhält man 10/30 = 0.33333. Auf diese Weise rechnen wir die ersten k = 17 Schritte (Verschiebungen) in n = 50 Pantuns durch und führen für diese (17x50) 850 Ausprägungen eine Varianzanalyse durch. Wir bezeichnen hier

x_{ij} = Übereinstimmungsmaß im Schritt i (i = 1,2,...,17) des Pantuns j (j = 1,2,...,50).

$\bar{x}_i$ = Durchschnittliche Übereinstimmung im Schritt i, d.h.

$$\bar{x}_i = (1/n) \sum_{j=1}^{n} x_{ij} .$$

$\bar{x}$ = Durchschnittliches Übereinstimmungsmaß in der ganzen Stichprobe.

SAQ_Z = Summe der Abweichungsquadrate zwischen den Schritten $= n \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$ mit k-1 Freiheitsgraden.

SAQ_I = Summe der Abweichungsquadrate innerhalb der Schritte $= \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$ mit N-k Freiheitsgraden

SAQ_T = Die gesamte Summe der Abweichungsquadrate = $SAQ_Z + SAQ_I = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x})^2$ mit N-1 Freiheitsgraden.

Anschaulichkeitshalber geben wir die einzelnen $\bar{x}_i$ im Schritt i an:

Schritt i	$\bar{x}_i$
1	0.3542
2	0.3620
3	0.3546
4	0.3464
5	0.3430
6	0.3661
7	0.3408
8	0.3760
9	0.3382
10	0.3690
11	0.3237
12	0.3740
13	0.3632
14	0.3689
15	0.3447
16	0.5298
17	0.3547

Die Resultate der Rechnung findet man in Tabelle 8.2.

Tabelle 8.2

Varianzanalyse der Assonanz in 50 Pantuns (nach Altmann 1963)

Variabilität	SAQ	FG	Varianz	F-Test	P
zwischen Schritten (Z)	1.6008	16	0.1000	7.479	1.35×10^{-16}
innerhalb der Schritte (I)	11.1434	833	0.0134		
Total	12.7442	849			

Der F-Test zeigt, daß die Variabilität zwischen den 17 Schritten als signifikant groß zu betrachten ist. Man kann nun testen, in welchem Schritt sich die Ausprägung ($\bar{x}_i$) von dem Durchschnitt ($\bar{x}$) stark unterscheidet. Wie man sieht, ist die Ausprägung am größten im Schritt 16 ($x_{16} = 0.5298$). Wir überprüfen die Differenz mit Hilfe eine t-Tests nach der Formel

$$t = \frac{\bar{x}_{16} - \bar{x}}{s/\sqrt{n}} , \qquad (8.2)$$

wo $s^2 = SAQ_I/(N-k) = 0.013377$, und bekommen

$$t = \frac{0.5298 - 0.3652}{0.115/\sqrt{50}} = 10.07 ,$$

was mit 833 Freiheitsgraden einer Wahrscheinlichkeit von $P \approx 1.39\times10^{-20}$ entspricht.

Der Test gibt global an, daß genau in Schritt 16 eine starke Assonanz besteht, was einem Parallelismus der beiden Halbstrophen entspricht.

Wie jedoch die Stärke dieser Assonanz in einzelnen Positionen beschaffen ist, das muß mit gesonderten Tests für jede Position ermittelt werden.

8.3. Der Chiquadrat-Test

Untersuchungen mit dieser Methode wurden von Sebeok, Zeps (1959) in der tscheremissischen Volkspoesie und von Altmann (1963) in der malayischen Volkspoesie durchgeführt.

Zur detaillierten Überprüfung der Assonanzstruktur des Pantuns verfahren wir folgendermaßen. Wir wählen zufällig 100 Pantuns, die für die Untersuchung geeignet sind (z.B. keine monosyllabischen Wörter enthalten). Im Malayischen gibt es 6 Vokale (a, ĕ, i, o, u, e), d.h. 36 vokalische Muster [aa, aĕ, ai, ...]. Wir wählen folgende Bezeichnungen:

N = Zahl der unterschiedlichen Pantuns (hier 100);

M = Menge der vokalischen Muster, M = {aa, aĕ, ai, ...};

v = ein beliebiges vokalisches Muster $v \in M$, $|v| = 36$;

i,j = zwei Positionen im Pantun, $i \neq j$; $i,j = 1,2,\ldots,16$;

$f_i(v)$, $f_j(v)$ = beobachtete Anzahl der Pantuns, die in der Position i bzw. j das Muster v haben;

$f_{ij}(v)$ = erwartete Anzahl der Pantuns, die in den Positionen i und j gleichzeitig das Muster v haben.

Diese Größe berechnen wir unter der Annahme der Unabhängigkeit, wie in früheren Kapiteln, als

$$f_{ij}(v) = \frac{f_i(v) f_j(v)}{N} . \qquad (8.3)$$

Summiert man (8.1) über alle Muster v in den Positionen i, j, dann bekommt man

$$E_{ij}(A) = \sum_{v \in M} f_{ij}(v) , \qquad (8.4)$$

d.h. die erwartete Anzahl der Pantuns mit einer Assonanz in den Positionen i und j. Die Größe (8.1) muß für jedes Muster separat berechnet

werden. Dagegen erhalten wir durch einfache Auszählung der Assonanzen (beliebiges v) aus den Pantuns die Zahl

$O_{ij}(A)$ = beobachtete Zahl der Pantuns, die in den Positionen i und j eine Assonanz haben.

Aus diesen Zahlen setzen wir das Chiquadrat-Kriterium wie folgt zusammen:

$$X^2 = \frac{[O_{ij}(A) - E_{ij}(A)]^2}{E_{ij}(A)} + \frac{[(N-O_{ij}(A)) - (N-E_{ij}(A))]^2}{N - E_{ij}(A)}$$

$$= \frac{N[O_{ij}(A) - E_{ij}(A)]^2}{E_{ij}(A)[N - E_{ij}(A)]} \qquad (8.5)$$

Tabelle 8.3

Assonanz im malayischen Pantun

Erstes Glied i	Zweites Glied des Positionspaares													
	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8	i+9	i+10	i+11	i+12	i+13	i+14	i+15
1	2.55	4.15	0.02	0.68	0.04	0.13	0.59	***12.48***	1.77	0.01	0.66	2.62	0.92	4.15	7.09
2	6.14	0.97	0.66	1.15	0.00	0.12	8.46	***53.75***	2.20	1.60	0.12	0.12	4.11	0.25	
3	0.43	5.98	0.02	2.61	0.09	2.21	0.00	2.36	0.57	0.01	0.04	0.73	0.57		
4	3.31	0.01	0.43	0.20	1.50	0.05	1.99	***119.75***	0.12	1.79	0.17	0.70			
5	1 59	0.08	0.01	1.03	1.20	1.53	2.17	1.06	0.58	2.11	4.90				
6	1.06	4.61	0.01	2.38	0.73	0.26	2.39	***34.45***	0.01	4.56					
7	0.61	1.38	0.38	0.02	0.00	0.28	0.04	1.87	0.26						
8	0.89	0.11	0.63	0.50	0.24	0.47	0.47	***70.86***							
9	0.64	0.01	3.07	1.68	0.36	0.06	1.39								
10	0.24	0.04	0.99	0.49	1.33	2.94									
11	6.11	0.70	0.11	0.18	2.48										
12	0.48	0.84	0.18	0.81											
13	0.01	0.01	0.05												
14	1.40	0.00													
15	1.13														

Die hervorgehobenen Werte sind signifikant auf $\alpha = 0.0005$

oder einfach

$$X^2 = \frac{N(O - E)^2}{E(N - E)} \,. \qquad (8.6)$$

Diese Größe ist wie ein Chiquadrat mit 1 Freiheitsgrad verteilt. Der kritische Wert auf $\alpha = 0.05$ ist 3.84. Wenn ein berechnetes X^2 nach (8.6) größer als 3.84 ist, unter der Bedingung, daß $O > E$, dann kann man von einer signifikanten Assonanz sprechen. Die Resultate dieser Untersuchung sind in Tabelle 8.3 enthalten.

Obwohl man vereinzelt in mehreren Spalten signifikante Werte findet, kann man eine systematische Assonanzstruktur nur für die Paare (i,i+8) annehmen, d.h. eben für die phonischen Parallelismen.

Da das vokalische Muster "a-a" im Malayischen so häufig ist, daß es kaum eine phonische Assoziation hervorruft, wurden die Rechnungen auch ohne die "a-a" Muster durchgeführt. Die Resultate sind in Tabelle 8.4 dargestellt. Hier tritt die Spalte i+8 noch stärker in Erscheinung als in Tabelle 8.3.

Tabelle 8.4

Assonanz im malayischen Pantun ohne das Muster "a-a"

Erstes Glied i	Zweites Glied des Positionspaares													
	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8	i+9	i+10	i+11	i+12	i+13	i+14	i+15
1	4.19	1.72	0.02	0.49	0.27	0.00	0.07	***24.88***	1.80	0.48	0.82	4.10	1.07	5.00	4.45
2	0.43	1.05	1.52	0.00	0.54	1.30	4.83	***71.76***	0.46	0.58	0.07	0.71	3.54	0.89	
3	3.78	6.23	0.01	2.13	3.10	1.99	0.76	10.41	1.10	0.04	1.59	1.41	2.61		
4	2.52	0.88	0.02	2.00	1.59	0.09	1.19	***75.52***	0.61	0.74	0.03	0.31			
5	0.01	0.10	0.09	0.02	1.82	1.58	2.02	4.68	3.52	1.46	6.80				
6	1.01	3.44	0.15	2.47	3.54	0.05	0.22	***37.36***	0.35	5.12					
7	0.61	2.47	2.04	0.00	0.22	2.26	0.00	2.57	0.86						
8	0.41	0.37	0.30	0.55	0.16	0.07	0.12	***79.38***							
9	0.76	1.64	6.53	1.32	0.01	0.37	3.56								
10	1.65	0.67	2.42	0.06	0.65	3.58									
11	3.53	0.13	0.08	0.21	2.68										
12	0.00	0.16	0.59	0.13											
13	1.75	0.09	0.01												
14	0.62	0.25													
15	0.57														

Die hervorgehobenen Werte sind signifikant auf $\alpha = 0.0005$

Vielleicht noch stärker als die Assonanz ist die Tendenz zum inneren Reim im malayischen Pantun in parallelen Positionen. Betrachtet man als Reim die Identität der letzten zwei Laute des Wortes, so erhält man die Resultate, wie in Tabelle 8.5 dargestellt. Am stärksten ist diese Tendenz natürlich in den Positionspaaren (4,12) und (8,16), wo man sie (deterministisch) erwartet, aber man sieht auch, daß sie in allen (i,i+8) Positionspaaren stärker ist als die Assonanz.

Durch eine derartige einfache Berechnung kann man also unerwartet ganz neue Aspekte einer Textsorte entdecken. Die hier verwendeten Methoden kann man natürlich auch für die Untersuchung anderer "paralleler" Erscheinungen verwenden.

Tabelle 8.5

Reim im malayischen Pantun

Erstes Glied	Zweites Glied des Positionspaares														
i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8	i+9	i+10	i+11	i+12	i+13	i+14	i+15
1	7.44	3.76	1.31	1.75	1.87	3.30	0.83	***59.03***	2.75	1.95	0.35	0.67	0.01	1.83	2.50
2	0.37	4.57	0.08	0.48	3.19	0.69	4.74	***415.24***	5.52	0.05	0.91	2.42	0.13	0.63	
3	0.93	1.15	2.51	0.00	0.85	1.06	0.03	***27.18***	0.86	0.76	0.17	1.06	0.83		
4	1.06	0.52	0.50	3.54	0.49	0.18	0.83	***1462.40***	0.00	0.67	1.83	3.46			
5	1.37	1.59	0.23	0.11	0.12	0.13	4.57	***17.16***	0.97	0.94	0.05				
6	1.23	0.03	0.85	0.29	0.49	0.04	2.22	***182.35***	2.10	0.01					
7	1.34	0.34	7.55	2.15	0.09	0.77	0.46	***17.19***	1.39						
8	1.13	0.18	0.28	2.68	0.04	0.00	7.02	***1713.13***							
9	0.87	0.54	0.28	0.02	0.01	2.02	1.19								
10	0.34	1.50	0.01	0.52	0.16	0.11									
11	0.65	0.26	1.72	0.38	0.26										
12	0.23	0.96	0.62	3.65											
13	0.07	0.31	0.04												
14	2.68	0.00													
15	6.61														

Die hervorgehobenen Werte sind signifikant auf $\alpha = 0.0005$

9. ZYKLISCHE WIEDERHOLUNG

Zyklische Wiederholungen rufen eine Art Wellenbewegung im Text hervor. Sie sind am leichtesten dann zu handhaben, wenn die Eigenschaften der sich zyklisch wiederholenden Elemente numerische Werte annehmen können. Da dies in vielen Fällen möglich ist, eröffnet sich hier ein sehr umfangreiches Forschungsfeld, das nicht nur an Methoden, sondern auch an Problemen sehr reich ist.

Die bekanntesten Untersuchungen dieser Art stammen aus der Poetik, wo man auch heute noch oft die Proportionen der akzentuierten Silben an allen Positionen des Verses zählt und dann eine schwingende Kurve zeichnet. Selten ist man weitergegangen (jedoch vgl. z.B. Grotjahn 1979), obwohl es nötig wäre, nicht nur den Datenverlauf mit einer Kurve zu erfassen, sondern die Kurve auch zu begründen. Es bieten sich hier die Fourier-Analyse, die Theorie der Zeitreihen, der Markov-Ketten u.a. an, so daß wir uns notgedrungen einschränken müssen.

Hier seien einige Beispiele solcher Verläufe genannt.

(i) Bezeichnet man die Zahl der Daktyle in den ersten vier Positionen des Hexameters mit 0, 1, 2, 3, 4, dann kann man z. B. die ersten 30 Verse des Gedichts von Bridges (vgl. § 6.1) folgendermaßen schreiben:

1 1 1 2 1 2 3 1 1 0 1 2 1 0 2 2 1 1 2 3 3 3 2 1 2 2 1 2 1 1

Die graphische Darstellung ist in Abbildung 9.1 zu finden. Die Frage ist, ob es in dieser Folge eine periodische Schwingung gibt, und wenn ja, welcher Art.

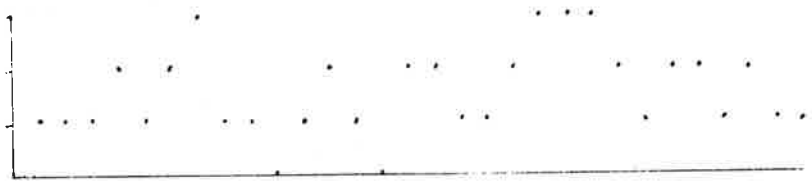

Abb. 9.1. Zahl der Daktyle in den ersten 30 Versen des Gedichts von Bridges

(ii) In Bottos Gedicht "Smrt' Jánošíkova" (Slovakisch) ergeben sich die Proportionen der betonten Silben auf den einzelnen Positionen des Verses wie folgt (Kochol 1968):

82, 27, 31, 52, 43, 5, 90, 16, 37, 41, 39, 5

Die graphische Darstellung in Proportionen ist in Abbildung 9.2 zu sehen.

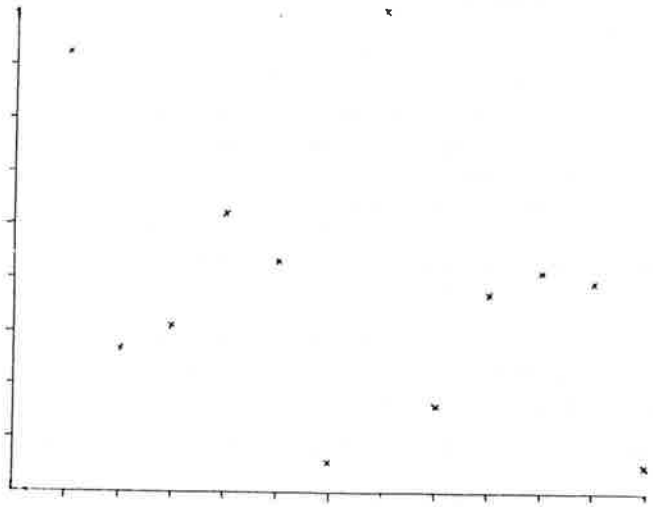

Abb. 9.2. Betonungsverlauf im Bottos Gedicht (Kochol 1968)

(iii) In Heisenbergs "Der Teil und das Ganze" (1. Kapitel) lautet die Folge von Satzlängen, gemessen in der Zahl der Teilsätze wie folgt:

1 1 6 7 2 9 2 3 1 3 5 2 5 1 1 3 2 1 5 1 1 4 1 3 3 ...

Die graphische Darstellung ist in Abbildung 9.3 zu finden.

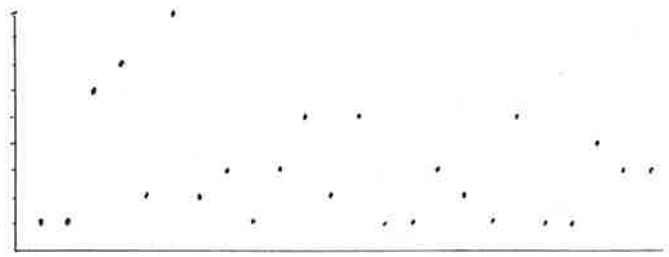

Abb. 9.3. Satzlängenfolgen bei Heisenberg

Nach der Betrachtung solcher Verläufe stellt sich automatisch die Frage, ob man in ihnen irgendwelche Regularitäten entdecken kann, ob es sich um einfache oder um überlagerte Schwingungen handelt, welche Regularitäten für unterschiedliche Einheiten und Textsorten gelten, ob es Unterschiede zwischen Texten gleicher Sorten in unterschiedlichen Sprachen gibt, wovon die Schwingungsperiode abhängt usw.

Die Daten in Beispiel (ii) unterscheiden sich von denen in Beispiel (i) und (iii) dadurch, daß es sich hier um einen Verlauf durch die Positionen des Verses handelt, wobei nicht nur eine Position in Betracht gezogen wird, wie in Kapitel 3, sondern alle gleichzeitig. Die Positionen sind "numerierbar" und festgelegt. In den Beispielen (i) und (iii) ist dies nicht der Fall, man kann die Reihe an beliebiger Stelle anfangen und an beliebiger Stelle beenden; falls sie lang genug ist, müßte die Rechnung immer zu ähnlichen Resultaten führen. In Beispiel (ii) wird der Verlauf von der Sprache festgelegt, denn im Slovakischen liegt der Hauptakzent auf der ersten Silbe, die Nebenakzente auf den ungeraden Silben; in (i) und (iii) pendeln sich die Dichter möglicherweise in einen Rhythmus ein, der recht kompliziert sein und durch Korrekturen ständig neue Abweichungen aufnehmen kann; hier darf man davon ausgehen, daß der aktuelle Wert der Variablen von den vorangehenden Werten abhängt.

Alle diese Verläufe lassen sich als Zeitreihen auffassen, wobei wir in den Texten wahrscheinlich nur mit solchen zu tun haben werden, die in einem Gleichgewicht um einen Mittelwert bleiben und als stationär bezeichnet werden. Trends gehören in einen anderen Bereich.

In Anbetracht der zahlreichen Methoden, die in den letzten Jahrzehnten für die Untersuchung von Zeitreihen entwickelt worden sind (vgl. z.B. Box, Jenkins 1970; Pandit, Wu 1983; Schlittgen, Streitberg 1984; Grotjahn 1981), werden wir uns notgedrungen einschränken müssen, da es

zunächst sowohl an Daten mangelt als auch an theoretischen Einsichten über die sprachliche Natur der textbildenden Prozesse.

9.1 Fourier-Analyse

Will man eine zyklische Regularität lediglich beschreiben, so eignet sich dazu sehr gut die Fourier-Analyse, mit deren Hilfe man die Amplituden der durch Störung (Rauschen) verdeckten Frequenzen ermitteln kann. Als Resultat bekommen wir eine Überlagerung von sinus- und cosinus-Schwingungen, mit denen wir die beobachtete Reihe approximieren, d.h., den deterministischen Teil der Regularität von dem stochastischen trennen können. Wir nehmen also an, daß wir die beobachteten Werte y_x ($x = 1,2,...,N$) mit Hilfe des Modells

$$y_x = A_0 + \sum_{i=1}^{q}[A_i \cos(2\pi f_i x) + B_i \sin(2\pi f_i x)] + e_x \qquad (9.1.1)$$

erfassen können. Hier sind A_0, A_i, B_i ($i = 1,...,q$) Parameter, f_i ist die i-te Harmonische der grundlegenden Frequenz 1/N, d.h.

$$f_i = \frac{i}{N},$$

und q berechnet sich fogendermaßen: Wenn N gerade ist, dann ist $q = N/2$; wenn N ungerade ist, dann ist $q = (N-1)/2$.

Die einzelnen Koeffizienten schätzt man mit Hilfe der Methode der kleinsten Quadrate als

$$A_0^* = \bar{y} \qquad (9.1.2)$$

$$A_i^* = \frac{2}{N} \sum_{x=1}^{N} y_x \cos(2\pi f_i x) \qquad (9.1.3)$$

$$B_i^* = \frac{2}{N} \sum_{x=1}^{N} y_x \sin(2\pi f_i x), \qquad (9.1.4)$$

wo $f_i = i/N$ und $i = 1,2,...,q$. Die *Intensität* $I(f_i)$ bei der Frequenz f_i ergibt sich als

$$I(f_i) = \frac{N}{2}(A_i^2 + B_i^2). \qquad (9.1.5)$$

Wenn N gerade ist, dann ergeben sich die q-ten Koeffizienten als

$$A_q^* = \frac{1}{N} \sum_{x=1}^{N} (-1)^x y_x$$
$$B_q^* = 0 \qquad (9.1.6)$$

und die Intensität ist

$$I(f_q) = NA_q^2. \qquad (9.1.7)$$

Die Intensitäten ergeben das *Periodogram* der Reihe, und ihre Summe gleicht der Summe der quadratischen Abweichungen der gemessenen Werte von ihrem Mittelwert, d.h.

$$\sum_{i=1}^{q} I(f_i) = \sum_{x=1}^{N} (y_x - \bar{y})^2, \qquad (9.1.8)$$

was als Kontrolle bei den Rechnungen dienen kann. Gleichzeitig stellt $I(f_i)$ den Anteil der Koeffizienten A_i und B_i an der Gesamtvarianz dar.

Die Rechnung illustrieren wir an der Akzentulerung der Silben in Bottos "Smrt' Jánošíkova" nach Kochol (1968), vgl. Tabelle 9.1.

Variable X gibt die Positionen im Vers an. Die rohen Daten, die Proportionen darstellen, sind in der zweiten Zeile (y_x) aufgeführt.

Wir berechnen A_i und B_i nach den Formeln (9.1.3) und (9.1.4). Zu diesem Zweck müssen wir erst $\cos(2\pi(1/N)x)$ und $\sin(2\pi(1/N)x)$ ermitteln, vgl. Tabelle 9.1.

$$A_1 = \frac{2}{12}[82(0.87) + 27(0.5) + ... + 39(0.87) + 5(1.0)] = -1.74$$

$$B_1 = \frac{2}{12}[82(0.5) + 27(0.87) + ... + 39(-0.5) + 5(0.0)] = 1.85$$

Tabelle 9.1

Berechnung der Koeffizienten der Fourier–Reihe für die Daten von Kochol

x	1	2	3	4	5	6	7
y_x	82	27	31	52	43	5	90
cos(2πx/N)	0.87	0.5	0.0	-0.5	-0.87	-1.0	-0.87
sin(2πx/N)	0.5	0.87	1.0	0.87	0.5	0.0	-0.5

x	8	9	10	11	12
y_x	16	37	41	39	5
cos(2πx/N)	-0.5	0.0	0.5	0.87	1.0
sin(2πx/N)	-0.87	-1.0	-0.87	-0.5	0.0

Alle Koeffizienten sind in Tabelle 9.2 aufgeführt. Die Kontrolle ergibt $s^2 = (1/N)\Sigma(y_x - \bar{y})^2 = 641$, was mit 7692/12 = 641 übereinstimmt.

Im nächsten Schritt geht es darum, aus den berechneten Koeffizienten einige wenige so zu wählen, daß die Gesamtvarianz maximal reduziert wird. Die Signifikanz der Koeffizienten oder das I(f) kann man auch testen (vgl. Anderson 1971:102ff), aber es wurde des öfteren gezeigt, daß es zu widersprüchlichen Resultaten führt (vgl. Tintner 1965:223 ff). Man kann die Koeffizienten also mit der trial-and-error Prozedur wählen und kombinieren, oder man eliminiert einige aufgrund von linguistischen Annahmen.

Wie man in Abbildung 9.2 sieht, ist der Vers in zwei rhythmische symmetrische Teile von jeweils 6 Silben aufgeteilt. Die Halbverse kann man dann auf drei verschiedene Weisen gleichmäßig aufteilen, nämlich auf

Tabelle 9.2

Fourier–Analyse der Daten aus Tabelle 9.1

i	f_i	Periode	A_i	B_i	$I(f_i)$	% s^2
1	0.08	12	-1.74	1.85	38.36	0.50
2	0.17	6	0.17	5.77	200.17	2.60
3	0.25	4	0.00	0.33	0.67	0.01
4	0.33	3	-19.50	20.21	4731.50	61.51
5	0.42	2.4	1.74	-4.51	139.97	1.82
6	0.50	2	-14.67	0.00	2581.33	33.56
				$\Sigma_i I(f_i)$	7692.00	

Perioden von jeweils 2, 2.4 oder 3 Positionen. Periode 4 und 12 fallen daher aus, was man auch in Tabelle 9.2 an niedrigen $I(f_i)$ sehen kann. Wie man sich leicht überzeugt, kann man eine relativ gute Anpassung mit den Koeffizienten A_i und B_i, i = 2, 4, 5, 6 erreichen.

In Tabelle 9.3 ist die Anpassung mit der Kurve

$$y_x^* = 39 + 0.17\cos(2\pi x2/12) + 5.77\sin(2\pi x2/12) - 19.50\cos(2\pi x4/12) + 20.21\sin(2\pi x4/12) + 1.73\cos(2\pi x5/12) - 4.51\sin(2\pi x5/12) - 14.67\cos(2\pi x6/12)$$

dargestellt. In Abbildung 9.4 sieht man die graphische Darstellung.

Tabelle 9.3

Anpassung der Fourier–Reihe an die Daten von Tabelle 9.1

x	1	2	3	4	5	6	7
y_x	82	27	31	52	43	5	90
y_x^*	82.25	26.27	29.49	49.54	40.25	3.27	89.75

x	8	9	10	11	12
y_x	16	37	41	39	5
y_x^*	16.73	38.51	43.46	41.45	6.73

Diese Anpassung ergibt eine Residualsumme der quadratischen Abweichungen von 39.03, was zwar "optisch" gut aussieht, aber zu viele Koeffizienten kostet. Ein Polynom siebter Ordnung hätte wohl gleichgute Dienste geleistet.

Auf eine weitere Analyse zyklischer Wiederholungen werden wir hier verzichten. Zahlreiche Zeitreihen scheinen nur das stochastische Element zu enthalten, Modelle wurden bisher nicht entwickelt, linguistische Annahmen sind unbekannt. Man muß wohl abwarten, bis Einzeluntersuchungen neue Impulse einbringen.

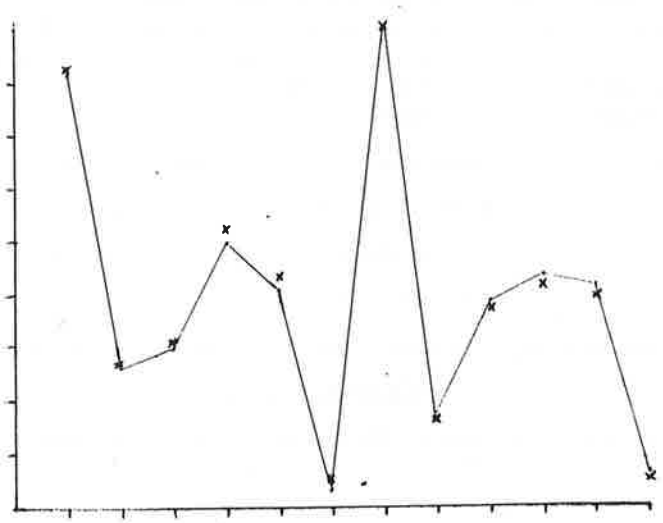

Abb. 9.4. Betonungsverlauf in Bottos "Smrt' Jánošíkova"

SCHLUSSWORT

A start in mathematization or mathematical modelling, however unrealistic, is better than either a prolix but unenlightening description or a grandiose verbal sketch.

Bunge 1967: 469

Unser Bemühen war es, die notwendigste Mathematik so darzustellen, daß es auch einem "qualitativen" Textlinguisten möglich sein sollte, seine Zählungen in die Formeln einzusetzen, um die erwünschten Resultate zu bekommen. Sollte man trotzdem der Überzeugung sein, daß man Wiederholungsstrukuren ohne Mathematik untersuchen kann, so beraubt man sich der Möglichkeit, eine Theorie aufzubauen bzw. eine gegebene zu testen.

Obwohl man bei den meisten Wiederholungsarten bisher sowohl wenige Daten als auch wenige Hypothesen hat, soll man möglichst frühzeitig mit der Mathematisierung anfangen, denn dies bringt derartige Vorteile (vgl. Bunge 1967: 474–476), auf die eine reife Wissenschaft nicht verzichten kann. Ja, gerade die Mathematisierung ist ein Zeichen der Reifung einer Disziplin. Die recht elementare Mathematik, die hier benutzt wurde, reicht jedoch keineswegs aus, um eine breit angelegte Texttheorie aufzubauen. Es wurde mit ihr eher eine Tür geöffnet, eine Richtung angedeutet. Ginge man aber in dieser Richtung weiter, so bestünde die Hoffnung, daß man nicht jahrzehntelang auf der Stufe der Begriffsbildung bleibt – wie es der Fall in der qualitativen Textanalyse oder in der sog. Standardlinguistik ist – sondern allmählich in die tieferen Bereiche der Textbildung eindringt und diese Disziplin an allgemeinere Disziplinen, wie z.B. Synergetik oder Systemtheorie, anschließt.

Wir hoffen, daß die Menge der Methoden und der Probleme, die hier präsentiert wurden, einen ersten informativen Überblick über die Breite dieses Gebiets darstellt und einen Ansporn sowohl zur Weiterentwicklung der Modelle, als auch eine Anregung zu weiteren Zählungen und Messungen an verschiedenen Texten in verschiedenen Sprachen bringt.

Die induktive Weiterentwicklung dieser Forschung kann man folgendermaßen vorantreiben:

(1) Für die hier verwendeten Einheiten untersucht man weitere Texte und testet die Resultate, d.h., man überprüft die Gültigkeit des gegebenen Modells für die gegebene Einheit.

(2) Man wendet die dargestellten Methoden auf weitere textologische Einheiten an und testet die Resultate, d.h., man prüft, bei welchen Einheiten welche Wiederholungsmechanismen wirken.

(3) Versagt ein Modell, dann behält man es vorläufig, erweitert es aber um einen neuen Parameter (mit oder ohne Interpretation) und testet das erweiterte Modell.

Deduktiv kann man die Weiterentwicklung folgendermaßen vorantreiben:

(1) Man konstruiert theoretisch weitere Wiederholungsmuster und prüft ihre Existenz in Texten, denn die hier aufgeführten sind sicherlich nicht die einzig möglichen.

(2) Man sucht nach weiteren Faktoren, die in die Modelle eingehen sollten, um ihnen eine sinnvollere, vollständigere Interpretation zu verleihen.

(3) Man sucht nach Zusammenhängen zwischen den Wiederholungsarten – die hier noch gar nicht angesprochen wurden – und konstruiert allgemeinere Gesetze, unter die man mehrere Wiederholungsmechanismen subsumieren kann.

(4) Man leitet adäquatere Modelle ab, indem man eventuell von anderen Annahmen ausgeht.

(5) Irgendwann greift man zur Axiomatisierung. Dies ist natürlich noch Zukunftsmusik, da sich philologische Wissenschaften langsamer entwickeln als die Naturwissenschaften. Sollte man aber auf diese Ziele verzichten, so ist es fraglich, welche man sich eigentlich stellen kann.

Die Wiederholungsanalyse in Texten ist eine Disziplin, die schon von Anfang an mit Mathematik verbunden wurde und die man ohne Mathematik nicht ernsthaft betreiben kann. Sie stellt also eine philologische Disziplin dar, in der die müßige, sophistische und völlig irrelevante Diskussion über die "Vorrangigkeit" der Suche nach "qualitativen Strukturen" überhaupt nicht auftauchen kann. Dieser Umstand wird sicherlich dazu beitragen, daß ihre Entwicklung schneller vorangehen wird.

LITERATUR

AITKEN, A.J., BAILEY, R.W., HAMILTON-SMITH, N. (Ed.) (1973) The computer and literary studies. Edinburgh, Edinburgh University Press

ALTMANN, G. (1963) Phonic structure of Malay pantun. Archiv orientální 31, 274-286

ALTMANN, G. (1968) Some phonic features of Malay shaer. Asian and African Studies 4, 9-16

ALTMANN, G. (1973) Mathematische Linguistik. In: KOCH, W. (Hrsg.), Perspektiven der Linguistik I. Stuttgart, Kohlhammer 208-232

ALTMANN, G. (1978) Zur Anwendung der Quotiente in der Textanalyse. Glottometrika 1, 91-106

ALTMANN, G., BURDINSKI, V. (1982) Towards a law of word repetitions in text-blocks. Glottometrika 4, 146-167

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., LEHFELDT, W. (1980) Einführung in die quantitative Phonologie. Bochum, Brockmeyer

ALTMANN, G., SCHWIBBE, M., KAUMANNS, W., KÖHLER, R., WILDE, J. (1982) Das Menzerathsche Gesetz in infomationsverarbeitenden Systemen. Stuttgart, Olms 1988

ALTMANN, G., ŠTUKOVSKÝ, R. (1965) The climax in Malay pantun. Asian and African Studies 1, 13-20

ANDERSON, T.W. (1971) The statistical analysis of time series. New York, Wiley

ANTOSCH, F. (1969) The diagnosis of literary style with the verb-adjective ratio. In: DOLEŽEL, BAILEY 1969, 57-65

ARAPOV, M.V., EFIMOVA, E.N., ŠREJDER, Ju.A. (1975a) O smysle rangovych raspredelenij. Naučno-techničeskaja informacija Ser. 2, Nr. 1, 9--20

ARAPOV, M.V., EFIMOVA, E.N., ŠREJDER, Ju.A. (1975b) Rangovye raspredelenija v tekste i jazyke. Naučno-techničeskaja informacija Ser. 2, Nr. 2, 3-7

AUSTERLITZ, R. (1961) Parallelism. In: DAVIE et al. 439-443

BASHARIN, G.P. (1959) On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability and Its Applications 4, 333-336

BEKTAEV, K.B., LUK'JANENKOV, K.F. (1971) O zakonach raspredelenija edinic pis'mennoj reči. In: PIOTROWSKI,R.G. (Hrsg.), Statistika reči i avtomatičeskij analiz teksta. Leningrad, Nauka 47-112

BELONOGOV, G.G. (1962) O nekotorych statističeskich zakonomernostjach v russkoj pi'smennoj reči. Voprosy jazykoznanija 11, 100-101

BERRIEN, F.K. (1968) General and social systems. New Brunswick N.J., Rutgers

BERRY-ROGGHE,G.L.M., (1973) The computation of collocations and their relevance in lexical studies. In: AITKEN, BAILEY, HAMILTON-SMITH 103-112

BEST, K-H., KOHLHASE, J. (Hrsg.) (1983) Exakte Sprachwandelforschung. Göttingen, Hedorot

BOCK, H.H. (1974) Automatische Klassifikation. Göttingen, Vandenhoeck & Rupprecht

BODER, D.P. (1940) The adjective-verb quotient: a contribution to the psychology of language. Psychological Revue 3, 309-343

BOWMAN, K.O., HUTCHESON, K., ODUM, E.P., SHENTON, L.R. (1969) Comments on the distribution of indices of diversity. In: PATIL, PIELOU, WATERS 315-359

BOX, G.E.P., JENKINS, G.M., (1970) Time series analysis, forecasting and control. San Francisco, Holden-Day

BRADLEY, J.V. (1968) Distribution-free statistical tests. Englewood Cliffs, N.J., Prentice Hall

BRAINERD, B. (1972a) Article use as an indirect indicator of style among English-language authors. In: JÄGER,S. (Hrsg.), Linguistik und Statistik. Braunschweig, Vieweg 11-32

BRAINERD, B. (1972b) On the relation between types and tokens in literary texts. Journal of Applied Probability 9, 507-518

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

BREIDT, R. (1973) Lassen sich Perseverationen durch Hirnschädigungen erklären? Psychiatrica clinica 6, 357-369

BROOKES, B.C. (1982) Quantitative analysis in the humanities: The advantage of ranking techniques. In: GUITER, ARAPOV 65-115

BUNGE, M. (1961) The weight of simplicity in construction and assaying of scientific theories. Philosophy of Science 28, 120-149

BUNGE, M. (1967) Scientific research I. Berlin, Springer

BUSEMANN, A. (1925) Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. Jena, Fischer

CARROLL, J.B. (1960) Vectors of prose style. In: SEBEOK,T.A. (Ed.) Style in language. Cambridge, Mass., The M.I.T. Press 1960, 283-292

CARROLL, J.B. (1968) Word-frequency studies and the lognormal distribution. In: ZALE,E.M. (Ed.), Proceedings of the Conference on Language and Language Behavior. New York

ČEBANOV, S.G. (1974) O podčinenii rečevych ukladov "indoevropejskoj" gruppy zakonu Puassona. Doklady Akademii Nauk SSSR. Novaja serija 55/2

COCHRAN, W.G. (1950) The comparison of percentages in matched samples. Biometrika 37, 256-266

CRAMER, P. (1968) Word association. New York, Academic Press

DANEŠ, F., VIEHWEGER,D. (Hrsg.) (1977) Probleme der Textgrammatik II. Berlin, Akademie Verlag

DANNHAUER, H-M., WICKMANN, D. (1972) Quantitative Bestimmung semantischer Umgebungsfelder in einer Menge von Einzeltexten. Literaturwissenschaft und Linguistik 2, 29-43

DAVID, F.N. (1950) Two combinatorial tests of whether a sample has come from a given population. Biometrika 37, 97-110

DAVID, J., MARTIN, R. (Hrsg.) (1977) Etudes de statistique linguistique. Paris, Klincksieck

DAVIE, D. et al. (Hrsg.) (1961) Poetics, poetyka, poetika. Warszawa, Panstwowe Wydawnicztwo Naukowe

DIJK, T.A.v. (1980) Textwissenschaft: eine interdisziplinäre Einführung. Tübingen, Niemeyer

DJADJULI (1961) Transkripsi Sjair Tjinta Berahi. Bahasa dan Budaja 9, 91-133

DOLEŽEL, L., BAILEY, R.W. (Eds.) (1969) Statistics and style. New York, Elsevier

DOLPHIN, C. (1977) Evaluation probabiliste des cooccurrences. In: DAVID, MARTIN 1977, 21-34

DRESSLER, W.U., BEAUGRANDE, R.A.de (1981) Introduction to textlinguistics. London, Longman

DROBISCH, W.M. (1866) Ein statistischer Versuch über die Formen des lateinischen Hexameters. Berichte über die Verhandlungen der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Philologisch-historische Klasse 18, 75-139

ESTOUP, J.B. (1916) Gammes sténographiques. Paris, Institut Sténographique

FAGEN, R.M., GOLDMAN, R.N. (1977) Behavioral catalogue analysis methods. Animal Behavior 25, 261-274

FISCHER, H. (1969) Entwicklung und Beurteilung des Stils. In: KREUZER, GUNZENHÄUSER 1969, 171-183

FRANCIS, I.S. (1966) An exposition of a statistical approach to Federalist dispute. In: LEED 1966, 38-78

FRUMKINA, R.M. (1962) O zakonach raspredelenija slov i klassov slov. In: MOLOŠNAJA, T.N. (Hrsg.), Strukturno tipologičeskie issledovanija. Moskva, ANSSSR 1962, 124-133

FUCKS, W. (1968) Nach allen Regeln der Kunst. Stuttgart, Deutsche Verlagsanstalt

FUCKS, W. (1970) Über den Gesetzesbegriff einer exakten Literaturwissenschaft, erläutert an Sätzen und Satzfolgen. Zeitschrift für Literaturwissenschaft und Linguistik 1, 113-137

FUCKS, W. (1971) Possibilities of exact style analysis. In: STRELKA, J. (Ed.), Patterns of literary style. University Park, Pennsylvania State University Press 51-75

FUCKS, W. (1955) Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Köln/Opladen, Westdeutscher Verlag

GANI, J. (1975) Stochastic models for type counts in a literary text. In: GANI,J. (Ed.), Perspectives in Probability and Statistics. London, Academic Press 313-323

GEOFFROY, A., LAFON, P., SEIDEL, G., TOURNIER, M. (1973) Lexicometric analysis of co-occurrences. In: AITKEN, BAILEY, HAMILTON-SMITH 1973, 113-133

GIBBONS, J.D. (1971) Nonparametric statistical inference. New York, McGraw-Hill

GONDA, J. (1959) Stylistic repetition in the Veda. Amsterdam, N.V. Noord Hollandsche Uitgevers Maatschappij

GOTTMAN, J.M., PARKHURST, J.T. (1980) A developmental theory of friendship and acquaintanceship processes. In: COLLINS,W.A. (Ed.), Development of cognition, affect, and social relations. Hillsdale, New Jersey, Erlbaum 197-253

GROOT, A.W. de (1946) Algemene Versleer. Den Haag

GROTJAHN, R. (1979) Linguistische und statistische Methoden in Metrik und Textwissenschaft. Bochum, Brockmeyer

GROTJAHN, R. (1980) The theory of runs as an instrument for research in quantitative linguistics. Glottometrika 2, 11-43

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

GUITER, H., ARAPOV, M.V. (Eds.) (1982) Studies in Zipf's law. Bochum, Brockmeyer

GUNZENHÄUSER, R. (1969) Zur literaturästhetischen Theorie G.D.Birkhoffs. In: KREUZER, GUNZENHÄUSER 1969, 295-311

HAIGHT, F.A., JONES, R.B. (1974) A probabilistic treatment of qualitative data with special reference to word association tests. Journal of Mathematical Psychology 11, 237-244

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

HALLIDAY, M.A.K., HASAN, R. (1976) Cohesion in English. London, Longman

HARWEG, R. (1974) Textlinguistik. In: KOCH,W.A. (Hrsg.), Perspektiven der Linguistik II. Stuttgart, Kröner 88-116

HERDAN, G. (1962) The calculus of linguistic observations. The Hague, Mouton

HERDAN, G. (1964) Quantitative linguistics. London, Butterworth

HERDAN, G. (1966) The advanced theory of language as choice and chance. Berlin, Springer

HERFINDAHL, O. (1950) Concentration in the steel industry. Diss., New York, Columbia University

HOOKE, R., JEEVES, T.A. (1961) Direct search solution of numerical and statistical problems. Journal of the Association for Computer Machines 8, 212-229

HŘEBÍČEK, L. (1985) Text as a unit and co-references. In: BALLMER, Th.T. (Ed.), Linguistic dynamics. Berlin, New York, de Gruyter 190-198

HŘEBÍČEK, L. (1986) Cohesion in Ottoman poetic texts. Archiv orientální 54, 252-256

HUTCHESON, K. (1970) A test for comparing diversities based on the Shannon formula. Journal of Theoretical Biology 29, 151-154

JAKOBSON, R. (1971) Unterbewußte sprachliche Gestaltung in der Dichtung. Zeitschrift für Literaturwissenschaft und Linguistik 1, 101-112

JOHNSON, N.L., KOTZ, S. (1969) Discrete distributions. Boston, Houghton Mifflin

KALININ, V.M. (1956) Funkcionaly, svjazannye s raspredeleniem Puassona, i statističeskaja struktura teksta. Trudy Matematičeskogo Instituta imeni V.A. Steklova 79, 182-197

KALININ, V.M. (1964) O statistike literaturnogo teksta. Voprosy jazykoznanija 13, Nr.1, 122-127

KATZ, L. (1965) Unified treatment of a broad class of discrete probability distributions. In: PATIL 1965, 175-182

KEMP, C.D., KEMP, A.W. (1956a) Generalized hypergeometric distributions. Journal of the Royal Statistical Society B 18, 202-211

KEMP, C.D., KEMP, A.W. (1956b) The analysis of point quadrat data. Australian Journal of Botany 4, 167-174

KENDALL, M.G., STUART, A. (1967) The advanced theory of statistics. London, Griffin

KOCH, W.A. (1969) Vom Morphem zum Textem. Hildesheim, Olms

KOCH, W.A. (1971) Taxologie des Englischen. München, Fink

KOCH, W.A. (1974) Tendenzen der Linguistik. In KOCH, W.A. (Hrsg), Perspektiven der Linguistik II. Stuttgart, Kröner 190-311

KOCHOL, V. (1968) Syntax a metrum. In: LEVÝ,J., PALAS,K. (Hrsg.), Theorie verše II. Brno, Universita J.E.Purkyně 167-178

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

KÖHLER, R., ALTMANN, G. (1983) Systemtheorie und Semiotik. Zeitschrift für Semiotik 5, 424-431

KRÁLÍK, J. (1977) An application of exponential distribution law in quantitative linguistics. Prague Studies in Mathematical Linguistics 5, 223-235

KREUZER, H., GUNZENHÄUSER, R.(Hrsg.) (1969:3) Mathematik und Dichtung. München, Nymphenburger

KU, H.H. (1963) A note on contingency tables involving zero frequencies and the 2I test. Technometrics 5, 398-400

KULLBACK, S., KUPPERMAN, M., KU, H.H. (1962) An application of information theory to the analysis of contingency tables, with a table of 2n ln n, n=1(1)10,000. Journal of Research of the National Bureau of Standards - B. Mathematics and Mathematical Physics 66B, 217-243

LÁNSKÝ, P., RADIL-WEISS, T. (1980) A generalization of the Yule-Simon model, with special reference to word association tests and neural cell assembly formation. Journal of Mathematical Psychology 21, 53-65

LEED, J. (Ed.) (1966) The computer and literary style. Kent, Ohio, Kent State UP

MAAS, H.-D. (1972) Über den Zusammenhang zwischen Wortschatzumfang und Länge des Textes. Zeitschrift für Literaturwissenschaft und Linguistik 8, 73-96

MANDELBROT, B. (1953) An information theory of the statistical structure of language. In: JACKSON,W. (Ed.), Communication Theory. New York, Academic Press 503-512

MANDELBROT, B. (1954a) Structure formelle des textes et communication. Word 10, 1-27

MANDELBROT, B. (1954b) Simple games of strategy occurring in communication through natural languages. IRE Transactions, PGIT-3, 124-137

MANDELBROT, B. (1954c) On recurrent noise limiting coding. In: Information Networks, the Brooklyn Polytechnic Institute Symposium 205-221

MANDELBROT, B. (1957) Linguistique statistique macroscopique. In: APOSTEL,L., MANDELBROT, B., MORF, A., Logique, langage et théorie de l'information. Paris, Presses Universitaires de France 1-78

MANDELBROT, B. (1961) On the theory of word frequencies and on related Markovian models of discourse. In: JAKOBSON,R. (Ed.), Structure of Language and its Mathematical Aspects. Providence, Rhode Island, American Mathematical Society 190-219

MANDELBROT, B. (1966) Information theory and psycholinguistics: A theory of word frequencies. In: LAZARSFELD,P.F., HENRY,N.W. (Eds.), Readings in mathematical social science. Chicago, Science Research Associates 350–368

MAŠKINA, L.E. (1968) O statističeskich metodach issledovanija leksiko-grammatičeskoj distribucii. Minsk, Diss.

MASON, D.I. (1961) Sound-repetition terms. In: DAVIE et al. 1961, 189–199

McINTOSH, R.P. (1967) An index of diversity and the relation of certain concepts to diversity. Ecology 48, 392–404

McNEIL, D.R. (1973) Estimating an author's vocabulary. Journal of the American Statistical Association 68, 92–96

MILLER, G.A. (1957) Some effects of intermittent silence. The American Journal of Psychology 70, 311–314

MILLER, G.A., CHOMSKY, N. (1963) Finitary models of language users. In: BUSH,R.R., GALANTER,E., LUCE,R.D. (Eds.), Handbook of Mathematical Psychology II. New York, Wiley 1963,419–491

MILLER, G.A., MADOW, W.G. (1963) On the maximum likelikood estimate of the Shannon-Wiener measure of information. In: LUCE,R.D., BUSH,R.R. GALANTER,E. (Eds.), Readings in Mathematical Psychology I. New York, Wiley 1963, 448–469

MITTENECKER, E. (1953) Perseveration und Persönlichkeit I,II. Zeitschrift für angewandte Psychologie 1, 5–31, 265–284

MOOD, A.M. (1940) The distribution theory of runs. Annals of Mathematical Statistics 11, 367–392

MORTON, A.Q., LEVISON, M. (1966) Some indicators of authorship in Greek prose. In: LEED 1966, 141–179

MOSTELLER, F., WALLACE, D.L. (1964) Inference and disputed authorship: The Federalist. Reading, Mass, Addison-Wesley

MÜLLER, W. (1971) Wortschatzumfang und Textlänge. Eine kleine Studie zu einem vielbehandelten Problem. Muttersprache 81, 266–276

MULLER, CH. (1965) Du nouveau sur les distributions lexicales: la formule de Waring-Herdan. Cahiers de Lexicologie 1, Nr.6, 35–53

MULLER, CH. (1968) Initiation à la statistique linguistique. Paris, Librairie Larousse

MULLER, CH. (1977) Observation, prévision et modèles statistiques. In: DAVID,J., MARTIN,R. (Hrsg.), Etudes de statistique linguistique. Paris, Klincksieck 9–19

NELDER, J.A., MEAD, R. (1964) A simplex method for function minimization. Computer Journal 7, 308–313 (8,1965,27)

NEŠITOJ, V.V. (1975) Dlina teksta i ob'em slovarja. Pokazateli leksičeskogo bogatstva teksta. In: Metody izučenija leksiki. Minsk, BGU 110–118

NEWMAN, L.I., POPPER, W. (1918) Studies in Biblical parallelism. Berkeley, UCP

NÖTH, W. (1974) Kybernetische Regelkreise in Linguistik und Textwissenschaft. Grundlagenstudien aus Kybernetik und Geisteswissenschaft 15, 75–86

NÖTH, W. (1975) Homeostasis and Equilibrium in Linguistics and Text Analysis. Semiotica 14, 222–244

NÖTH, W. (1977) Dynamik semiotischer Systeme. Stuttgart, Metzler 1977

NÖTH, W. (1978) Systems Analysis of Old English Literature. Journal for Descriptive Poetics and Theory of Literature (PTL) 3, 117–137

NÖTH, W. (1983) System theoretical principles of the evolution of the English language and literature. In: DAVENPORT,M., HANSEN, E., NIELSEN,F. (Eds.), Current topics in English historical linguistics. Odense, Univerity Press 103–122

OOMEN, U. (1971) Systemtheorie der Texte. Folia Linguistica 5,12–34

ORD, J.K. (1967) On a system of dicrete distributions. Biometrika 54, 649-656

ORD, J.K. (1972) Families of frequency distributions. London, Griffin

ORLOV, Ju.K. (1982) Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik) In: ORLOV, BORODA, NADAREJŠVILI 1982, 1-55

ORLOV, Ju.K., BORODA, M.G., NADAREJŠVILI, I.Š. (1982) Sprache, Text, Kunst. Quantitative Analysen. Bochum, Brockmeyer

OSGOOD, CH.E. (1959) The representational model and relevant research methods. In: I de SOLA POOL (Ed.), Trends in content analysis. Urbana, University of Illinois Press 33-88

PALEK, B., FISCHER, B. (1977) Ein Modell der Referenzstruktur des Textes. Studia Grammatica 18, 74-102

PALERMO, D.S., JENKINS, J.J. (1964) Word association norms. Minneapolis, University of Minnesota Press

PANDIT, S.M., WU, S.M., Time series and system analysis with applications. (1983) New York, Wiley

PANTOEN MELAJOE (1929) Weltevreden

PAŠKOVSKIJ, V.E., SREBRJANSKAJA, I.I. (1971) Statističeskie ocenki pis'mennoj reči bol'nych šizofreniej. In: Inženernaja lingvistika. Leningrad

PATIL, G.P. (Ed.) (1965) Classical and contagious discrete distributions. New York, Pergamon

PATIL, G.P. JOSHI, S.W. (1968) A dictionary and bibliography of discrete distributions. Edinburgh, Oliver & Boyd

PATIL, G.P., PIELOU, E.C., WATERS, W.E. (Eds.) (1971) Statistical ecology 3. University Park, The Pennsylvania State University Press

PIOTROVSKAJA, A.A., PIOTROVSKIJ, R.G. (1974) Matematičeskie modeli v diachronii i tekstoobrazovanii. In: Statistika reči i avtomaticeskij analiz teksta. Leningrad, Nauka 361-400

PIOTROWSKI, R.G. (1984) Text, Computer, Mensch. Bochum, Brockmeyer

PIOTROWSKI, R.G., BEKTAEV, K.B., PIOTROWSKAJA, A.A. (1985) Mathematische Linguistik. Bochum, Brockmeyer

PRŮCHA, J. (1967) On word-class distribution in Czech utterances. Prague Studies in Mathematical Linguistics 2, 65-76

RAPOPORT, A. (1982) Zipf's law re-visited. In: GUITER, ARAPOV 1982, 1-28

RATKOWSKY, D.A., HALSTEAD, M.H., HANTRAIS, L. (1980) Measuring vocabulary richness in literary works: A new proposal and a reassessment of some earlier measures. Glottometrika 2, 125-147

RIEGER, B. (1971) Wort- und Motivkreise als Konstituenten lyrischer Umgebungsfelder. Eine quantitative Analyse semantisch bestimmter Textelemente. Zeitschrift für Literaturwissenschaft und Linguistik 4, 23-41

RIEGER, B. (1974) Eine tolerante Lexikonstruktur. Zur Abbildung natürlich-sprachlicher Bedeutung auf unscharfe Mengen in Toleranzräumen. Zeitschrift für Literaturwissenschaft und Linguistik 16, 31-47

SACHS, L. (1972:3) Statistische Auswertungsmethoden. Berlin, Springer

SCHLISMANN, A. (1948) Sprach- und Stilanalyse mit einem vereinfachten Aktionsquotienten. Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik 2

SCHLITTGEN, R., STREITBERG, B.H.J. (1987:2) Zeitreihenanalyse. München, Oldenbourg

SCHMIDT, F. (1972) Numerische Textkritik: Goethes und Schillers Anteil an der Abfassung des Aufsatzes "Die Piccolomini". Zeitschrift für Literaturwissenschaft und Linguistik 5, 59–70

SCHWEIZER, H. (1979) Sprache und Systemtheorie. Tübingen, Narr

SEBEOK, T.A., ZEPS, V.J. (1959) On non-random distribution of initial phonemes in Cheremis verse. Lingua 8, 370–384

SEGAL, D.M. (1961) Nekotorye utočnenija verojatnostej modeli Cipfa. Mašinnyj perevod i prikladnaja lingvistika 5, 51–55

SICHEL, H.S. (1971) On a family of discrete distributions particularly suited to represent long-tailed frequency data. In: LAUBSCHER, N.F. (Ed.), Proceedings of the Third Symposium on Mathematical Statistics. Pretoria, S.A.C.S.I.R. 51–97

SICHEL, H.S. (1974) On a distribution representing sentence-length in written prose. Journal of the Royal Statistical Society A 137, 25–34

SICHEL, H.S. (1975) On a distribution law for word frequencies. Journal of the American Statistical Association 70, 542–547

SIEGEL, S. (1956) Nonparametric statistics for the behavioral sciences. New York, McGraw-Hill

SIMON, H.A. (1955) On a class of skew distribution functions. Biometrika 42, 425–440

SIMPSON, E.H. (1949) Measurement of diversity. Nature 163, 688

SKINNER, B.F. (1939) The alliteration in Shakespeare's sonnets: A study in literary behavior. Psychological Record 3, 186–192

SKINNER, B.F. (1941) A quantitative estimate of certain types of sound-patterning in poetry. The American Journal of Psychology 54, 64–79

SOMMERS, H.H. (1962) Analyse statistique du style. Louvain, Paris, Nauwelaerts

SPANG-HANSSEN, H. (1956) The study of gaps between repetitions. In: HALLE,M. (Ed.), For Roman Jakobson. The Hague, Mouton 1956, 497–502

STRAUSS, U. (1980) Struktur und Leistung der Vokalsysteme. Bochum, Brockmeyer

STRAUSS, U., SAPPOK, Ch., DILLER, H.J., ALTMANN, G. (1984) Zur Theorie der Klumpung von Textentitäten. Glottometrika 7, 73–100

ŠTUKOVSKÝ, R., ALTMANN, G. (1964) Fonická povaha slovenského rymu. Litteraria 7, 65–80

ŠTUKOVSKÝ, R., ALTMANN, G. (1965) Vyvoj otvoreného rymu v slovenskej poézii. Litteraria 8, 156–161

ŠTUKOVSKÝ, R., ALTMANN, G. (1966) Die Entwicklung des slowakischen Reimes im XIX. und XX. Jahrhundert. In: LEVÝ,J., PALAS,K. (Hrsg.), Teorie verše I. Brno 259–261

SWED, F.S., EISENHART, C. (1943) Tables of testing randomness of grouping in a sequence of alternatives. Annals of Mathematical Statistics 14, 66–87

TEŠITELOVÁ, M. (1967) On the role of nouns in lexical statistics. Prague Studies in Mathematical Linguistics 2, 121–131

TINTNER, G. (1965:2) Econometrics. New York, Wiley

TULDAVA, Ju. (1980) K voprosu ob analitičeskom vyraženii svjazi meždu ob'emom slovarja i ob'emom teksta. In: Lingvostatistika i kvantitativnye zakonomernosti teksta. Tartu 113–144

UHLÍŘOVÁ, L. (1967) Statistics of word order of direct object in Czech. Prague Studies in Mathematical Linguistics 2, 37–49

WILDGEN, W. (1985) Archetypensemantik. Tübingen, Narr

WILKINSON, R.J. (1907) Malay literature. Part I. Kuala Lumpur

WILKINSON, R.J., WINSTEDT, R.O. (1914) Pantun Melayu. Singapore

WORONCZAK, J. (1961) Statistische Methoden in der Verslehre. In: DAVIE,D. et al. 1961, 607-624

WORONCZAK, J. (1967) On an attempt to generalize Mandelbrot's distribution. In: To honor Roman Jakobson III. The Hague. 2254-2268

YNGVE, V. (1956) Gap analysis and syntax. IRE Transactions PGIT-2, 106-112

YULE, U.G. (1944) The statistical study of literary vocabulary. Cambridge, UP

ZIPF, G.K. (1935) The psycho-biology of language. Boston, Houghton Mifflin

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

ZÖRNIG, P. (1984a) The distribution of the distance between like elements in a sequence I. Glottometrika 6, 1-15

ZÖRNIG, P. (1984b) The distribution of the distance between like elements in a sequence II. Glottometrika 7, 1-14

ZÖRNIG, P. (1987) A theory of distances between like elements in a sequence. Glottometrika 8, 1-22

Sachverzeichnis

Namensverzeichnis

Quantitative Linguistics

Aim and Scope: Application of Mathematical Methods in Research on Linguistics, Literature and Related Areas.

Modes of Publication: Irregular intervalls, circa 4 volumes per year.

Available volumes:

Vol. 1: Altmann, G. (ed.), *Glottometrika* 1. 1978; VII+231 pp., DM 24,80

Vol. 2: Grotjahn, R., *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* 1979; V+296 pp., DM 34,80

Vol. 3: Grotjahn,R. (ed.), *Glottometrika* 2. 1980; V+218 pp., DM 24,80

Vol. 4: Strauss, U., *Struktur und Leistung der Vokalsysteme.* 1980; III+158 pp., DM 19,80

Vol. 5: Matthäus, W. (ed.), *Glottometrika* 3. 1980; 236 pp., DM 29,80

Vol. 6: Grotjahn, R., Hopkins, E. (eds.), *Empirical research on language teaching and language acquisition.* 1980; 231 pp., DM 29.80

Vol. 7: Altmann, G., Lehfeldt, W., *Einführung in die quantitative Phonologie.* 1980: X+401 pp., DM 29,80

Vol. 8: Altmann, G., *Statistik für Linguisten.* 1980; III+239 pp., DM 29.80

Vol. 9: Hopkins, E., Grotjahn, R. (eds.), *Studies in language teaching and language acquisition.* 1981; 220 pp., DM 29,80

Vol. 10: Skorochod'ko, E.F., *Semantische Relationen in der Lexik und in Texten.* 1981; 218 pp., DM 29,80

Vol. 11: Grotjahn, R. (ed.), *Hexamater studies.* 1981; VI+263 pp., 29,80

Vol. 12: Rieger, B. (ed.), *Empirical semantics I. A collection of new approaches in the field.* 1981; XIII+375 pp., DM 49.80

Vol. 13: Rieger, B. (ed.), *Empirical Semantics II. A collection of new approaches in the field.* 1981; XII+442 pp., DM 49.80

Vol. 14: Lehfeldt, W., Strauss, U. (eds.), *Glottometrika* 4. 1982; 200 pp., DM 29,80

Vol. 15: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen.* 1982; 330 pp., DM 49,80

Vol. 16: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf's law.* 1982; 262 pp., DM 39,80

Vol. 17: Arapov, M.V., Cherc, M.M., *Mathematische Methoden in der historischen Linguistik.* 1983; 171 pp., DM 24,80

Vol. 18: Brainerd, B. (ed.), *Historical linguistics.* 1983; II+236 pp., DM 29,80

Vol. 19: Winkler, P. (ed.), *Investigations on the speech process.* 1983; III+312 pp., DM 34,80

Vol. 20: Köhler, R., Boy, J. (eds.), *Glottometrika* 5. 1983; 228 pp., DM 29,80

Vol. 21: Goebl, H. (ed.), *Dialectology.* 1984; IV+335 pp., DM 49,80

Vol. 22: Alekseev, P.M., *Statistische Lexikographie.* 1984; 157 pp., DM 24,80

Vol. 23: Schwibbe, G., *Intelligenz und Sprache: Zur Vorhersagbarkeit des intellektuellen Niveaus kontentanalytischer Indikatoren.* 1984; 200 pp., DM 24,80

Vol. 24: Piotrowski, R.G., *Text - Computer - Mensch.* 1984; IX+422 pp., DM 49,80

Vol. 25: Boy, J., Köhler, R. (eds.), *Glottometrika* 6. 1984; 195 pp., DM 24,80

Vol. 26: Rothe, U. (ed.), *Glottometrika* 7. 1984; 176 pp., DM 29,80

Vol. 27: Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A., *Mathematische Linguistik.* 1985; 514 pp., DM 64,80

Vol. 28: Piotrowski, R.G., Popeskul, A.N., Chazinskaja, M.S., Rachubo, N.P., *Automatische Wortschatzanalyse.* 1985; 187 pp., DM 29,80

Vol. 29: Andersen, S., *Sprachliche Verständlichkeit und Wahrscheinlichkeit.* 1985; 194 pp., DM 29,80

Vol. 30: Embleton, Sh.M., *Statistics in historical linguistics.* 1986; VIII+194 pp., DM 29,80

Vol. 31: Köhler, R., *Zur sprachlichen Synergetik: Struktur und Dynamik der Lexik.* 1986; 201 pp., DM 29,80

Vol. 32: Fickermann, I. (ed.), *Glottometrika* 8. 1987; 211 pp., DM 29,80

Vol. 33: Wildgen, W., Mottron, L., *Dynamische Sprachtheorie. Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese.* 1987; III+423 pp., DM 59,80

Vol. 34: Grotjahn, R., Klein-Braley, C., Stevenson, D.K. (eds.), *Taking their measure: The validity and validation of language tests.* 1987; X+274pp., DM 39.80

Vol. 35: Schulz, K.P. (ed.), *Glottometrika* 9. 1988; 271 pp.

In preparation:

Piotrowski, R., Lesochin, M., Luk'janenkov, K., *introduction of elements of mathematics to linguistics.*

Tuldava, Ju., *Probleme und Methoden der quantitativen Analyse der Lexik.*

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

Mizutani, Sh. (ed.), *Japanese contributions to quantitative linguistics.*

Altmann, G., *Wiederholungen in Texten.*

Altmann, G., Hammerl, R., *Diskrete Wahrscheinlichkeitsverteilungen* I,II.

B B S

Bochumer Beiträge zur Semiotik

Ziele: Interdisziplinäre Beiträge zu praktischen und theoretischen Themen der Semiotik.
Erscheinungsweise: Unregelmäßige Abstände, ca. 5 - 10 Bände pro Jahr: Monographien, Aufsatzsammlungen zu festgesetzten Themen, Kolloquiumsakten usw.
Herausgeber: Walter A. Koch (Bochum)
Herausgeberbeirat: Karl Eimermacher (Bochum), Achim Eschbach (Essen), Udo L. Figge (Bochum), Roland Harweg (Bochum), Elmar Holenstein (Bochum), Werner Hüllen (Essen), Frithjof Rodi (Bochum).

Bände: lieferbar (*) und in Vorbereitung (bis 1987):

***Bd. 1:** HOLENSTEIN, Elmar, *Sprachliche Universalien*. xix + 250 S., paperback (pb) DM 44.80, ISBN 3-88339-419-X
***Bd. 2:** ZHOU, Hengxiang, *Determination und Determinantien: Eine Untersuchung am Beispiel neuhochdeutscher Nominalsyntagmen*. xii + 267 S., pb DM 44.80, ISBN 3-88339-412-2
***Bd. 3:** KOCH, Walter A., *Philosophie der Philologie und Semiotik*. Ca. 270 S., pb ca. DM 44.80, hardcover (hc) ca. DM 59.80, ISBN 3-88339-413-0
Bd. 4: KOCH, Walter A. (ed.), *For a Semiotics of Emotion*. Ca. 180 S., pb ca. DM 29.80, hc ca. DM 44.80, ISBN 3-88339-415-7
***Bd. 5:** ESCHBACH, Achim (ed.), *Perspektiven des Verstehens*. Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-414-9
Bd. 6: CANISIUS, Peter (ed.), *Perspektivität in Sprache und Text*. Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-416-5
Bd. 7: EISMANN, Wolfgang, GRZYBEK, Peter (eds.), *Semiotische Studien zum Rätsel*. Ca. 280 S., pb ca. DM 44.80, ISBN 3-88339-417-3
Bd. 8: KOCH, Walter A. (ed.), *Semiotik in den Einzelwissenschaften*. Ca. 1000 S., hc ca. DM 194.80, ISBN 3-88339-418-1
***Bd. 9:** SENNHOLZ, Klaus, *Grundzüge der Deixis*. xxvi + 314 S., pb DM 59.80, ISBN 3-88339-462-9
***Bd. 10:** KOCH, Walter A., *Evolutionäre Kultursemiotik*. xxii + 321 S., pb DM 59.80, ISBN 3-88339-463-7
***Bd. 11:** CANISIUS, Peter, *Monolog und Dialog*. Ca. 380 S., pb ca. DM 64.80, ISBN 3-88339-464-5
Bd. 12: JOB, Ulrike, *Regulative Verben im Französischen: Ein Beitrag zur semantischen Rekonstruktion des internen Lexikons*. Ca. 230 S., pb ca. DM 39.80, ISBN 3-88339-487-4
***Bd. 13:** SCHMIDT, Ulrich, *Impersonalia, Diathesen und die deutsche Satzgliedstellung*. Ca. 370 S., pb ca. DM 64.80, ISBN 3-88339-494-7
Bd. 14: KOCH, Walter A., POSNER, Roland (eds.), *Semiotik und Wissenschaftstheorie*. Ca. 350 S., pb ca. DM 59.80, ISBN 3-88339-554-4
Bd. 15: FIGGE, Udo L. (ed.), *Semiotik: Interdisziplinäre und historische Aspekte*. Ca. 250 S., pb ca. DM 44.80, ISBN 3-88339-555-2

Neuere und detailliertere Informationen zur Reihe (z.B. aktuelle Preisliste) sowie Bestellungen (Reihe oder Einzelbände) beim Verlag:
Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630-Bochum-Querenburg. Tel. (0234) 701360 oder 701383.

B P X

Bochum Publications in Evolutionary Cultural Semiotics

Aim and Scope: Transdisciplinary contributions to the analysis of sign processes and accompanying events from the perspective of the evolution of culture.
Modes of Publication: Irregular intervals, circa 5 to 10 volumes per year. Monographs, collections of papers on topical issues, proceedings of colloquies etc.
General Editor: Walter A. Koch (Bochum).
Advisory Editors: Karl Eimermacher (Bochum), Achim Eschbach (Essen).
Advisory Board: Yoshihiko Ikegami (Tokyo), Vjačeslav Vs. Ivanov (Moscow), Rolf Kloepfer (Mannheim), Roland Posner (Berlin), Thomas A. Sebeok (Bloomington), Vladimir N. Toporov (Moscow), Jan Wind (Amsterdam), Irene P. Winner (Cambridge, Mass.), Thomas G. Winner (Cambridge, Mass.).

Volumes: Available (*) and in preparation (up to 1987):

***Vol. 1:** YAMADA-BOCHYNEK, Yoriko, *Haiku East and West: A Semiogenetic Approach*. xiv + 591 pp., pb DM 94.80, ISBN 3-88339-404-1
Vol. 2: ESCHBACH, Achim, KOCH, Walter A. (eds.), *A Plea for Cultural Semiotics*. Ca. 320 pp., pb ca. DM 59.80, ISBN 3-88339-405-X
Vol. 3: KOCH, Walter A., *Cultures: Universals and Specifics*. Ca. 170 pp., pb ca. DM 34.80, ISBN 3-88339-407-6
Vol. 4: KOCH, Walter A. (ed.), *Simple Forms: An Encyclopaedia of Simple Text-Types in Lore and Literature*. Ca. 700 pp., pb (paperback) ca. DM 129.80, hc (hardcover) ca. DM 144.80, ISBN 3-88339-406-8
Vol. 5: WINNER, Irene P., *Cultural Semiotics: A State of the Art*. Ca. 130 pp., pb ca. DM 24.80, ISBN 3-88339-408-4
***Vol. 6:** KOCH, Walter A., *Evolutionary Cultural Semiotics*. Ca. 370 pp., pb ca. DM 69.80, hc ca. DM 84.80, ISBN 3-88339-409-2
Vol. 7: KOCH, Walter A. (ed.), *Culture and Semiotics*. Ca. 220 pp., pb ca. DM 44.80, hc ca. DM 59.80, ISBN 3-88339-421-1
Vol. 8: EIMERMACHER, Karl, GRZYBEK, Peter (eds.), *Cultural Semiotics in the Soviet Union*. Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-410-6
Vol. 9: VOGEL, Susan, *Children's Humour: A Semiogenetic Approach*. Ca. 270 pp., pb ca. DM 49.80, ISBN 3-88339-411-4
Vol. 10: KOCH, Walter A. (ed.), *Semiotics in the Individual Sciences*. Ca. 1000 pp., hc ca. DM 199.80, ISBN 3-88339-484-X
Vol. 11: KOCH, Walter A. (ed.), *Geneses of Language. Acta Colloquii*. Ca.400 pp., pb ca. DM 69.80, ISBN 3-88339-485-8
Vol. 12: KOCH, Walter A. (ed.), *The Nature of Culture. Proceedings of the International and Interdisciplinary Symposium, October 7-11, 1986, Ruhr-University Bochum*. 2 vols., each ca. 500 pp., pb each ca. DM 84.80, ISBN 3-88339-553-6
***Vol. 13:** KOCH, Walter A., *Genes vs. Memes*. Ca. 100 pp., pb ca. DM 24.80, ISBN 3-88339-551-X

For more recent and more detailed information on the series (e.g. the current price-list) and for orders for the whole series or individual volumes please contact the publisher: **Studienverlag Dr. Norbert Brockmeyer, Querenburger Höhe 281, D-4630-Bochum, Fed. Rep. Germany.** Tel. (0234) 701360 or 701383.