Studies
in Quantitative Linguistics
2

**Vivien Altmann**
**Gabriel Altmann**

# Anleitung zu quantitativen Textanalysen

Methoden und Anwendungen

**RAM - Verlag**

# Anleitung zu quantitativen Textanalysen

**Methoden und Anwendungen**

**von**

**Vivien Altmann**

**und**

**Gabriel Altmann**

**2008**
**RAM-Verlag**

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII +134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.

ISBN: 978-3-9802659-5-9



RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

# Vorwort

Die Tatsache, dass es im Deutschen kein Kompendium der quantitativen Einzeltextanalyse gibt, führte uns dazu, einen Versuch zu starten und zumindest einige Aspekte des Textes von einem bestimmten Sichtwinkel zu betrachten und als ein „geschlossenes“ Anleitungsbuch dem Leser zu übergeben. Da es nicht möglich ist, alle Aspekte eines Textes zu erfassen, haben wir uns auf nur die besser ausgearbeiteten Bereiche beschränkt, nämlich auf Rhythmus, Phonik, Wortdynamik und Denotationsanalyse und haben zahlreiche Methoden eingeführt, die alternativ benutzt werden können.

Das Buch, das wir eher als Anleitung verstehen, kann in einem Semester mit oder ohne Lehrer bewältigt werden, es reichen elementare Kenntnisse der Statistik und die entsprechende Software. Wir haben uns bemüht, den Leser zu stimulieren, indem wir viele Aspekte der Forschung zu zeigen versuchten. Eine konsequente Weiterführung und Erweiterung einiger Ideen, sowohl historisch als auch synchron und zwischensprachlich könnte zu tieferen Erkenntnissen der Textdynamik führen. Wir haben mit Absicht ein Gedicht als Forschungsobjekt gewählt, weil man daran sowohl die Phonik als auch den Rhythmus mit etwas mehr „Berechtigung“ untersuchen kann. Sie sind die Grundlage der meisten Poesie. In der Prosa spielen sie bisher nicht die gleiche Rolle. Wichtig war auch die Tatsache, dass der Gedichttext kurz war, damit entsprechende Methoden zum Zuge kommen. Man kann alle mutatis mutandis auch an Prosa anwenden.

Für den Unterrichtenden der Poetik soll das Buch eine Anleitung für Vorlesungen und Übungen gewähren. Die große Menge von Formeln und Rechnungen soll nicht zu der Ansicht verleiten, dass quantitative und qualitative Poetik unterschiedliche Ziele verfolgen, denn in beiden geht es darum, Regularitäten und Tendenzen zu erfassen – jeweils mit entsprechenden Mitteln.

Wir haben uns bemüht, dem Leser eine Zahl von alternativen Methoden anzubieten, die er nicht alle anwenden muss, wenn er einen Text analysiert. Eine Auswahl, im meisten Fällen eine einzige Methode reicht, um ein Phänomen nachzuweisen und zu erfassen. Alle Rechnungen wurden auf eine möglichst nicht-technische Weise Schritt für Schritt durchgeführt, damit zumindest die Berechnung einer Formel möglich ist, wenn dem Leser die statistischen Hintergründe nicht hinreichend einleuchtend sind. Verweise auf Literatur helfen, das Studium der betreffenden Methode oder den Sinn eines Verfahrenstiefer kennen zu lernen. Es wäre nicht sinnvoll gewesen, die Ableitungen bestimmter Verfahren breit darzustellen, da dies das Lesen des Textes erschwert hätte. Es war eher das Ziel, zu zeigen, was mit einfachen Methoden feststellbar und machbar ist. Die Kenntnis einiger einfachen Verfahren wird heutzutage in empirischen Wissenschaften generell erwartet und vorausgesetzt. Ist sie nicht vorhanden, dann kann sich der Leser das entsprechende Buch aus der Unmenge der Lehrbücher der Statistik oder der Graphentheorie auszuwählen.

Für die Textkorrektur und zahlreiche Ratschläge bedanken wir uns herzlichst bei Herrn Werner Lehfeldt, der mit göttlicher Geduld alle sprachlichen Fehler ausgemerzt hat. Der übriggebliebene Rest sind unsere eigenen üblichen Sünden, für die wir zwar bestraft werden sollten, aber erst später… Herrn Karl-Heinz Best danken wir für die Überwachung des phonischen Teils des Buches, der sich an seine Transkription hält.

V. und G. Altmann

J.W.v. Goethe, Erlkönig

Wer reitet so spät durch Nacht und Wind?
Es ist der Vater mit seinem Kind;
Er hat den Knaben wohl in dem Arm.
Er fasst ihn sicher, er hält ihn warm.

Mein Sohn, was birgst du so bang dein Gesicht?
Siehst, Vater, du den Erlkönig nicht?
Den Erlkönig mit Kron und Schweif?
Mein Sohn, es ist ein Nebelstreif.

Du, liebes Kind, komm, geh mit mir!
Gar schöne Spiele spiel ich mit dir;
Manch bunte Blumen sind an dem Strand,
Meine Mutter hat manch gülden Gewand.

Mein Vater, mein Vater, und hörest du nicht,
Was Erlenkönig mir leise verspricht?
Sei ruhig, bleibe ruhig, mein Kind:
In dürren Blättern säuselt der Wind

Willst, feiner Knabe, du mit mir gehn?
Meine Töchter sollen dich warten schön;
Meine Töchter führen den nächtlichen Reihn
Und wiegen und tanzen und singen dich ein.

Mein Vater, mein Vater, und siehst du nicht dort
Erlkönigs Töchter am düstern Ort?
Mein Sohn, mein Sohn, ich sehe es genau:
Es scheinen die alten Weiden so grau.

Ich liebe dich, mich reizt deine schöne Gestalt;
Und bist du nicht willig, so brauch ich Gewalt.
Mein Vater, mein Vater, jetzt faßt er mich an!
Erlkönig hat mir ein Leids getan!

Dem Vater grausets, er reitet geschwind,
Er hält in Armen das ächzende Kind,
Erreicht den Hof mit Mühe und Not:
In seinen Armen das Kind war tot.

# Inhalt

# 1. Einführung

Das Ziel dieses Buches besteht darin, die Möglichkeiten quantitativer Sprachanalyse anhand eines kurzen Textes zu zeigen. Üblicherweise untersucht man lange Texte, in denen sich Tendenzen klarer ausprägen können, die Maßzahlen und verschiedenen Funktionen haben größere Stabilität und kleinere Variabilität. Am besten geeignet für die Analyse sind jedoch Texte „mittlerer" Größe – was immer das auch zu bedeuten mag –, weil sich in sehr langen Texten viele klassischen Tests als ineffizient erweisen und in kurzen Texten die Macht des Tests zu klein wird. In einer derartigen Situation muss man bei kurzen Texten oft zu verteilungsfreien Methoden greifen, die es auch hier ermöglichen, zuverlässige Schlüsse zu ziehen und Erscheinungen zu entdecken, die man auf anderem Wege nicht feststellen kann.

Es ist heutzutage gang und gäbe, Textanalysen an Korpora durchzuführen, da solche in großen Mengen zur Verfügung stehen. Nutzt man jedoch ein Korpus als ganzes, so ist seine Verwendung für bestimmte Fragestellungen aus mehreren Gründen problematisch:

(i) Es stellt eine Mischung von heterogenen Texten dar, so dass eine Charakterisierung von Einzeltexten ausgeschlossen ist, es sei denn, die Texte im Korpus sind getrennt und werden vollständig erfasst.

(ii) Induktive Verfahren führen in Korpora zur Charakterisierung übergeordneter Entitäten, zu induktiven Verallgemeinerungen, z.B. zu einem Genre, wobei man oft Schlüsse über Grundgesamtheiten zieht, die es gar nicht gibt.

(iii) Gesetze gelten für homogene Texte, d.h. für Einzeltexte, bei denen die Randbedingungen besser identifizierbar sind, während sie in einem Korpus stark verwischt werden, so dass in ihm das Testen von Gesetzen im Korpus Beschränkungen unterworfen, nicht aber völlig ausgeschlossen ist.

All dies vermindert jedoch keineswegs den Wert von entsprechend gestalteten Korpora oder deren Anwendung für andere Zwecke. Für bestimmte Fragestellungen sind sie sogar unabdingbar. Auf höheren Ebenen der Sprache, z.B. in der Syntax, kann man sie erfolgreich auch für die Suche nach allgemeinen Gesetzen ausnutzen.

In Einzeltexten kann man wortwörtlich eine unendliche Menge von Eigenschaften finden. Ihre „Entdeckung" hängt von dem von uns benutzten Begriffsapparat ab, ob man nun die reale Existenz solcher Eigenschaften akzeptiert oder nicht. Die Erkenntnisgewinnung verläuft beim Menschen nur über Begriffe, von denen nur äußerst wenige in Wörtern kodifiziert sind. So kann man in Texten vermutete Eigenschaften erst dann untersuchen, wenn wir sie begrifflich erfasst haben. Mit anderen Worten, die Zahl der begrifflich erfassten Spracheigenschaften hängt von unserem Wissensstand ab. Viele Texteigenschaften lassen sich am exaktesten quantitativ erfassen, was auch ihre weitere Verarbeitung wie Darstellung, Testen, Interpretation sehr erleichtert und exakter macht. Andere lassen sich nur quantitativ erfassen, z.B. Längen, funktionale Abhängigkeiten,

Übergangsabhängigkeiten u.ä., so dass die Anwendung quantitativer Methoden auf einer bestimmten Stufe der Forschung unumgänglich wird.

Betrachtet man die quantitative Analyse nicht nur als Charakterisierungsmöglichkeit und induktives Testen von isolierten Hypothesen, sondern auch als eine Möglichkeit für die Überprüfung deduktiver Gesetzeshypothesen, so erweitert sich ihr Aufgabengebiet beträchtlich, und ihr gnoseologischer Status gewinnt an Wert.

In den folgenden Kapiteln werden wir nur einen relativ kurzen, sehr bekannten Text analysieren, nämlich Goethes Gedicht „Der Erlkönig“. Dieses aus 32 Versen in acht Strophen bestehende und im Knittelvers mit dem Reimschema *aabb* geschriebene Gedicht soll uns als Beispiel dienen. Vergleiche mit anderen Texten werden nicht angestrebt, wir überlassen dieses Problem anderen Forschern. Würde man das Problem jedoch von dieser Seite aufgreifen, so könnte man die geschichtliche Entwicklung deutscher Texte, speziell der Poesie, um viele Aspekte bereichern. Man betrachte dieses Buch eher als ein Lehrbuch, in dem Probleme und Lösungswege gezeigt werden, jedoch die Ableitung der mathematischen Mittel meistens ausbleibt, da das Buch kein Lehrbuch der Mathematik sein will.

Bei der Beschreibung des Gedichtes verfahren wir so, dass wir bei der niedrigsten Ebene anfangen, nämlich bei rein formalen Entitäten wie Längen, Folgen von Längen oder dem Rhythmus, der, ebenso wie in der Musik, nur die Begleitung darstellt. Dann gehen wir zur phonischen Ebene über, die die bedeutungslose Lautung des Textes darstellt, dann analysieren wir die Wörter, die die Bedeutung tragen, und zum Schluss suchen wir nach den Eigenschaften der denotativen Struktur des Gedichtes. Morphologie und Syntax werden nur berührt, aber nicht tiefer untersucht. Jede dieser Ebenen besitzt Eigenschaften, die die anderen nicht besitzen. Es ist aber möglich, zumindest im Modellbereich nach Gemeinsamkeiten zwischen den Ebenen zu suchen. Es ist, natürlich, das Ziel jeglicher Forschung zu zeigen, dass das Ganze eines Gedichts in der Tat ein Ganzes ist, in dem alles mit allem zumindest indirekt zusammenhängt. Um dieses Ziel zu erreichen, muss man leider zu Anfang analytisch vorgehen und Ebene um Ebene nach Regularitäten durchsuchen und diese formal erfassen. Dies wird in den Kapiteln 2 bis 6 systematisch angestrebt.

Im zweiten Kapitel betrachten wir den Rhythmus des Textes. Hinter dem gleichsam an der Oberfläche liegenden Knittelvers oder hinter einem festen Rhythmus versteckt sich eine Menge anderer (latenter) Tendenzen, die man nur durch Tests ermitteln kann. Der Vers ist nur eine rhythmische Einheit, andere Einheiten kann man im Verlauf der Verse oder in Strophen als Ganzen oder sogar im Gedicht als Ganzem entdecken. Dieser Aspekt wurde bisher weniger untersucht, daher widmen wir ihm besondere Aufmerksamkeit.

Im dritten Kapitel befassen wir uns mit der phonischen Seite des Textes. Auch wenn ein Autor seinen Text mehrmals korrigiert und ändert, nehmen wir trotzdem an, dass sich die Phonik schließlich in einen Zustand einpendelt, der dem Gehörempfinden des Autors zufriedenstellend vorkommt. Der Autor handelt

zwar intuitiv, zählt und rechnet nicht, kümmert sich eventuell um den von ihm angestrebten Wohlklang, weiß aber nicht, dass er nach bestimmten Gesetzen handelt, die er nicht kennen kann. Er würde sich vermutlich wundern, wenn man ihm sagte, was man alles in seinem Text entdeckt hat.

Im vierten Kapitel widmen wir uns der Verteilung der Wörter und dem Fluss der Information im Text. Beide Probleme sind hinreichend bekannt, jedoch gibt es zahlreiche Aspekte, die wir nur streifen werden. Diese Forschungsrichtung ist von den hier vorgestellten am weitesten entwickelt. Sie wurde nicht nur von Linguisten, sondern auch von Mathematikern, Geographen, Soziologen, Physikern und anderen Spezialisten vorangetrieben, die in ihren Disziplinen analoge Phänomene entdeckten und formal erfassen wollten. sie ist gleichzeitig einer der strittigsten Bereiche der quantitativen Linguistik, weil man das zur Debatte stehende Problem unter so vielen Gesichtspunkten angehen kann.

Im fünften Kapitel beschäftigen wir uns mit der neuartigen denotativen Analyse, stellen den Text als einen Graphen dar und untersuchen seine Struktur mit Hilfe von Begriffen aus der Graphentheorie. Die Relationen im Text werden sowohl statistisch als auch deterministisch ermittelt und durch Kanten des Graphen dargestellt. Die Eigenschaften des Graphen werden als Texteigenschaften interpretiert. Diese Disziplin ist am wenigsten fortgeschritten, da sie noch sehr jung ist.

Im sechsten Kapitel wird angedeutet, wie man grammatische Phänomene für die Charakterisierung des Textes verwenden kann.

Das Buch eignet sich nur dann zum Selbststudium, wenn man nur nach Anwendungs- und Auswertungsmöglichkeiten sucht, man kann es in diesem Sinne als eine Art Kochbuch bezeichnen. Alle Methoden lassen sich bequem in einem Semester an einen Text anwenden. Es wird empfohlen, alle Probleme an anderen Beispielen durchzurechnen und die dabei erzielten Ergebnisse mit denen aus dem Erlkönig zu vergleichen. Es ist kein Lehrbuch der Statistik oder der Graphentheorie, sondern eher eine Anleitung zu selbständigen Analysen, die auch von Anfängern durchgeführt werden können.

Das Problem der Poetik besteht unter anderem auch darin, dass man bestimmte Ebenen und Entitäten für den jeweils gegebenen Aspekt als wichtig betrachtet, während andere einfach ignoriert werden. Im Rahmen einer Arbeitsteilung zwischen den Disziplinen ist dies auch berechtigt: Die Poetik ist nur eine der Disziplinen, die sich mit poetischen Texten beschäftigen. Die allgemeine Textologie strebt jedoch nach einem einheitlichen Blick auf Texte, wobei ihr die Poetik als Spezialdisziplin hilfreich sein kann. Der einfachste Weg führt über die quantitative Erfassung von Textphänomenen, die man dann leicht auf gemeinsame Nenner bringen kann. Die quantitative Betrachtung solcher Phänomene erweitert nicht nur den Untersuchungsrahmen, sondern gibt auch die Möglichkeit, eine Theorie anzustreben, d.h. ein System von Gesetzen, die die Textegenerierung steuern. Vieles ist bereits erreicht worden, vieles liegt noch vor uns.

# 2. Rhythmus

Ob Prosa oder Poesie, jedes hinreichend lange Textsstück hat seinen Rhythmus, der durch unterschiedliche Entitäten gestaltet werden kann, durch Betonung, Silben-, Vers- oder Satzlänge, grammatische Gliederung mit Intonation, Pausen usw. Der Rhythmus kann deterministisch sein, z.B. bei festen Metren, konstanter Silbenzahl im Vers u.ä., er kann aber auch „schwächer“ sein und Tendenzen aufweisen, die sich nur statistisch erfassen lassen. Man stellt über solche Tendenzen Hypothesen auf und versucht, diese zu testen. Die Aufgabe eines Textwissenschaftlers besteht nicht nur in der Bildung von Begriffen, unter die er Erscheinungen subsumiert, sondern vor allen Dingen in der Formulierung von Hypothesen, wobei seiner Phantasie keine Grenzen gesetzt sind. Zwar zwingen ihn die jeweils herrschende Doktrin und die nach ihr festgelegte „einzig mögliche“ Ordnung der Daten zur Zügelung seiner Phantasie, auf der anderen Seite wissen wir aber, dass kein Wissen gesichert ist und es für kein wissenschaftliches Problem eine endgültige Lösung gibt. Seine Phantasie sollte daher eher von der Testbarkeit – der wichtigsten Eigenschaft wissenschaftlicher Hypothesen – geleitet sein und nicht von vage formulierten, schulabhängigen oder ideologisch gefärbten, sich oft widersprechenden Ansichten. Dies ist eben der Bereich, in dem ihm der Statistiker eine hilfreiche Hand reichen kann. Zwar kann er ihm die heuristische Arbeit nicht abnehmen, er vermag es aber, viel für die Objektivität, Akzeptabilität und Systematisierung seiner Resultate zu tun, d.h., er kann ihm helfen, zumindest bis an die Schwelle einer Theorie zu gelangen.

Im folgenden werden wir die Verse des „Erlkönig“ als Folgen von betonten und unbetonten Silben auffassen und in diesen Folgen nach Tendenzen suchen. Der „Erlkönig“ dient uns dabei nur als Beispiel, unser Hauptinteresse gilt der Vermittlung einfacher nützlicher statistischer Methoden.

## 2.1. Rhythmische Muster

Rhythmisch kann man die Verse des „Erlkönig“ als Folgen von betonten und unbetonten Silben angeben. Lässt man die betonten Silben aus und vermerkt lediglich die Anzahl der unbetonten vor einer betonten, so ergibt sich z.B. für

„Wer reitet so spät durch Nacht und Wind?“

das Muster der unbetonten Silben als

1 2 1 1.

Kodiert man alle Verse auf diese Weise, dann bekommt man für das ganze Gedicht dieses Resultat:

(I)

| | | | | |
|---|---|---|---|---|
| 1. | 1211 | | 17. | 1121 |
| 2. | 1121 | | 18. | 2121 |
| 3. | 1112 | | 19. | 2122 |
| 4. | 1121 | | 20. | 1222 |
| 5. | 1122 | | 21. | 1222 |
| 6. | 1112 | | 22. | 1121 |
| 7. | 1121 | | 23. | 1112 |
| 8. | 1111 | | 24. | 1212 |
| 9. | 1111 | | 25. | 1322 |
| 10. | 1121 | | 26. | 1222 |
| 11. | 1121 | | 27. | 1222 |
| 12. | 2112 | | 28. | 1121 |
| 13. | 1222 | | 29. | 1122 |
| 14. | 1122 | | 30. | 1122 |
| 15. | 1112 | | 31. | 1112 |
| 16. | 1112 | | 32. | 1121 |

In Vers 25 gab es bei unseren Informanten unterschiedliche Lesarten. Zwecks Einheitlichkeit haben wir die obige gewählt. An dieser Stelle ist zu bemerken, dass ein Expertenurteil nur dann „besser“ ist als ein statistisch begründetes Urteil, wenn es auf einer Theorie basiert. Da es aber keine Theorie des Rhythmus oder der „poetischen Rede“ gibt, sollte man eher eine statistische Lösung poetologischer Probleme anstreben.

Diese rhythmischen Muster können auf verschiedene Weisen weiter zusammengefasst und kodiert werden. Hier wählen wir folgende Zusammenfassung nach der Silbenzahl (Zahl der unbetonten Silben):

(II)

| Kod | Kombination | Silbenzahl |
|---|---|---|
| a: | 1111 | 4 |
| b: | 1112 | 5 |
| | 1121 | 5 |
| | 1211 | 5 |
| c: | 1122 | 6 |
| | 1212 | 6 |
| | 2112 | 6 |
| | 2121 | 6 |
| d: | 1222 | 7 |
| | 2122 | 7 |
| e: | 1322 | 8 |

Benutzt man zur Kodierung diese Buchstaben, die Längenklassen nach Silbenzahl darstellen, dann ergibt sich das Gedicht als Folge von rhythmischen Mustertypen

(III) bbbb cbba abbc dcbb bcdd dbbc eddb ccbb

oder äquivalent numerisch nach der Silbenzahl dargestellt als:

(IV) 5555 6554 4556 7655 5677 7556 8775 6655.

Die Häufigkeit einzelner Typen (Längenklassen) ist

(V)

| | |
|---|---|
| a | 2 |
| b | 16 |
| c | 7 |
| d | 6 |
| e | 1. |

Die Aufgabe der induktiven quantitativen Analyse besteht nun darin, in den Grunddaten (I)-(V) nach Regularitäten zu suchen und diese darzustellen.

## 2.2. Ausnutzung

Die Ausnutzung kann auf zwei Weisen aufgefasst werden: als Anzahl einzelner Typen, die überhaupt konstruiert worden sind, oder als deren Häufigkeit im Gedicht.

### 2.2.1. Konstruktion

Wie man in Schema (II) sieht, werden die einzelnen Typen unterschiedlich ausgenutzt. Nur bei Typ *a* werden alle Möglichkeiten (d.h. 1) erschöpft, bei Typ *b* wurde 2111 nicht realisiert usw. Die konstruktionelle Ausnutzung berechnen wir so, dass wir die Zahl der realisierten Typen in Beziehung zu der Zahl der möglichen setzen. Die Zahl der theoretischen Möglichkeiten der Typenbildung berechnet sich mit Hilfe des Multinomialkoeffizienten als

$$(2.1) \quad T = \frac{n!}{k_1!k_2!...k_r!},$$

wobei

*n* die Zahl der Elemente in der Sequenz ist, d.h. hier ist im Vers immer $n = 4$;

$k_1, k_2,\ldots,k_r$ die Anzahl einzelner Ziffern in der Sequenz darstellt. So ist bei Typ *b* die Ziffer 1 dreimal vorhanden, d.h. $k_1 = 3$, die Ziffer 2 ist einmal vorhanden, d.h. $k_2 = 1$ usw.

*r* die Zahl unterschiedlicher Ziffern symbolisiert: in Typ *a* ist $r = 1$, weil wir hier nur Einser haben; in den Typen *b*, *c* und *d* ist $r = 2$, weil es hier sowohl Einser als auch Zweier gibt, und in Typ *e* ist $r = 3$.

Falls $r = 2$, dann vereinfacht sich der Multinomialkoeffizient (2.1) auf den Binomialkoeffizienten

$$(2.2)\quad T = \binom{n}{k_1} = \binom{n}{k_2} .$$

In unserem Fall erhalten wir die Zahlen der theoretischen Möglichkeiten folgendermaßen: In Typ *a* gibt es 4 Einser, d.h. $k_1 = 4$, woraus nach (2.2)

$$\text{Typ } a: \qquad T_a = \binom{4}{4} = 1$$

folgt. Die anderen Zahlen ergeben sich analog als

Typ *b*: Hier gibt es dreimal 1 und einmal 2, d.h. $T_b = \left(\frac{4!}{3!1!}\right) = \binom{4}{1} = 4$

Typ *c*: Hier gibt es zwei Zweier und zwei Einser, d.h.

$$T_c = \left(\frac{4!}{2!2!}\right) = \binom{4}{2} = 6$$

Typ *d*: Hier gibt es 3 Zweier und eine 1, d.h. $T_d = \left(\frac{4!}{3!1!}\right) = \binom{4}{3} = 4,$

während Typ *e* mit Hilfe des Multinomialkoeffizienten berechnet werden muss:

$$\text{Typ } e: \; T_e = \frac{4!}{1!2!1!} = 12 .$$

So erhalten wir für die Ausnutzung der Typen das Resultat in der letzten Spalte von Tabelle 2.1.

Tabelle 2.1
Ausnutzung der Typen

| Typ | Länge | Typenzahl | | Ausnutzung |
|---|---|---|---|---|
| | | beobachtet *b* | theoretisch *t* | *b/t* |
| a | 4 | 1 | 1 | 1/1 = 1.00 |
| b | 5 | 3 | 4 | 3/4 = 0.75 |
| c | 6 | 4 | 6 | 4/6 = 0.67 |
| d | 7 | 2 | 4 | 2/4 = 0.50 |
| e | 8 | 1 | 12 | 1/12 = 0.08 |

Wie man sieht, ist die Typenausnutzung um so kleiner, je länger der Vers ist. Es wäre interessant zu untersuchen, ob sich dieses Verhältnis auch in längeren Gedichten nachweisen lässt. Solange aber keine weiteren Daten vorliegen, wäre es etwas müßig, deduktiv nach einer „Ausnutzungskurve“ zu suchen, zumal *b/t* offensichtlich einen Wendepunkt hat und möglicherweise etwas komplizierter aussehen wird. Vorläufig lässt sich dieser Trend einfach mit einer Geraden erfassen, die die Form

$$\text{Ausnutzung} = b/t = A - BL$$

hat, wo $L$ die Länge bedeutet und $A$ und $B$ Parameter sind. In unserem Fall ergibt sich mit Hilfe der Methode der kleinsten Quadrate (s. Anhang II)

$$\text{Ausnutzung} = 1.854 - 0.209\,L \qquad \text{für } L = 4,5,\ldots,8.$$

Der Determinationskoeffizient (s. Anhang I) $D = 0.98$ deutet an, dass eine Gerade die Daten vorläufig sehr gut erfasst. Die theoretischen Werte findet man in Tabelle 2.2 und die graphische Darstellung in Abbildung 2.1.

Tabelle 2.2
Beobachtete und berechnete Ausnutzung der Typen

| Länge | Beobachtete Ausnutzung | Berechnete Ausnutzung |
|---|---|---|
| 4 | 1.00 | 1.018 |
| 5 | 0.75 | 0.809 |
| 6 | 0.67 | 0.600 |
| 7 | 0.50 | 0.391 |
| 8 | 0.08 | 0.182 |
| | $A = 1.854$, $B = -0.209$, $D = 0.98$ | |

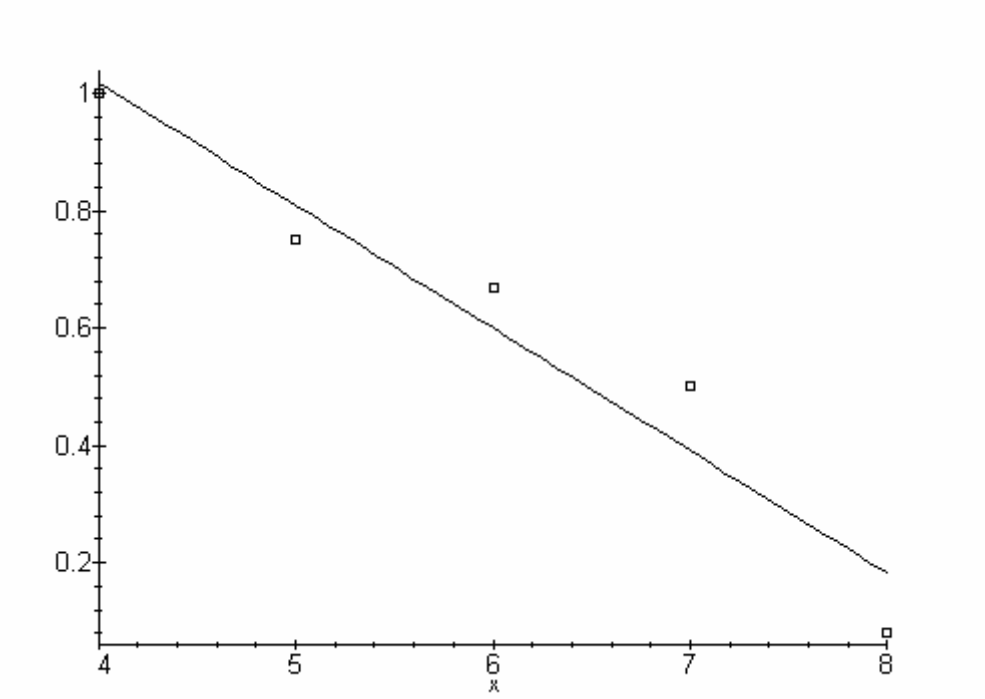


Abbildung 2.1. Ausnutzung der Typen

An dieser Stelle kann man die Vermutung aussprechen, dass diese Abhängigkeit rein lokal ist, d.h. nur für das gegebene Gedicht gilt. Dies hängt sicherlich auch mit dessen Länge zusammen, denn in längeren Gedichten gibt es für alle Muster eines Typs mehr Chancen verwirklicht zu werden. Die Einbeziehung der Gedichtlänge in die Rechnung würde eine Charakteristik der rhythmischen Strenge des Gedichts ergeben. An einem Ende dieser Skala stünden Gedichte mit deterministischem Rhythmus, am anderen Ende solche mit völlig freiem Rhythmus. Unter diesem Standpunkt könnte man die Versifikation sowohl geschichtlich als auch sprachbedingt untersuchen, wobei nicht nur die von uns gewählte, sondern jede beliebige andere Musterbildung interessant wäre. Theoretisch interessant wäre diejenige Musterbildung, die es uns erlauben würde, sprachbezogene Schlüsse zu ziehen bzw. die Entwicklung in Form einer Kurve darzustellen.

Man kann sich gut vorstellen, dass sich mutatis mutandis auch der Prosarhythmus untersuchen lässt. Hier können wir eine derartige Untersuchung nur andeuten. Man teile den Text in Sätze auf, wobei jeder Satz jetzt in etwa die Rolle der Strophe übernimmt. Man stelle die einzelnen Muster und ihre Anzahl fest und vergleiche diese mit den theoretischen Anzahlen, die sich genauso wie in der Poesie aus den Multinomialkoeffizienten ergeben. Zu beachten ist, dass dieses Verfahren nur die Ausnutzung der Typen betrifft. Es besteht aber auch die Möglichkeit, dass jeder Satz eine Konstruktion sui generis ist, wenn der Text nicht lang genug ist. Jegliche Urteile wären hier voreilig.

### 2.2.2. Typenfrequenz

Im vorigen Abschnitt haben wir nur die Realisierung von Typen berücksichtigt. Die einzelnen rhythmischen Verstypen kommen aber in einem Gedicht mit unterschiedlicher Häufigkeit vor. Es gibt wiederum mehrere Möglichkeiten nach Regularitäten zu suchen.

(a) Wir benutzen Schema (I) aus Abschnitt 2.1 und stellen folgende Häufigkeiten der elf vorhandenen Muster fest:

| Muster(Typ) | Häufigkeit |
|---|---|
| 1111 | 2 |
| 1112 | 6 |
| 1121 | 9 |
| 1211 | 1 |
| 1122 | 4 |
| 1212 | 1 |
| 2112 | 1 |
| 2121 | 1 |
| 1222 | 5 |
| 2122 | 1 |
| 3211 | 1 |

Diese Anordnung ist sozusagen alphabetisch, und aus ihr lassen sich beliebige Regularitäten nur durch eine entsprechende Umordnung ermitteln. Eine andere mögliche Anordnung dieser Daten ist beispielsweise die Rangordnung, d.h. ihre Ordnung nach der Häufigkeit, was in der quantitativen Linguistik ein übliches Verfahren darstellt. So erhalten wir die Rangordnung in Tabelle 2.3.

Tabelle 2.3
Ranghäufigkeitsverteilung der rhythmischen Muster

| Muster | Rang | Häufigkeit |
|---|---|---|
| 1121 | 1 | 9 |
| 1112 | 2 | 6 |
| 1222 | 3 | 5 |
| 1122 | 4 | 4 |
| 1111 | 5 | 2 |
| 1211 | 6 | 1 |
| 1212 | 7 | 1 |
| 2112 | 8 | 1 |
| 2121 | 9 | 1 |
| 2122 | 10 | 1 |
| 3211 | 11 | 1 |

Allgemein gilt, dass es für jede „korrekt“ ermittelte Klasse sprachlicher Erscheinungen eine Ranghäufigkeitsverteilung gibt, jedoch ist nicht immer a priori bekannt, welches Modell einer allgemeineren Theorie sich in den gegebenen Daten realisiert. Daher verfährt man bei der Suche nach dem Modell zuerst induktiv und sucht mit Hilfe einer Software (vgl. Altmann-Fitter 1997) nach der Klasse der adäquaten Verteilungen, die man als zunächst mögliche Hypothesen beibehält. Es lässt sich auf diese Weise zeigen, dass für die obigen Daten eine große Zahl von theoretischen Verteilungen adäquat ist. Dies ergibt sich üblicherweise, wenn der Datenumfang recht klein ist (hier $N$ = 32) und der monotone Verlauf recht regulär. In einem solchen Fall – wenn es noch keine ausgereifte Theorie gibt – akzeptiert man vorläufig die Verteilung mit der kleinsten Anzahl von Parametern und mit der besten Anpassung. Diesem Kriterium entspricht hier am besten die 1-verschobene geometrische Verteilung

$$(2.3) \quad P_x = pq^{x-1}, \quad x = 1,2,3,\ldots,$$

deren Werte in der letzten Spalte von Tabelle 2.4 zu finden sind. Die Werte für $x$ = 7 und 8, und für $x$ = 9 und 10 wurden zusammengefasst, damit die theoretische Klassengröße mindestens 1 beträgt. Der Parameter $p$ und das Resultat des Chiquadrat-Tests sind in der letzten Zeile der Tabelle zu finden. Wegen $P$ = 0.98 kann man diese Verteilung vorläufig gut akzeptieren. Die Resultate aus Tabelle 2.4 sind in Abbildung 2.2 graphisch dargestellt.

Tabelle 2.4
Anpassung der geometrischen Verteilung
an die Rangordnung der rhythmischen Muster

| Rang | Beobachtet | Berechnet nach (2.3) |
|---|---|---|
| 1 | 9 | 8.70 |
| 2 | 6 | 6.34 |
| 3 | 5 | 4.61 |
| 4 | 4 | 3.36 |
| 5 | 2 | 2.45 |
| 6 | 1 | 1.78 |
| 7 | 1 | 1.30 |
| 8 | 1 | 0.95 |
| 9 | 1 | 0.69 |
| 10 | 1 | 0.50 |
| ≥11 | 1 | 1.33 |
| $p$ = 0.2719, $q$ = 1 – $p$ = 0.7281, $X^2$ = 1.27, FG = 7, $P$ = 0.98 | | |

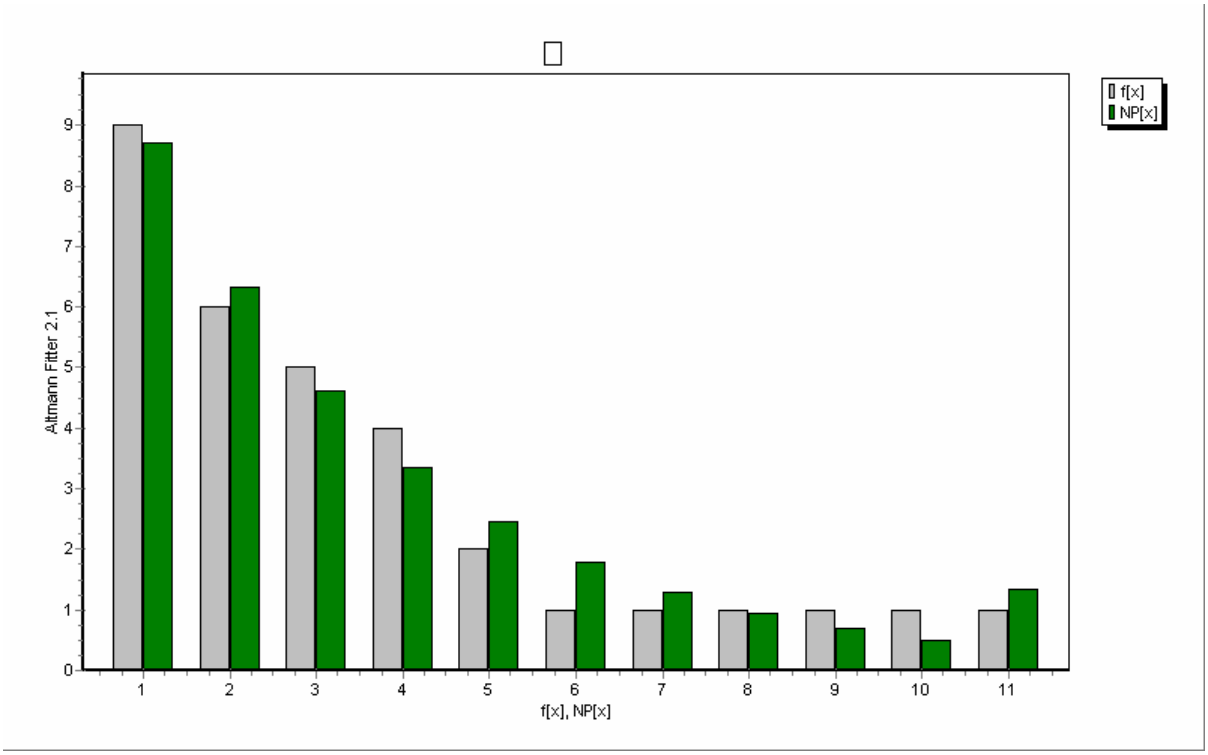


Abbildung 2.2. Anpassung der geometrischen Verteilung an die Rangordnung der rhythmischen Muster

(b) Ein ähnliches Bild bietet sich, wenn wir die Rangverteilung der zusammengefassten Typen in Schema (V) betrachten. So erhalten wir die in Tabelle 2.5 und in Abbildung 2.3 dargestellten Resultate.

Tabelle 2.5
Ranghäufigkeitsverteilung der nach Länge zusammengefassten Muster (Typen)

| Typ (Muster) | Rang | Beobachtete Häufigkeit | Berechnete Häufigkeit (nach 2.3) |
|---|---|---|---|
| b | 1 | 16 | 16.18 |
| c | 2 | 7 | 8.00 |
| d | 3 | 6 | 3.96 |
| a | 4 | 2 | 1.96 |
| e | 5 | 1 | 1.91 |
| $p = 0.5055$, $X^2 = 1.62$, FG = 3, $P = 0.65$ | | | |

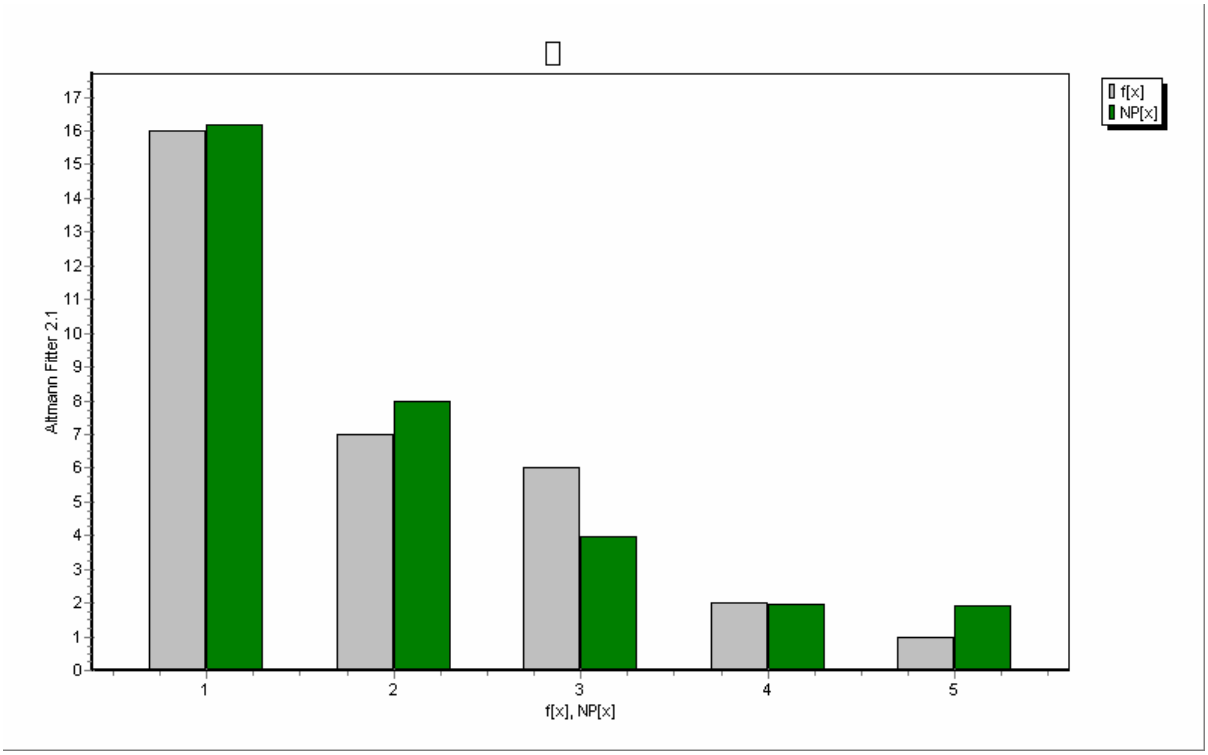


Abbildung 2.3. Ranghäufigkeitsverteilung der nach Länge zusammengefassten Typen

Zu beachten ist, dass für diese Zwecke besonders die sog. Partialsummenverteilungen geeignet sind (vgl. Wimmer, Altmann 2000a,b; Wimmer, Šidlík, Altmann 1999). Wir begnügen uns hier mit der obigen einfachen Lösung.

(c) Ein ganz anderes Bild erhalten wir, wenn wir nicht die Rangordnung, sondern die Länge der Typen als unabhängige Zufallsvariable betrachten. In diesem Falle erhalten wir die entsprechende Verteilung aus Schema (V), wie in Tabelle 2.6 und Abbildung 2.4 dargestellt.

Tabelle 2.6
Verteilung der rhythmischen Muster nach der Länge

| Länge des Musters | Beobachtete Häufigkeit | Berechnete Häufigkeit (nach 2.4) |
|---|---|---|
| 4 | 2 | 1.82 |
| 5 | 16 | 14.25 |
| 6 | 7 | 10.53 |
| 7 | 6 | 4.08 |
| 8 | 1 | 1.32 |
| $a$ = 0.8157, $b$ = 0.1042, $X^2$ = 2,40, FG = 2, $P$ = 0.30 | | |

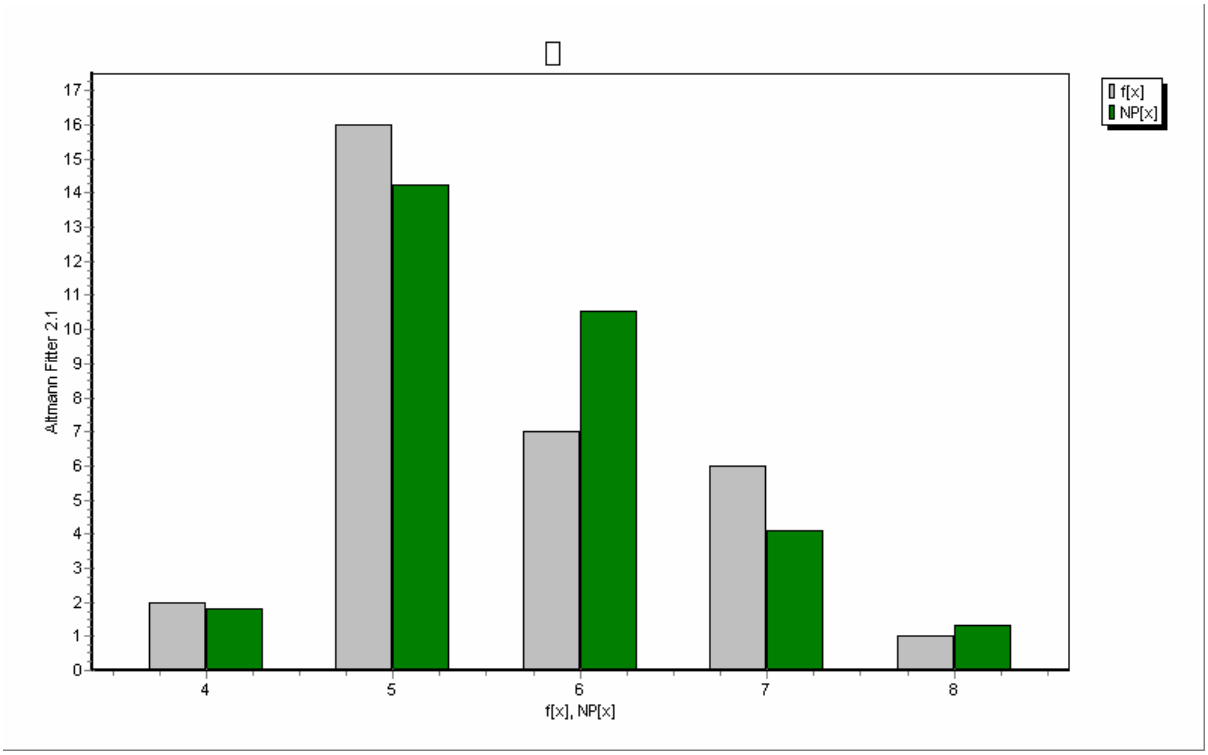


Abbildung 2.4. Anpassung der Hyperpoisson-Verteilung an die Verteilung der Musterlängen

Die beobachtete Verteilung ist nicht monoton fallend, daher müsste man entweder die geometrische Verteilung modifizieren – wenn man bei dem gegebenen Modell bleiben möchte –, was die Zunahme eines weiteren Parameters bedeutet, oder nach einer kompakten Verteilung mit zwei Parameters suchen. Die Software hat für diesen Fall die Hyperpoisson-Verteilung als am geeignetsten gefunden, die in der Linguistik sehr oft benutzt wird, nämlich (hier in der 4-verschobenen Form)

$$(2.4)\quad P_x = \frac{a^{x-4}}{b^{(x-4)}\,{}_1F_1(1;b;a)}, \quad x = 4,5,6,...,$$

wo $k^{(r)} = k(k+1)(k+2)\ldots(k+r-1)$ die Faktorialpotenz (in Formel (2.4) setzt man $k = b$ und $r = x-4$ ein) und ${}_1F_1(1;b;a) = 1+\frac{a}{b}+\frac{a^2}{b(b+1)}+...$ die hypergeometrische Funktion sind. Die nach dieser Formel berechneten Werte sind in der letzten Spalte von Tabelle 2.6 zu finden. Der Chiquadrat-Test zeigt, dass die Anpassung sehr gut ist.

Nur weitere Untersuchungen können zeigen, inwieweit man diese Verteilungen praktisch verwenden bzw. in eine Theorie einbetten kann. Vorläufig kön-

nen wir nur behaupten, dass die Häufigkeitsverteilung der rhythmischen Muster weder „normal“ noch „uniform“ ist, sondern einem ausgeprägten Modell folgt.

Die Modellierung eines Verlaufes, einer Verteilung, ist aus vielen Gründen wichtig und sollte so früh wie möglich angestrebt werden. Erstens gibt sie uns ein Bild darüber, ob die Daten überhaupt strukturiert sind oder chaotisch verlaufen. Falls sie strukturiert sind, darf man hoffen, die Struktur zu finden. Falls sie chaotisch sind, hat man die Hoffnung, den Charakter des Chaos zu bestimmen. Zweitens erlaubt uns das Modell zunächst Verallgemeinerungen vorzunehmen. Diese wiederum ermöglichen es uns, auf eine höhere Generierungsebene zugelangen, d.h. nach sehr allgemeinen Faktoren zu suchen, die die Erscheinung erzeugen, beispielsweise von linguistischen Daten auf psychologische Faktoren zu schließen, allgemeine Zufallsprozesse zu berücksichtigen, wie sie auch in anderen Wissenschaften verwendet werden. Das wiederum ermöglicht es für die Textanalyse einen Anschluss z.B. an die Systemtheorie zu finden. Drittens geht es in der Wissenschaft immer um Erklärungen, und diese kann man nur mit Hilfe von Gesetzen suchen, die Teile einer Theorie sind. Voraussetzung für die Etablierung eines Gesetzes ist eine testbare Hypothese, und diese muss in der Textwissenschaft mit quantifizierten und messbaren Begriffen so formuliert werden, dass sie leicht in die Sprache der Mathematik übersetzbar ist. Der Test zeigt ihre Chance in eine Theorie aufgenommen zu werden, bedeutet aber nicht sofort, dass sie den Status eines Gesetzes bereits erlangt hat. Dazu muss noch einiges getan werden, nämlich sie muss deduktiv abgeleitet werden, d.h., sie muss aus einer Theorie, aus anderen Gesetzen oder aus Axiomen folgen.

In der Textwissenschaft verfährt man zunächst induktiv, d.h., man versucht Begriffe zu quantifizieren (operationalisieren) und aufgrund von Messungen an individuellen Texten arbeitet man sich dann zu einer zunächst sehr vagen Hypothese durch, z.B. „es gibt einen Zusammenhang zwischen X und Y“ oder „X folgt einer ´ordentlichen´ Wahrscheinlichkeitsverteilung“ oder „in der Menge A kann man eine Rangordnung der Elemente etablieren“ usw. Man führt anschließend die entsprechenden Operationen durch, wobei man sich heutzutage meistens verschiedener Softwares bedient.

Solange in einer Hypothese Beobachtungsbegriffe oder spezifische Begriffe vorhanden sind, z.B. „im Deutschen gilt“ oder „im Erlkönig ist X verteilt als“, stehen wir auf einer theoretischen Vorstufe. Auch viele analysierte Texte „verbessern“ diese Situation nicht, auch wenn die Forschung auf diese Weise voranschreitet. Erst wenn der versteckte Mechanismus entdeckt wird, die Regularität deduktiv abgeleitet werden kann und für alle Texte, die bestimmte Bedingungen erfüllen, gilt, können wir von einer Theorie sprechen. Ansätze dazu sind bereits vorhanden.

Das Ziel diese Buches besteht aber eher darin, die induktiven Möglichkeit zu zeigen, mit deren Hilfe wir imstande sind, die Existenz einer regulären Erscheinung aufzuspüren.

## 2.3. Globale Maße

Auf dem Wege zur Quantifizierung ist es üblich, eine Erscheinung zumindest global zu charakterisieren, um wenigstens feststellen zu können, in welchem Intervall sich eine Eigenschaft bewegt. Dies gibt uns ein erstes Bild der Eigenschaft und ermöglicht es später, deren Zusammenhänge mit anderen Eigenschaften zu erforschen. Üblicherweise übernimmt man hier die bekannten Maßzahlen aus der Statistik wie Mittelwert, Streuung, Variationskoeffizient, Koeffizienten der Schiefe und des Exzesses, Median u.a. (vgl. z.B. Altmann, Lehfeldt 1980: 142ff.). In der Linguistik hat es sich seit Herdan (1956) eingebürgert, für diese Zwecke zwei Maße zu benutzen, nämlich die Wiederholungsrate (repeat rate, Herfindahlsches Konzentrationsmaß) und die Entropie. Ihre Benutzung sollte ein Anlass sein, mehr Kontakt zu anderen Wissenschaften zu suchen, wo sie täglich benutzt werden. Wenngleich wir dies dem Leser überlassen, können wir zeigen, dass diese zwei Maße auch einer Regularität folgen.

*Die Wiederholungsrate*

Die Wiederholungsrate wird definiert als

$$(2.5) \quad R = \sum_x p_x^2 ,$$

wo $p_x$ die Wahrscheinlichkeiten oder im empirischen Fall die relativen Häufigkeiten sind und $x$ die (diskrete) Variable ist, die hier alle ihre Werte durchläuft. Wir schätzen $R$ mit Hilfe der relativen Häufigkeiten ab und bekommen

$$(2.5a) \quad R = \frac{1}{N^2} \sum_x f_x^2 ,$$

wo $f_x$ die absoluten Häufigkeiten und $N$ ihre Summe, d.h. $N = \sum f_x$ sind. Das Maß ist sehr leicht zu berechnen. So bekommen wir für die Typenfrequenzen aus Tabelle 2.3/2.4

$$\sum_x f_x^2 = 9^2 + 6^2 + 5^2 + 4^2 + 2^2 + 6(1^2) = 168 .$$

Da die Summe aller Häufigkeiten $N = 32$, erhalten wir

$$R = 168/32^2 = 0.1641.$$

Für Tabelle 2.5/2.6 erhalten wir

$$\sum_x f_x^2 = 16^2 + 7^2 + 6^2 + 2^2 + 1^2 = 346,$$

woraus

$$R = 346/32^2 = 0.3338.$$

Wie Altmann und Lehfeldt (1980: 151ff.) gezeigt haben, ist aufgrund der geometrischen Verteilung der theoretische Wert von $R$ zumindest im phonischen Bereich gleich

$$(2.6) \quad R_t = \frac{2}{K},$$

wo $K$ die Inventargröße von Entitäten oder Klassen von Entitäten bezeichnet. Es ist noch zu prüfen, ob dies auch für andere Bereiche der Sprache und Texte gilt. Hier werden wir dies sehr einfach überprüfen. In Tabelle 2.4 haben wir 11 Klassen (Typen, Muster), d.h., die Inventargröße ist 11, daher ergibt sich der theoretische Wert (2.6) als $R_t = 2/11 = 0.18$, so dass der beobachtete Wert ($R = 0.16$) dicht darunter liegt. Für Tabelle 2.5, wo das Inventar $K = 5$ ist, erhalten wir $R_t = 2/5 = 0.4$, und auch hier liegt der beobachtete Wert ($R = 0.33$) nah darunter, wie man in Abbildung 2.5 sehen kann.

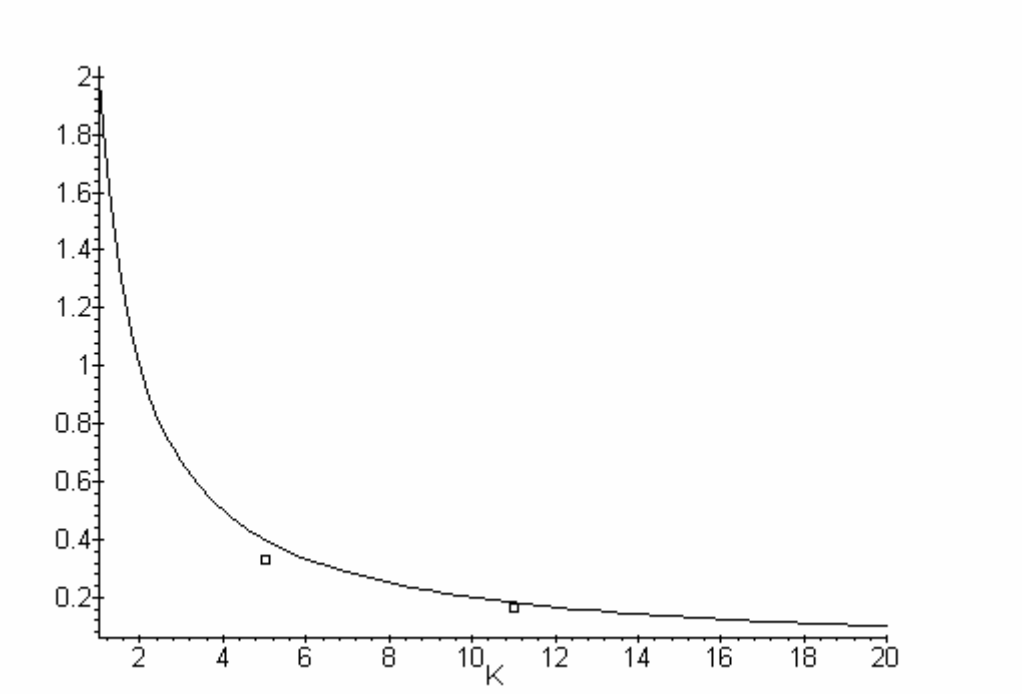


Abbildung 2.5. Theoretische Kurve $R_t = 2/K$ der Wiederholungsrate und die Lage von Erlkönig (Punkte)

Die Kurve zeigt uns, dass textuelle Daten so angeordnet sind, dass die Wiederholungsrate von der Inventargröße abhängt, womit ein erforschungswerter Zusammenhang thematisiert wird. Man kann schließen, dass, wenn die Hypothese gilt, die Inventargröße darüber entscheidet, wie die gegebene Entität verteilt sein wird. Die obige theoretische Kurve gilt nur dann, wenn die Daten geometrisch verteilt sind, was nicht immer der Fall zu sein braucht. Folgen sie z.B. der Zipf-Mandelbrotschen Verteilung, dann ergibt sich eine andere Kurve, jedoch mit sehr ähnlichem Verlauf (vgl. Zörnig, Altmann 1983).

*Entropie*

Etwas langwieriger, aber keineswegs schwieriger ist die Berechnung der Entropie, die aus der Informationstheorie übernommen wurde und sich daher an eine weit entwickelte Disziplin anlehnt. Sie wird in der Shannonschen Form als

$$(2.7) \quad H = -\sum_x p_x \, ld \; p_x$$

definiert, wobei $p_x$ wieder die Wahrscheinlichkeiten sind und *ld* der Logarithmus zur Basis 2. Man kann aber auch andere Logarithmen benutzen und eventuell (aber nicht unbedingt) in den dyadischen Logarithmus transformieren und zwar mit Hilfe von $ld\ x = log_a\ x/log_a\ 2$, wo *a* eine beliebige Basis bedeutet. Beispielsweise ist mit natürlichen Logarithmen $ld\ x = ln\ x/ln\ 2$. Da wir *H* auch mit relativen Häufigkeiten schätzen, ergibt sich für den empirischen Fall aus (2.7) – indem wir $p_x = f_x/N$ einsetzen –

$$(2.7.a) \; H = ld\ N - \frac{1}{N}\sum_x f_x \, ld \; f_x \, .$$

Für Tabelle 2.3 erhalten wir

$$H = ld\ 32 - [9\ ld\ 9 + 6\ ld\ 6 + 5\ ld\ 5 + 4\ ld\ 4 + 2\ ld\ 2 + (6)1\ ld\ 1]/32 =$$

$$= \; 5 - [28.5293 + 15.5098 + 11.6096 + 8 + 2 + 0]/32$$

$$= \; 5 - 2.0515 = 2.9485.$$

Für Tabelle 2.5 erhalten wir

$$H = ld\ 32 - [16\ ld\ 16 + 7\ ld\ 7 + 6\ ld\ 6 + 2\ ld\ 2 + 1\ ld\ 1]/32 =$$

$$= 5 - (64 + 19.6515 + 15.5098 + 2 + 0)/32 = 1.8387.$$

Auch für die Entropie gibt es theoretische Erwartungen, die in der Phonologie gelten. Aufgrund der geometrischen Verteilung haben Altmann und Lehfeldt (1980: 172) die Funktion

$$(2.8) \quad H_t = -ld\left[\left(\frac{4}{K+2}\right)\left(\frac{K-2}{K+2}\right)^{\frac{K-2}{4}}\right]$$

abgeleitet. Für $K = 11$ bekommen wir

$$H_t = -ld\left[\left(\frac{4}{13}\right)\left(\frac{9}{13}\right)^{\frac{9}{4}}\right] = 2.8941$$

und für $K = 5$

$$H_t = -ld\left[\left(\frac{4}{7}\right)\left(\frac{3}{7}\right)^{\frac{3}{4}}\right] = 1.7241.$$

Die empirischen Werte liegen jetzt etwas oberhalb der theoretischen Kurve, jedoch nah genug, um die theoretische Kurve zu akzeptieren (s. Abb. 2.6).

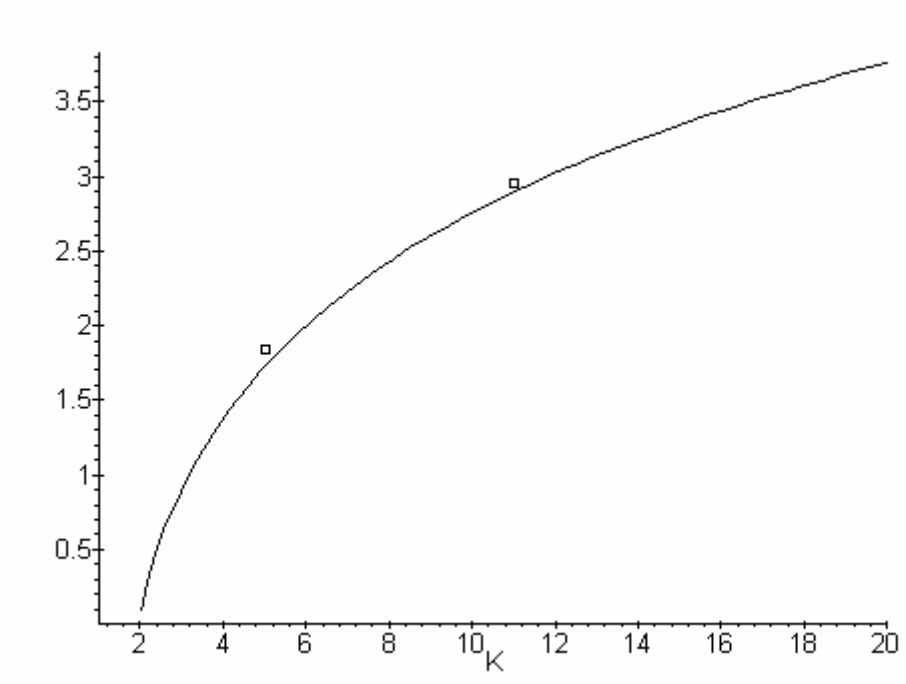


Abbildung 2.6. Theoretische Kurve für die Entropie und die Lage des „Erlkönig“

Wieder hängt die theoretische Kurve davon ab, welche Verteilung ihr zugrundeliegt. Für die Zipf-Mandelbrot-Verteilung haben Zörnig und Altmann (1984) eine etwas andere, jedoch wieder sehr ähnlich verlaufende Kurve erhalten, was wieder darauf hindeutet, dass die Entropie mit dem Inventarumfang zusammenhängt. Dieses Resultat bestätigt auch die Forschung im Worthäufigkeitsbereich, wo das Inventar die Zahl unterschiedlicher Wörter im Text (Types) darstellt. Cohen, Mantegna und Havlin (2004) haben als erste Approximation eine empirische Kurve vorgeschlagen, die die Abhängigkeit sehr gut erfasst, theoretisch aber noch nicht begründet ist.

In diesem Bereich ergeben sich interessante, untersuchungswerte Probleme, die man weiter verfolgen kann, wie. z.B.:

(a) Wie gestalten sich die Verteilungen von rhythmischen Mustern und die Wiederholungsrate bzw. die Entropie in anderen Gedichten von Goethe?

(b) Wie gestalten sie sich bei anderen Autoren?

(c) Gibt es eine Entwicklung oder zumindest Variabilität in der deutschen Literatur, oder ist diese Erscheinung hier konstant?

(d) Wie verhält es in anderen Sprachen, in denen eine ähnliche Versifikation möglich ist? Haben wir es hier mit einem allgemeinen Phänomen zu tun, oder ist das Deutsche (bzw. nur der „Erlkönig“) ein Fall sui generis?

(e) Kann man die Problematik auf den Rhythmus der Prosa übertragen?

Man kann Verteilungen nicht nur für rhythmische Erscheinungen aufstellen, sondern für alles, was variabel ist, und daher ergibt sich hier ein breites Forschungsfeld. Im vorliegendem Fall ist die einzige Bedingung, dass die Variable diskret ist bzw. sich als diskret darstellen lässt.

## 2.4. Klimax im Vers

Wie man in den einzelnen Schemata (I) – (IV) sieht, ist das Gedicht eine Folge von rhythmischen Mustern. Die klassische Poetik untersucht nur die Muster eines Verses, stellt dessen Form fest und klassifiziert den Typ des Verses. Man kann sich aber auch fragen, ob das Gedicht als Ganzes bestimmte rhythmische Tendenzen aufweist, die man als nichtzufällig bezeichnen kann. Bei Versen mit deterministischer rhythmischer Regularität ist dies automatisch gegeben, aber bei Versen, in denen ungleiche Versfüße in anscheinend irregulärer Anordnung vorhanden sind, muss eine mögliche Tendenz erst entdeckt werden. Probleme dieser Art hat man bereits im 19. Jahrhundert erörtert (s. z.B. Drobisch 1966, 1968a,b), man findet sie sehr oft bei den Strukturalisten erwähnt, aber auch quantitative Linguisten haben es unternommen, dieses Problem in exakterer Form zu behandeln (vgl. Altmann, Štukovský 1963, 1965).

Stellen wir uns hier die Frage, ob es eine allgemeine Tendenz gibt, die vier unbetonten Positionen vom Anfang bis zum Ende des Verses mit steigender An-

zahl der Silben auszufüllen. An einzelnen Versen kann man eine derartige Tendenz nicht nachweisen, es gibt einzelne Muster, die gegen sie sprechen, z.B. 2121. Betrachtet man aber das ganze Gedicht und stellt die Zahl der unbetonten Silben in einzelnen Positionen (p) fest, dann sieht man in Schema (I), dass in der ersten Position 29-mal die 1 und 3-mal die 2 vorkommt. Die Summe in der ersten Spalte ist also 29 + 6 = 35, und der Durchschnitt ergibt sich dann als 35/32 = 1.09. Auf die gleiche Weise erhalten wir für die einzelnen Spalten die in der dritten Spalte der Tabelle 2.7aufgeführten Resultate.

Tabelle 2.7
Durchschnittliche Silbenzahl in einzelnen unbetonten Positionen

| Position $p$ | Silbenzahl $S$ | Durchschnitt $D$ | Berechnetes $D_p$ |
|---|---|---|---|
| 1 | 35 | 1.09 | 1.13 |
| 2 | 41 | 1.28 | 1.31 |
| 3 | 53 | 1.66 | 1.51 |
| 4 | 51 | 1.59 | 1.70 |
| | $a$ = 0.935, $b$ = 0.191, $D$ = 0.99 | | |

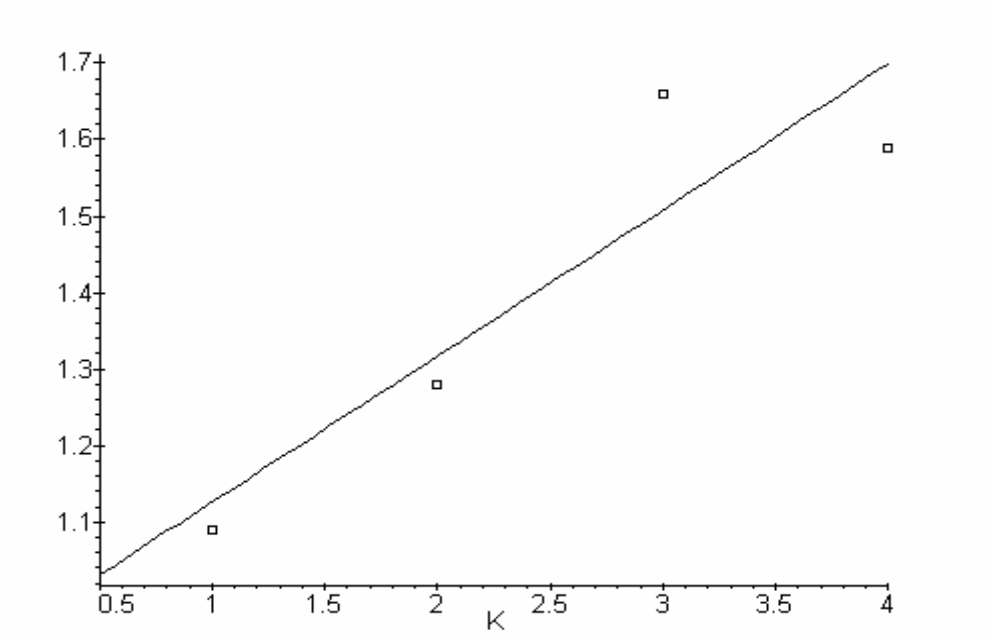


Abbildung 2.7. Rhythmische Klimax im Erlkönig

Wie man sieht, ist hier keine perfekte Klimax vorhanden (Position 4 ist kleiner als Position 3), sondern eher ein spezifischer Fall, bei dem die zweite Hälfte des Verses als Ganze größer ist als die erste. Dennoch erweist sich dieser mit einer Geraden erfasste Trend noch immer als signifikant, wie der Wert des Determinationskoeffizienten $D$ = 0.99 zeigt. Die Gerade lautet in unserem Fall

$$D_p = 0.935 + 0.191p.$$

Die Werte der berechneten Geraden sind in der letzten Spalte von Tabelle 2.7 und in Abbildung 2.7 zu sehen.

Die Berechnung der Parameter einer Kurve und deren Anpassung führt man heute mit der entsprechenden Software durch. Allgemeine statistische Softwares benutzen dazu die Methode der kleinsten Quadrate, die in den meisten Fällen ausreicht, spezielle Softwares verwenden die iterative Optimierung, die die Anpassung noch etwas verbessern kann. Der Determinationskoeffizient, den wir einfachheitshalber statt des *F*-Tests oder des *t*-Tests benutzen, sagt lediglich, wie groß die durch die Kurve erklärte Summe der quadratischen Abweichungen ist. Die Formel lautet

$$D = 1 - \frac{\sum_x (y_x - \hat{y}_x)^2}{\sum_x (y_x - \overline{y})^2},$$

wobei $\hat{y}_x$ der aus der theoretischen Kurve berechnete Wert der Variablen und $\overline{y}$ der Mittelwert der Variablen sind. *D* bewegt sich in dem Intervall <0; 1>. Man pflegt Werte größer als 0.9 als sehr gut, Werte über 0.8 als gut zu betrachten. Unser Wert *D* = 0.99 ist also sehr gut. Da wir den Determinationskoeffizienten ständig benutzen werden, findet man ihn auch im Anhang I.

## 2.5. Ein erster Blick auf Iterationen

Betrachten wir Schema (III), in dem die rhythmischen Muster mit den Buchstaben *a, b, c, d, e* kodiert sind:

(III) *bbbb cbba abbc dcbb bcdd dbbc eddb ccbb*.

Als erstes stellt sich die Frage, ob eine derartige Folge zufällig entstanden ist, oder ob man in ihr Tendenzen erkennen kann, die man als nichtzufällig bezeichnen darf. Hat man nämlich eine gegebene Anzahl unterschiedlicher Buchstaben, dann kann man diese auf verschiedene Weisen hintereinander anordnen. Es gibt jedoch recht feste Regeln, die es uns ermöglichen zu entscheiden, ob eine derartige Folge als zufällig oder als nicht zufällig zu betrachten ist. Eine Folge der obigen Art kann zahlreiche Aspekte aufweisen.

Bezeichnen wir als Iteration (engl. *run*) eine Sequenz von gleichen Buchstaben in (III), wobei auch ein einfaches Symbol eine Iteration ist, so bekommen wir (ohne Rücksicht auf die leeren Stellen, die Strophenenden andeuten)

*bbbb c bb aa bb c d c bbb c ddd bb c e dd b cc bb*

d.h. insgesamt $r = 18$ Iterationen. Um zu testen, ob wir es hier mit einer signifikant großen oder kleinen Anzahl von Iterationen zu tun haben, können wir bei $n = 32$ Elementen zu der Normalverteilung übergehen und testen

$$(2.9) \quad u = \frac{v - E(v)}{\sigma_v},$$

wobei $v = n - r$, $E(v)$ die theoretische Erwartung und $\sigma_v$ die Standardabweichung symbolisieren.

Bezeichnen wir mit $k_i$ ($i \in T = \{a, b, c, d, e\}$) die Anzahl der einzelnen Buchstaben in der obigen Sequenz, d.h.

$k_a = 2$
$k_b = 16$
$k_c = 7$
$k_d = 6$
$k_e = 1,$

so dass $n = \sum_{i \in T} k_i = 32$, d.h. die Zahl der Verse insgesamt. Weiter definieren wir die Größen

$$F_2 = \sum_{i \in T} k_i (k_i - 1)$$
$$F_3 = \sum_{i \in T} k_i (k_i - 1)(k_i - 2),$$

die wir für weitere Berechnungen brauchen. Die Erwartung ergibt sich als

$$(2.10) \quad E(v) = \frac{F_2}{n}$$

und die Standardabweichung als

$$(2.11) \quad \sigma_v = \sqrt{\frac{(n-3)F_2}{n(n-1)} + \frac{F_2^2}{n^2(n-1)} - \frac{2F_3}{n(n-1)}}.$$

Alle diese Größen berechnen wir mit Hilfe der Zahlen $k_i$ als

$F_2 = 2(1) + 16(15) + 7(6) + 6(5) + 1(0) = 314$
$F_3 = 0 + 16(15)14 + 7(6)5 + 6(5)4 + 0 = 3690$

$$E(v) = 314/32 = 9.8125$$

$$\sigma_v = \sqrt{\frac{(32-3)314}{32(31)} + \frac{314^2}{32^2(31)} - \frac{2(3690)}{32(31)}} = 2.2013,$$

woraus wegen $v = n - r = 32 - 18 = 14$

$$u = \frac{14 - 9.8125}{2.2013} = 1.90$$

folgt. Diese Zahl ist kleiner als 1.96, d.h., wir haben es auf der $\alpha = 0.05$-Ebene bei zweiseitigem Test mit einer zufälligen Zahl von Iterationen zu tun. Bei einseitigem Test ist der kritische Wert auf der $\alpha = 0.05$-Ebene $u = 1.64$, und wir hätten zu entscheiden, dass die Zahl der Iterationen nicht zufällig ist. Jedoch haben wir keine diesbezügliche Hypothese aufgestellt. Der Normaltest ist nur asymptotisch und bei kleinen Umfängen möglicherweise etwas verzerrt.

Unser Resultat liegt also an der Grenze, und in einem solchen Fall ist es empfehlenswert, entweder die Wahrscheinlichkeit umständlich, aber exakt zu berechnen (vgl. Mood 1940; Barton, David 1957) oder noch einen anderen Test durchzuführen. Hier wählen wir den zweiten Weg.

## 2.6. Der Diagonaltest

Wir betrachten das Problem nun von einer anderen Seite und fragen, ob es vielleicht eine Tendenz gibt, in einer Sequenz hinter einen Buchstaben vorzugsweise den gleichen zu stellen. Wenn eine solche Tendenz existiert, dann werden die Folgen von gleichen Buchstaben (Iterationen) länger, d.h., die Anzahl der Iterationen selbst wird dadurch sinken, deren Länge aber anwachsen. Dieses Problem ist eher aus der Linguistik bekannt, wo es unter dem Namen „Vokalharmonie“ untersucht wird.

Um diese Tendenz – „gleiche Buchstaben vorzugsweise hintereinander“ – zu testen, verfahren wir folgendermaßen. Mit $n_{ij}$ bezeichnen wir die Häufigkeit, mit der in Schema (III) hinter dem Buchstaben $i$ der Buchstabe $j$ steht, und erhalten z.B.

$$n_{aa} = 1, \qquad n_{ab} = 1, \qquad n_{ac} = 0,$$
$$n_{ba} = 0, \qquad n_{bb} = 9, \qquad n_{bc} = 5,$$

usw. Alle diese Übergänge sind in Tabelle 2.8 dargestellt, wo links die ersten Buchstaben, oben die Folgebuchstaben stehen. Die Zellen auf der Diagonalen ($n_{ii}$) bedeuten „gleiche Buchstaben hintereinander“.

Tabelle 2.8
Übergänge zwischen den Symbolen in Schema (III)

| | a | b | c | d | e | ∑ |
|---|---|---|---|---|---|---|
| a | 1 | 1 | 0 | 0 | 0 | 2 |
| b | 1 | 9 | 5 | 0 | 0 | 15 |
| c | 0 | 3 | 1 | 2 | 1 | 7 |
| d | 0 | 2 | 1 | 3 | 0 | 6 |
| e | 0 | 0 | 0 | 1 | 0 | 1 |
| ∑ | 1 | 15 | 7 | 6 | 1 | 31 |

Um die Überbelegung der Diagonale zu testen, hat Altmann (1987) zwei gleichwertige Tests vorgeschlagen, nämlich

$$(2.12)\quad u = \frac{S - \sum_i \frac{n_i.n._i}{n}}{\sqrt{\frac{A+B}{n^2(n-1)}}},$$

wo $S = n_{11} + n_{22} + \ldots + n_{kk}$ (d.h. die Summe der Zellen auf der Diagonalen)

$$A = \sum_i n_i.n._i(n - n_i.)(n - n._i)$$

$$B = 2\sum_{i<i'} n_i.n._i n_{i'}.n._{i'}$$

und

$$(2.13)\quad X^2 = \frac{n(nS - \sum n_i.n._i)^2}{\sum n_i.n._i(n^2 - \sum n_i.n._i)}$$

wobei $u^2 \approx X^2$. Hier symbolisiert $n_i.$ die Summe der $i$-ten Zeile der Tabelle und $n._i$ die Summe der $i$-ten Spalte. Die Funktion $S$ ist die Summe der Diagonalzellen, $n$ ist die Summe aller Zahlen in der Tabelle. Wir erhalten hier (man sieht, dass die Marginalsummen gleich sind!)

$$S = 1 + 9 + 1 + 3 + 0 = 14$$

$$\sum_i \frac{n_i.n._i}{n} = [2(2) + 15(15) + 7(7) + 6(6) + 1(1)]/31 = 10.1613$$

$$A = 2(2)(31\text{-}2)(31\text{-}2) + 15(15)(31\text{-}15)(31\text{-}15) + 7(7)(31\text{-}7)(31\text{-}7) +$$
$$+ 6(6)(31\text{-}6)(31\text{-}6) + 1(1)(31\text{-}1)(31\text{-}1) = 112588$$

$$B = 2[2(2)15(15) + 2(2)7(7) + 2^2 6^2 + 2^2 1^2 + 15^2 7^2 + 15^2 6^2 + 15^2 1^2 + 7^2 6^2 + + 7^2 1^2 + 6^2 1^2] = 44886.$$

Setzt man diese Zahlen in (2.12) ein, so erhält man

$$u = \frac{14 - 10.1613}{\sqrt{\frac{112588 + 44886}{31^2(30)}}} = \frac{3.8387}{2.3371} = 1.64,$$

und dieser Test sagt etwas klarer als der Test in Abschnitt 2.4, dass die Folge in diesem Sinne zufällig ist, da 1.64 viel kleiner als 1.96 ist.

Die Berechnung von Formel (2.13) ist etwas einfacher, denn

$$X^2 = \frac{31[31(14) - 315]^2}{315(31^2 - 315)} = \frac{438991}{203490} = 2.1573.$$

Hier haben wir 1 Freiheitsgrad mit dem kritischen Wert $\chi^2_{1(0.05)}$ = 3.84. Da unser Wert kleiner ist als 3.84, ist auf der Diagonalen keine Tendenz vorhanden. Weil die Zahlen zu klein sind, ist hier nur eine annähernde Übereinstimmung beider Tests gegeben ($\sqrt{2.1573}$ = 1.48, während $u$ = 1.64), die Aussage ist jedoch gleich.

## 2.7. Zörnigs Distanztest

In den Abschnitten 2.4 und 2.5 haben wir eigentlich nur die Distanzen der Länge 0 in Betracht gezogen, denn eine Null-Distanz ergibt eben eine Iteration. Da wir $v = n - r$ berücksichtigt haben, haben wir eigentlich getestet, ob die Zahl der Nicht-Null-Distanzen signifikant ist oder nicht. Der einseitige Test hat gezeigt, dass die Zahl der Nicht-Null-Distanzen signifikant größer ist als erwartet, daher ist die Zahl der Null-Distanzen (Iterationen) signifikant kleiner als erwartet bzw. bei zweiseitigem Test zufällig.

Wenn aber eine Folge zufällig ist, dann müssen auch die Distanzen zwischen jeweils gleichen Elementen zufällig verteilt sein. Um dies zu testen, hat Zörnig (1984, 1987) ein Verfahren vorgeschlagen, das es uns erlaubt, die theoretische Verteilung der Distanzen zu berechnen und mit der empirischen Verteilung zu vergleichen.

Betrachten wir wieder Schema (III)

*bbbbcbbaabbcdcbbbcdddbbceddbccbb*

und bezeichnen die Distanzgröße als die Zahl der Elemente, die zwischen jeweils zwei gleichen stehen. So gibt es zwischen den beiden *a*

1 Distanz der Größe 0;

zwischen den 16 *b*s haben wir der Reihe nach die Distanzen

0,0,0,1,0,2,0,3,0,0,4,0,4,2,0;

zwischen den 7 *c*s

6,1,3,5,4,0;

zwischen den 6 *d*s

5,0,0,4,0

und da es nur ein *e* gibt, gibt es hier keine Distanz. So erhalten wir die in Tabelle 2.9 angegebene empirische Verteilung der Distanzen.

Tabelle 2.9
Verteilung der Distanzen zwischen rhythmischen Mustern

| Distanz *x* | Anzahl |
|---|---|
| 0 | 14 |
| 1 | 2 |
| 2 | 2 |
| 3 | 2 |
| 4 | 4 |
| 5 | 2 |
| 6 | 1 |

Bezeichnen wir hier als

*n* die Länge der Sequenz
$k_i$ die Häufigkeit einzelner Elemente, $i \in T = \{a, b, c, d, e\}$,

so haben wir, wie bereits in Abschnitt 2.5 angegeben, folgende Anzahlen von Elementen:

$n = 32$
$k_a = 2$

$k_b$ =16
$k_c$ = 7
$k_d$ = 6
$k_e$ = 1.

Die theoretische Zahl der Distanzen $D_x$ der Länge $x$ erhalten wir als

$$(2.14)\quad D_x = \frac{(n-x-1)!}{n!}\sum_{i\in T} k_i(k_i-1)(n-k_i)_{(x)},$$

wo $r_{(x)} = r(r-1)\ldots(r-x+1)$ und $r_{(0)} = 1$ ist. In unserem Fall wird (2.14) aufgrund der obigen Zahlen zu

$$D_x = \frac{(32-x-1)!}{32!}[2(1)(32-2)_{(x)} + 16(15)(32-16)_{(x)} + 7(6)(32-7)_{(x)} + \\ + 6(5)(32-6)_{(x)} + 0].$$

Daraus berechnen wir

$$D_0 = \frac{31!}{32!}[2(1)+16(15)+7(6)+6(5)] = 9.81,$$

und an der Distanz 0 sehen wir, dass die Iterationstheorie eine Spezialfall von Zörnigs Distanztheorie ist, denn $D_0$ ist identisch mit (2.10). Weiter erhalten wir

$$D_1 = \frac{30!}{32!}[2(1)(32-2)+16(15)(32-16)+7(6)(32-7)+6(5)(32-6)] =$$

$$= \frac{1}{32(31)}[2(30)+240(16)+42(25)+30(26)] = 5.78$$

$$D_2 = \frac{29!}{32!}[2(30)29+240(16)15+42(25)24+30(26)25] = 3.50$$

usw., so dass wir schließlich alle Resultate von Tabelle 2.10 erhalten. Ob sich nun die beobachteten Häufigkeiten von den berechneten (theoretischen) unterscheiden, testen wir mit Hilfe des üblichen Chiquadrat-Tests und erhalten $X^2$ = 11.05. Die Zahl der Freiheitsgrade ist gleich der Zahl der verglichenen Klassen – Zahl der Parameter – 1, d.h. hier FG = 7 – 5 – 1 = 1. Ein Chiquadrat von 11.05

mit einem Freiheitsgrad ergibt $P = 0.0009$. Daran sehen wir, dass die Anordnung der Distanzen der empirischen Muster sich von der rein zufälligen Anordnung der Distanzen im Zörnigschen Modell signifikant unterscheidet. Sie ist daher keineswegs rein zufällig, sondern enthält unbekannte Konfigurationen, über die man weitere Hypothesen aufstellen kann. Vorläufig reicht uns die rein „optische" Feststellung, dass die Distanz 0, d.h. die unmittelbare Nachbarschaft, und die Distanz 4 zu Identität tendieren. Auch diese Hypothesen lassen sich mit Zörnigs Resultaten testen, wir verzichten hier darauf, weil die Daten einen recht kleinen Umfang haben.

Tabelle 2.10
Verteilung der Distanzen

| Distanz *x* | Beobachtet | Berechnet |
|---|---|---|
| 0 | 14 | 9.81 |
| 1 | 2 | 5.78 |
| 2 | 2 | 3.50 |
| 3 | 2 | 2.20 |
| 4 | 4 | 1.46 |
| 5 | 2 | 1.02 |
| $\geq 6$ | 1 | 3.23 |

## 2.8. Abhängigkeit der Musterfolgen

Die gleichen Daten wie in Tabelle 2.8 können wir dazu benutzen, um zu testen, ob in der Folge von Mustern Abhängigkeiten bestehen. Jedoch sind die Zahlen in der Tabelle so klein, dass der Test stark verzerrt wäre. Wir können die Daten aber dichotomisieren, indem wir „lange" und „kurze" Verse definieren. Die mittlere Länge der Muster ist 5.625, was man leicht aus Tabelle 2.6 errechnet. Daraus folgt, dass die Typen *a* und *b* mit jeweils 4 und 5 unbetonten Silben kurz (*K*), die Typen *c, d, e* hingegen „lang" (*L*) sind. Kodiert man Schema (III) oder Schema (IV) um, so erhält man

(VI) KKKK LKKK KKKL LLKK KLLL LLLK LLLK LLKK.

Man kann selbstverständlich zwei beliebige andere Symbole nehmen. Ein Teil der Information geht dabei natürlich verloren, aber möglicherweise geben sich neue Aspekte zu erkennen. Für die Auswertung einer derartigen binären Sequenz gibt es eine schier unendliche Menge von Methoden (vgl. z.B. Cox 1958; Gottman, Roy 1990; Bortz, Lienert, Boehnke 1990), hier werden wir von ihnen nur die einfachsten zeigen.

Wir tragen die Übergänge zwischen den beiden Buchstaben (*K* und *L*) in eine Vierfeldertafel mit Bezeichnungen ein

| | K | L |
|---|---|---|
| K | $n_{11}$ | $n_{12}$ |
| L | $n_{21}$ | $n_{22}$ |

d.h., $n_{11}$ ist die Zahl der Übergänge von *K* auf *K*, $n_{12}$ die der Übergänge von *K* auf *L* usw. Für unsere Daten in Schema (VI) erhalten wir das Resultat von Tabelle 2.11.

Tabelle 2.11
Übergänge zwischen kurzen und langen Versen

| | K | L |
|---|---|---|
| K | 12 | 4 |
| L | 5 | 10 |

Als *n* bezeichnen wir die Zahl aller Übergänge, die um 1 kleiner ist als die Zahl der Buchstaben, hier $n = 31$, was sich aus der Summe aller Zahlen in der Tabelle ergibt. Um zu überprüfen, ob es irgendeine Abhängigkeit zwischen einem Vorgängerbuchstaben und einem Nachfolgerbuchstaben gibt, setzten wir die Zahlen in Formel

$$\text{(2.15)}\quad \chi^2 = \frac{n\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2}\right)^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})}$$

ein und erhalten

$$\chi^2 = \frac{31\left(|12(10) - 4(5)| - \frac{31}{2}\right)^2}{16(17)15(14)} = \frac{221347.75}{57120} = 3.88.$$

Diese Größe ist wie ein Chiquadrat mit 1 Freiheitsgrad verteilt. Da unser berechnete Wert etwas größer ist als der kritische Wert 3.84 ($= \chi^2_{1(0.05)}$), nehmen wir an, dass es in der obigen Sequenz (VI) eine Tendenz gibt, kurze Muster nach kurzen und lange nach langen zu platzieren.

Alternativ kann man hier auch den Iterationstest für zwei unterschiedliche Kategorien benutzen, indem man wieder den asymptotischen Test verwendet, nämlich

(2.16) $$u = \frac{r - E(r)}{\sigma_r},$$

mit

(2.17) $$E(r) = 1 + \frac{2k_1k_2}{n}$$

und

(2.18) $$\sigma_r = \sqrt{\frac{2k_1k_2(2k_1k_2 - n)}{n^2(n-1)}}.$$

Hier bedeutet $k_1$ die Zahl der kurzen (*K*) Muster, $k_2$ die Zahl der langen (*L*) Muster, $n$ ist die Zahl aller Muster und $r$ die Zahl der Iterationen. Im obigen Fall bekommen wir aus Schema (VI) durch einfaches Auszählen

$k_1 = 17$
$k_2 = 15$
$n = 32$
$r = 11.$

Setzen wir diese Zahlen in (2.17) und (2.18) ein, so erhalten wir

$$E(r) = 1 + \frac{2(17)15}{32} = 16.9373,$$

$$\sigma_r = \sqrt{\frac{2(17)15[2(17)15 - 32]}{32^2(31)}} = 2.7712,$$

woraus sich (2.16) als

$$u = \frac{11 - 16.9373}{2.7712} = -2.14$$

ergibt. Diese Zahl ist viel kleiner als –1.96 und zeigt, dass die Zahl der Iterationen signifikant klein ist, d.h., kurze und lange Muster „klumpen sich“, was auch schon der vorherige Test etwas undeutlicher gezeigt hat. Umfangreiche Untersuchungen mit Hilfe der Iterationstheorie wurden im Bereich der Texte von Grotjahn (1980) durchgeführt.

Einen anderen Test schlägt Cox (1958) vor (vgl. auch Maxwell 1961: 137; Bortz, Lienert, Boehnke 1990: 563), wobei man ein Analogon von Tabelle 2.11 aufstellt, jedoch mit folgenden Definitionen:

$n_{11}$ = Zahl der Iterationen von kurzen Mustern = 6
$k_1$ = Zahl der kurzen Muster = 17
$n_{12} = k_1 - n_{11} = 17 - 6 = 11$
$n_{21} = n - k_1 - n_{11} + 1 = 32 - 17 - 6 + 1 = 10$
$n_{22} = n_{11} - 1 = 6 - 1 = 5.$

So erhalten wir die Daten in Tabelle 2.12.

Tabelle 2.12
Daten für den Cox-Test

| | | |
|---|---|---|
| 6 | 11 | 17 |
| 10 | 5 | 15 |
| 16 | 16 | 32 |

Die Wahrscheinlichkeit dieses Ereignisses kann man mit Hilfe der hypergeometrischen Verteilung als

$$(2.19)\quad P(n_{11}) = \frac{\binom{n_{11}+n_{21}}{n_{11}}\binom{n_{12}+n_{22}}{n_{12}}}{\binom{n}{n_{11}+n_{12}}} =$$

$$= \frac{(n_{11}+n_{21})!(n_{12}+n_{22})!(n_{11}+n_{12})!(n_{21}+n_{22})!}{n_{11}!n_{21}!n_{12}!n_{22}!n!}$$

berechnen. In unserem Fall erhalten wir

$$P(n_{11}) = \frac{16!16!17!15!}{6!11!10!5!32!} = 0.0618,$$

was größer ist als 0.05 und zeigt, dass die Übergänge an der Grenze der Zufälligkeit liegen (eher zufällig als einen Trend enthaltend).

Die Interpretation der Resultate in diesem Abschnitt ist nicht immer leicht. Der Testausgang bedeutet nicht, das wir die Wahrheit gefunden hätten, sondern gestattet uns, eine bestimmte vorläufige Entscheidung zu treffen. Besonders solche Werte, die an der Grenze der Signifikanz liegen, sind nicht sehr maßgebend;

denn die Signifikanzgrenze setzen wir selbst fest, sie findet sich nicht in den Daten. Zwar kann man die Wahrscheinlichkeit berechnen, mit der man einen Fehler begeht, wenn man eine bestimmte Entscheidung trifft, aber bei Daten von kleinem Umfang gerät man auch hier in Schwierigkeiten. Die Signifikanzgrenze ist nur ein Erfahrungswert, der sich in vielen Wissenschaften bewährt hat, ist aber nicht verbindlich für die Textanalyse. Mit ihrer Festsetzung tastet man sich nur langsam an das Verhalten von Textdaten.

## 2.9. Phasen

Eine Phase ist eine ununterbrochen wachsende oder eine ununterbrochen fallende Folge von Messwerten oder auch Rangzahlen. In einer Sequenz wie (IV) aus Abschnitt 2.1, die wir nochmals wiedergeben

(IV) 5555 6554 4556 7655 5677 7556 8775 6655,

kann man mehrere Phasen erkennen. Wenn man gleiche Längen unberücksichtigt lässt, d.h. als jeweils eine Einheit betrachtet, dann kann man sie als

(VII) 565456765676587565

kodieren. Symbolisiert man den Übergang zu einem höheren Wert als „+“und zu einem niedrigeren als „–“, dann erhält man die Sequenz

(VIII) + – – + + + – – + + – – + – – + –

Man kann nun untersuchen, ob die Länge der Phasen oder deren Anzahl zufällig ist. Man kann nämlich vermuten, dass der Dichter im Laufe des Schreibens Veränderungen im Rhythmus „vornimmt“ bzw. dass er zu bestimmten Phasen durch nachträgliche Veränderungen kommt. Im Folgenden werden wir zwei Aspekte der Phasen untersuchen.

### 2.9.1. Längen

In Sequenz (VIII) lässt man die erste und die letzte Phase aus und sieht dann, dass es hier

2 Phasen der Länge 1
5 Phasen der Länge 2
1 Phase der Länge 3

gibt. Insgesamt gibt es $n = 17$ Elemente (+ und –). Wir bezeichnen die Länge als $d$ ($d$ = 1,2,3). Um zu testen, ob die Anzahl der Phasen der Länge $d$ zufällig ist, führt man den Phasenverteilungstest (*runs-up-and-down*) von Wallis and Moore (1941) durch, indem man die erwarteten Anzahlen $e_d$ mit den beobachteten Anzahlen $o_d$ mit Hilfe eines Chiquadrat-Tests vergleicht.

Die erwartete (theoretische) Zahl der Phasen der Länge $d$ kann man als (s. Bortz, Lienert, Boehnke 1990: 572)

$$(2.20)\quad e_d = \frac{2(d^2+3d+1)(n-d-2)}{(d+3)!}$$

berechnen. Weiter ist

$$(2.21)\quad e_{d\geq 1} = 2\sum_{d=1}^{n-3}\frac{(d^2+3d+1)(n-d-2)}{(d+3)!} = 2\left(\frac{2n-7}{6}+\frac{1}{n!}\right),$$

was man zuletzt braucht. In unserem Fall haben wir

$$e_1 = \frac{2(1^2+3+1)(17-1-2)}{(1+3)!} = 5.83$$

$$e_2 = \frac{2(2^2+3(2)+1)(17-1-2)}{(2+3)!} = 2.38.$$

Um den erwarteten Wert für $d \geq 3$ zu erhalten, berechnen wir zuerst

$$e_{d\geq 1} = 2\left(\frac{2(17)-7}{6}+\frac{1}{17!}\right) = 9$$

und den erwarteten Wert für $d \geq 3$ erhalten wir einfach dadurch, dass wir von $e_{d\geq 1}$ die Werte von $e_1$ und $e_2$ subtrahieren, d.h.

$$e_{d\geq 3} = 9 - 5.83 - 2.38 = 0.79.$$

Wir fassen die Resultate in Tabelle 2.13 zusammen.

Tabelle 2.13
Verteilung der Phasenlängen

| d | $o_d$ | $e_d$ |
|---|---|---|
| 1 | 2 | 5.83 |
| 2 | 5 | 2.38 |
| ≥3 | 1 | 0.79 |

Den Vergleich der beobachteten und der theoretischen Werte führt man mit Hilfe des $\chi^2_p$ als

$$(2.22)\quad \chi^2_p = \sum_d \frac{(o_d - e_d)^2}{e_d}.$$

durch. In unserem Fall ist die Berechnung der Chiquadrat-Wertes etwas heikel, weil ein theoretischer Wert kleiner als 1 ist. Wir führen sie cum grano salis trotzdem durch und erhalten

$$\chi^2_p = \frac{(2-5.83)^2}{5.83} + \frac{(5-2.38)^2}{2.38} + \frac{(1-0.79)^2}{0.79} = 5.456.$$

Die Entscheidung wird in diesem Fall folgendermaßen herbeigeführt:

> Falls $\chi^2_p < 6.3$, dann berechnet man $\chi^2 = (6/7)\,\chi^2_p$ mit 2 Freiheitsgraden;
> sonst ist $\chi^2_p = \chi^2$ mit 2.5 Freiheitsgraden.

In unserem Fall rechnen wir also $\chi^2 = (6/7)5.456 = 4.68$, und dies ist mit zwei Freiheitsgraden nicht signifikant, da $\chi^2_{2(0.05)} = 5.99$. Das bedeutet, dass die Phasenlängen keinen evidenten Trend aufweisen. Einen zuverlässigeren Schluss erlauben unsere Daten wegen ihrer Kürze nicht.

Dieser Test zeigt, ob es eine wellenartige Bewegung der Längen im Gedicht gibt.

### 2.9.2. Häufigkeit

Bei der Beurteilung der Häufigkeit von Phasen zählt man auch die beiden Randphasen mit, was wir in Abschnitt 2.9.1 nicht getan haben. Um zu testen, ob die Zahl der Phasen signifikant groß oder klein ist, kann man bis zu $n = 25$ Beobachtungen die Tabellen von Eddington (1961) (s. Bortz, Lienert, Boehnke 1990: 771; – auf S. 772 ist *n* falsch angegeben) verwenden, was auch deshalb empfehlenswert ist, weil man hier die Wahrscheinlichkeiten rekursiv berechnen muss. In unserem Fall haben wir in der Sequenz (VIII) $n = 17$ Vorzeichen (Symbole) und darunter 10 Phasen (Iterationen). Die Wahrscheinlichkeit, dass man bei 17 Beobachtungen 10 oder weniger Phasen findet, ist nach der Tabelle $P(p \leq 10|n = 17) = 0.3770$. Da diese Zahl größer als 0.05 ist, schließen wir, dass es hier keine Tendenz gibt, die Phasen zu verlängern. Mit anderen Worten, man findet im „Erlkönig“ keine Phasentendenz in diesem Sinne.

Ist $n > 25$, so verfährt man wieder asymptotisch und berechnet die normalverteilte Größe mit Stetigkeitskorrektur als

(2.23) $$u = \frac{|\, p - E(p)\,| - 0.5}{\sigma_p},$$

wo

(2.24) $$E(p) = \frac{2n-1}{3}$$

und

(2.25) $$\sigma_p = \sqrt{\frac{16n-29}{90}}\,.$$

Diesen Test kann man nur bei längeren Texten anwenden. Würde man ihn für unsere Daten anwenden, dann bekäme man mit $n$ = 17 und $p$ = 10 zuerst $E(p)$ = (2(17)-1)/3 = 11, $\sigma_p$ = {[16(17)-29]/90}$^{1/2}$ = 1.6432, und schließlich $u$ = [|10-11|-0.5]/1.6432 = 0.3042. Dieser Wert ist bei zweiseitigem Test viel kleiner als der kritische Wert ($u$ = 1.96), daher kann man auch hier die Sequenz als zufällig betrachten.

## 2.10. Iterationslängentest

In Schema (VI) sieht man, dass die einzelnen Iterationen nicht gleich lang sind. Nach einer Iteration von 4 K kommt eine mit 6 K usw. Man kann sich fragen, ob eine bestimmte oder eine noch größere Länge nichtzufällig erscheint, d.h., ob es eine „Absicht“ gab, eine bestimmte Länge zu bevorzugen. Diese Hypothese kann man mit dem Test von Mood (1940) überprüfen.

In Schema (VI) betrachten wir die Iterationen von kurzen Elementen (K) und finden folgende Ausgangsdaten:

| | | |
|---|---|---|
| $n$ | = 32 | (Zahl aller Elemente) |
| $k_1$ | = 17 | (Zahl der K-Elemente) |
| $k_2$ | = 15 | (Zahl der L-Elemente) |
| $s$ | = 6 | (die längste Iteration von K). |

Die Überschreitungswahrscheinlichkeit für eine bestimmte Länge $s$ erhalten wir nach der Formel von Bradley (1968: 256) als

$$(2.26)\quad P(s) = \frac{\binom{k_2+1}{1}\binom{n-s}{k_2} - \binom{k_2+1}{2}\binom{n-2s}{k_2} + \binom{k_2+1}{3}\binom{n-3s}{k_2} - \ldots}{\binom{n}{k_2}}.$$

Setzen wir in diese Formel unsere Daten ein, so erhalten wir

$$P(s) = \frac{\binom{15+1}{1}\binom{32-6}{15} - \binom{16}{2}\binom{32-2(6)}{15}}{\binom{32}{5}} = 0.2152,$$

d.h., das Erscheinen einer Iteration mit der Mindestlänge $s$ = 6 ist rein zufällig, und es besteht kein Grund, hier eine Tendenz zu vermuten. In der obigen Formel fallen alle weiteren Glieder aus, weil

$$\binom{32-3(6)}{15} = \binom{14}{15} = 0.$$

Wenn $n$ > 30, so kann man den Test auch asymptotisch durchführen, und zwar mit Hilfe der Poisson-Verteilung als

$$(2.27)\quad P(s) = 1 - e^{-\lambda},$$

wo λ der Parameter der Poisson-Verteilung ist, den man als

$$(2.28)\quad \lambda = k_2\left(\frac{k_1}{n}\right)^s$$

berechnet. In unserem Fall erhalten wir

$$\lambda = 15(17/32)^6 = 0.3372,$$

woraus sich dann

$$P(6) = 1 - 2.7183^{-\,0.3372} = 0.2862$$

ergibt. Dieser Wert ist noch etwas größer als das obige Resultat, d.h., eine Sequenz der Länge 6 ist auch aufgrund dieses Tests rein zufällig.

## 2.11. Klimax im Gedicht

Wir haben festgestellt, dass es im Vers eine Art Klimax gibt, d.h., die Zahl der unbetonten Silben wächst vom Anfang bis zum Ende des Verses an. Es ergibt sich automatisch die Frage, ob eine derartige Tendenz im Rahmen des ganzen Gedichts in dem Sinne festzustellen ist, dass die Musterlänge vom Anfang bis zum Ende des Gedichts anwächst. Ein lineares Anwachsen kann man sich kaum vorstellen, es gibt daher nur die Möglichkeit, eine speziell ausgeprägte Tendenz zu entdecken. Zu diesem Zweck muss man eine Reihe von Tests durchführen, um möglichst viele Aspekte zu beleuchten. Im folgenden sind mehrere Möglichkeiten angegeben, die sich weiter variieren lassen.

### 2.11.1. Der U-Test

Betrachten wir wieder die Folge (IV), die wir bequemlichkeitshalber hier noch einmal aufführen

(IV) 5555 6554 4556 7655 5677 7556 8775 6655.

Sie repräsentiert die Verslängen, gemessen als Zahl der unbetonten Silben. Berechnet man den Durchschnitt aller dieser Zahlen in der Sequenz, so bekommt man den Wert 5.656. Nun kann man wieder die Sequenz dichotomisieren, indem man die Längen, die kleiner sind als 5.656, als *K* (kurz), und diejenigen, die länger sind als 5.656, als *L* (lang) bezeichnet. Dann bekommt man eine Sequenz mit zugeordneten Positionszahlen als

| K | K | K | K | L | K | K | K | K | K | K | L | L | L | K | K | K | L | L | L | L | L | L | K | L | L | L | K | L | L | K | K |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |

Um eine mögliche Klimax zu testen, stellen wir die Hypothese auf, dass die Rangzahlen der *K*-Zeichen eine signifikant kleinere Summe ($S_K$) ergeben als diejenige der *L*-Zeichen ($S_L$). Wir haben hier

$S_K = 1 + 2 + 3 + 4 + 6 + 7 + 8 + 9 + 10 + 11 + 15 + 16 + 17 + 24 + 28 + 31 + 32$
$= 224$

$S_L = 5 + 12 + 13 + 14 + 18 + 19 + 20 + 21 + 22 + 23 + 25 + 26 + 27 + 29 + 30$
$= 304$

und die Anzahl von kurzen Mustern beträgt $k_1 = 17$, die von langen $k_2 = 15$ und $n = k_1 + k_2 = 32$.

Den Test führen wir mit Hilfe des Mann-Whitneys (1947) U-Tests durch, indem wir das Kriterium

$$(2.29)\quad U = k_1 k_2 + \frac{k_1(k_1+1)}{2} - S_K$$

$$U' = k_1 k_2 + \frac{k_2(k_2+1)}{2} - S_L$$

berechnen. In unserem Fall erhalten wir

$U = 17(15) + 17(18)/2 – 224 = 184$
$U´ = 17(15) + 15(16)/2 – 304 = 71.$

Den kleineren der $U$, $U´$-Werte benutzen wir als Kriterium, und in den entsprechenden Tabellen (s. Bortz, Lienert, Boehnke 1990: 669, Tafel 6) finden wir, dass der kritische Wert für den $U´$-Wert auf der $\alpha = 0.05$-Ebene bei zweiseitigem Test 75 ist. Da unser beobachteter kleinerer Wert $U´$ den kritischen Wert unterschreitet, können wir vorläufig annehmen, dass es eine Gedichtsklimax gibt. Die Rangzahlen der langen Verse sind von denen der kurzen Verse signifikant unterschiedlich, hier signifikant größer.

Für $k_1, k_2 > 20$ benutzt man wieder den Normaltest nach

$$(2.30)\quad u = \frac{|U - E(U)| - 0.5}{\sigma_U}$$

mit

$$(2.31)\quad E(U) = \frac{k_1 k_2}{2}$$

und

$$(2.32)\quad \sigma_U = \sqrt{\frac{k_1 k_2 (k_1 + k_2 + 1)}{2}}\,.$$

Da wir hier die Information auf zwei Kategorien reduziert haben, ist dieser Test etwas grob, lässt sich aber schnell durchführen und gibt zumindest einen Hinweis, ob es sich lohnt, in dieser Richtung weiter zu forschen.

### 2.11.2. Der Rangkorrelationstest

Betrachten wir nun die Korrelation zwischen der Position und dem Rang, der einem Muster nach seiner Länge zugeordnet wird. Da es im „Erlkönig“ mehrere Muster gleicher Länge gibt, wird einer Gruppe gleichlanger Muster ihr durchschnittlicher Rang zugeordnet. So haben wir in Schema (IV) zwei Muster der Länge 4, die auf den Rängen 1 und 2 stehen würden. Sie erhalten beide den mittleren Rang 1.5. Auf den Rängen 3 bis 17 stehen 15 Muster der Länge 5, die alle den mittleren Rang 10 erhalten, usw. So bekommen wir die beiden Rangierungen, die in Tabelle 2.14 dargestellt sind. In der ersten Spalte steht die Positionsnummer im Gedicht, in der zweiten Spalte der dieser Position nach der Länge zugeordnete Rang, in der dritten Spalte der Unterschied $d_i$ der beiden Rangierungen, in der vierten stehen die Quadrate der Unterschiede, $d_i^2$ . Die fünfte Spalte benötigen wir für den nächsten Test.

Tabelle 2.14
Rangkorrelationen

| Position $p$ | Längenrang $l$ | $d_i$ | $d_i^2$ | $R(p_i)R(l_i)$ |
|---|---|---|---|---|
| 1 | 10 | -9 | 81 | 10 |
| 2 | 10 | -8 | 64 | 20 |
| 3 | 10 | -7 | 49 | 30 |
| 4 | 10 | -6 | 36 | 40 |
| 5 | 21.5 | -16.5 | 272.25 | 107.5 |
| 6 | 10 | -4 | 16 | 60 |
| 7 | 10 | -3 | 9 | 70 |
| 8 | 1.5 | 6.5 | 42.25 | 12 |
| 9 | 1.5 | 7.5 | 56.25 | 13.5 |
| 10 | 10 | 0 | 0 | 100 |
| 11 | 10 | 1 | 1 | 110 |
| 12 | 21.5 | -9,0 | 90.25 | 258 |
| 13 | 28.5 | -15.5 | 240.25 | 370.5 |
| 14 | 21.5 | -7.5 | 56.25 | 301 |
| 15 | 10 | 5 | 25 | 150 |
| 16 | 10 | 6 | 36 | 160 |
| 17 | 10 | 7 | 49 | 170 |
| 18 | 21.5 | -3.5 | 12.25 | 387 |
| 19 | 28.5 | -9.5 | 90.25 | 541.5 |
| 20 | 28.5 | -8.5 | 72.25 | 570 |
| 21 | 28.5 | -7.5 | 56.25 | 598.5 |
| 22 | 21.5 | 0.5 | 0.25 | 473 |
| 23 | 21,5 | 1.5 | 2.25 | 494.5 |

| | | | | |
|---|---|---|---|---|
| 24 | 10 | 14 | 196 | 240 |
| 25 | 32 | -7 | 49 | 800 |
| 26 | 28.5 | -2.5 | 6.25 | 741 |
| 27 | 28.5 | -1.5 | 2.25 | 769.5 |
| 28 | 10 | 18 | 256 | 280 |
| 29 | 21.5 | 7.5 | 56.25 | 623.5 |
| 30 | 21.5 | 8.5 | 72.25 | 645 |
| 31 | 10 | 11 | 121 | 310 |
| 32 | 10 | 12 | 144 | 320 |
| | | | $\sum d_i = 2179$ | $\sum R(p_i)R(l_i) = 9776$ |

Beim Testen müssen wir der Tatsache Rechnung tragen, dass es bei den Längenrängen Rangbindungen gibt, d.h. Zuordnungen von gleichem Rang an mehrere Positionen gleichzeitig. Um dies zu berücksichtigen, berechnen wir die Größe

$$(2.32)\quad T = \sum_{i=1}^{c} (t_i^3 - t_i)/12$$

wo $t_i$ die Anzahl der rangmäßig gleichgestellten Elemente ist, deren Kategorien von 1 bis c laufen. Wir haben Rang 1.5 zweimal, Rang 10 fünfzehnmal, Rang 21.5 achtmal und Rang 28.5 sechsmal, daher ist (für diese $c = 4$ Kategorien)

$$T = [(2^3 - 2) + (15^3 - 15) + (8^3 - 8) + (6^3 - 6)]/12 = 340.$$

Mit $n = 32$ setzen wir diese Zahlen in die Formel

$$(2.33)\quad \rho = \frac{2\left(\frac{n^3 - n}{12}\right) - T - \sum_{i=1}^{n} d_i^2}{2\sqrt{\left(\frac{n^3 - n}{12} - T\right)\left(\frac{n^3 - n}{12}\right)}},$$

mit der wir den Rangkorrelationskoeffizienten ρ berechnen, und erhalten

$$\rho = \frac{2\left(\frac{32^3 - 32}{12}\right) - 340 - 2179}{2\sqrt{\left(\frac{32^3 - 32}{12} - 340\right)\left(\frac{32^3 - 32}{12}\right)}} = 0.58.$$

Um festzustellen, ob dieser Wert signifikant ist, transformieren wir $\rho$ in eine Normalvariable

(2.34) $u = \rho\sqrt{n-1}$

und erhalten

$u = 0.58\sqrt{31} = 3.22.$

Da $u = 3.22 > 1.96$, schließen wir, dass zwischen Position und Längenrang eine signifikante positive Korrelation besteht, was wir bereits schon aufgrund des Ergebnisses des U-Test „geahnt“ haben.

Eine andere Möglichkeit bietet Hájeks (1969: 119, 137) Variante, bei der man die Summe der Rangprodukte berechnet, die wir in der letzten Spalte von Tabelle 2.14 sehen, nämlich

(2.35) $S = \sum_{i=1}^{n} R(p_i)R(l_i),$

wo $p_i$ = Position $i$, $l_i$ = Länge $i$, $R$ = Rang. Diese Größe wird auf die Normalvariable transformiert, und zwar in der Form

(2.36) $u = \dfrac{S - E(S)}{\sigma_S},$

wo

(2.37) $E(S) = 0.25n(n+1)^2$

(2.38) $\sigma_S = \sqrt{\dfrac{1}{144(n-1)}(n^3 - n - T)(n^3 - n)}.$

Da wir bereits alle Zahlen berechnet haben, erhalten wir

$E(S) = 0.25(32)33^2 = 8712$

$\sigma_S = \sqrt{\dfrac{1}{144(31)}(32^3 - 32 - 340)(32^3 - 32)} = 487.41,$

woraus

$$u = \frac{9776 - 8712}{487.41} = 2.18$$

folgt. Auch in diesem Fall haben wir $u = 2.18 > 1.96$, d.h., wir vermuten einen positiven Zusammenhang zwischen der Position in der Sequenz und der Musterlänge. Dieser Zusammenhang ist sicherlich auch qualitativ interpretierbar, man muss aber eine Eigenschaft finden, die quantifizierbar ist. Es bietet sich sofort die „Spannung“ im Gedicht an, die möglicherweise mit der Länge korreliert, jedoch lässt sich Spannung nur durch Urteile von Versuchspersonen ermitteln, ein Verfahren, dass hier nicht praktiziert, sondern dem Leser überlassen wird.

### 2.11.3. Cox und Stuarts $S_1$-Test

Bei diesem Test teilt man die Folge in zwei Hälften und vergleicht jeweils diejenigen zwei Werte der beiden Teile, die vom Zentrum gleichweit entfernt sind. In unserem Fall ist die Mitte nach der vierten Strophe, da das Gedicht insgesamt 8 Strophen hat. Ein steigender Trend setzt voraus, dass die Werte der zweiten Hälfte größer sind als die der ersten. In unserem Fall haben wir die Sequenz

5555 6554 4556 7655 | 5677 7556 8775 6655,

in der wir $x_i$ mit $x_{n-i+1}$ vergleichen, wobei $n = 32$, z.B.

$x_1 = 5$ mit $x_{32-1+1} = x_{32} = 5$
$x_2 = 5$ mit $x_{32-2+1} = x_{31} = 5$
$x_3 = 5$ mit $x_{32-3+1} = x_{30} = 6$
usw.

Wir definieren die Funktion

$$(2.39)\quad h_i = \begin{cases} 1 & wenn\ x_{n-i+1} > x_i \\ 0 & wenn\ x_{n-i+1} \le x_i \end{cases}.$$

Wenn der Test zweiseitig ist, dann definieren wir

$$(2.40)\quad h_i' = \begin{cases} 1 & wenn\ x_{n-i+1} > x_i \\ 0 & wenn\ x_{n-i+1} < x_i \\ 0.5 & wenn\ x_{n-i+1} = x_i \end{cases}.$$

Anschließend berechnen wir die Größe

(2.41) $S_1 = \sum_{i=1}^{n/2}(n-2i+1)h_i$ bzw. $S_1' = \sum_{i=1}^{n/2}(n-2i+1)h_i'$.

Für unsere Daten sind alle diese Werte in Tabelle 2.15 zusammengefasst.

Tabelle 2.15
Ausgangszahlen für den $S_1$-Test

| $i$ | $n$-$i$+1 | $x_i$ | $x_{n-i+1}$ | $h_i$ | $h'_i$ | $(n$-$2i$+1$)h_i$ | $(n$-$2i$+1$)h'_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 32 | 5 | 5 | 0 | 0.5 | (32-2+1)(0) = 0 | 3(0.5) |
| 2 | 31 | 5 | 5 | 0 | 0.5 | (32-4+1)(0) = 0 | 29(0.5) |
| 3 | 30 | 5 | 6 | 1 | 1 | (32-6+1)(1) = 27 | 27 |
| 4 | 29 | 5 | 6 | 1 | 1 | 25(1) = 25 | 25 |
| 5 | 28 | 6 | 5 | 0 | 0 | 23(0) = 0 | 0 |
| 6 | 27 | 5 | 7 | 1 | 1 | 21(1) = 21 | 21 |
| 7 | 26 | 5 | 7 | 1 | 1 | 19(1) = 19 | 19 |
| 8 | 25 | 4 | 8 | 1 | 1 | 17(1) 17 | 17 |
| 9 | 24 | 4 | 5 | 1 | 1 | 15(1) = 15 | 15 |
| 10 | 23 | 5 | 6 | 1 | 1 | 13(1) = 13 | 13 |
| 11 | 22 | 5 | 6 | 1 | 1 | 11(11) = 11 | 11 |
| 12 | 21 | 6 | 7 | 1 | 1 | 9(1) = 9 | 9 |
| 13 | 20 | 7 | 7 | 0 | 0.5 | 7(0) = 0 | 3.5 |
| 14 | 19 | 6 | 7 | 1 | 1 | 5(1) = 5 | 5 |
| 15 | 18 | 5 | 6 | 1 | 1 | 3(1) = 3 | 3 |
| 16 | 17 | 5 | 5 | 0 | 0.5 | 1(0) = 0 | 0.5 |
| | | | | | | 165 | 199 |

Wir testen $S_1$ wieder mit Hilfe der Normalverteilung mit

(2.42) $E(S_1) = \frac{n^2}{8}$

(2.43) $\sigma_{S_1} = \frac{n(n^2-1)}{24}$

und benutzen das Kriterium

(2.44) $u = \frac{|S_1 - E(S_1)| - 0.5}{\sigma_{S_1}}$

für den zweiseitigen Test, den wir aber auch einseitig auswerten können. In unserem Fall bekommen wir

$$u = \frac{\left|165 - \frac{32^2}{8}\right| - 0.5}{\sqrt{32(32^2 - 1)/24}} = \frac{36.5}{36.93} = 0.98.$$

Auch beim einseitigen Test ist dieses Resultat nicht signifikant, d.h., signalisiert keinen Trend. Berechnet man aber das nicht konservative Kriterium $S'_1$, dann erhält man

$$u = \frac{|199 - 128| - 0.5}{36.93} = 1.91,$$

was einseitig signifikant ist. Da wir aufgrund vorheriger Tests bereits die Existenz eines Verlängerungstrends (Gedichtsklimax) zumindest vermuten durften, ist der einseitige Test berechtigt. Wie man aber sieht, liegen latente Zusammenhänge unter der Oberfläche der Dinge, werden durch Akzidentalien (Störungen, Korrekturen usw.) verschleiert und lassen sich nur mit Mühe ans Licht bringen.

### 2.11.4. Der $S_2$-Test

Dieser von Cox und Stuart (1955) vorgeschlagene Test beruht auf der Binomialverteilung. Man vergleicht dabei die parallelen Werte der beiden Hälften der Sequenz. Wenn es einen steigenden Trend gibt und man den Wert der ersten Hälfte von dem parallelen Wert der zweiten subtrahiert, dann wird die Zahl der positiven Differenzen (+ Vorzeichen) signifikant groß. In unserem Fall haben wir

| Zweite Hälfte | | Erste Hälfte | | Differenz |
|---|---|---|---|---|
| 5 | - | 5 | = | 0 |
| 6 | - | 5 | = | + |
| 7 | - | 5 | = | + |
| 7 | - | 5 | = | + |
| 7 | - | 6 | = | + |
| 6 | - | 5 | = | + |
| 6 | - | 5 | = | + |
| 5 | - | 4 | = | + |
| 8 | - | 4 | = | + |
| 7 | - | 5 | = | + |
| 5 | - | 6 | = | - |
| 6 | - | 7 | = | - |
| 5 | - | 5 | = | 0 |
| 5 | - | 5 | = | 0 |

Es ergeben sich 9 „plus“-Vorzeichen. Da die Wahrscheinlichkeit eines positiven Vorzeichens unter der Nullhypothese gleich 0.5 ist, ergibt sich die Wahrscheinlichkeit, dass man bei $n = 16$ Fällen $x = 9$ oder mehr „plus“ Vorzeihen findet, als

$$P(X \geq 9) = \sum_{x=9}^{16} \binom{16}{x} \frac{1}{2^{16}} = 1 - \sum_{x=0}^{8} \binom{16}{x} \frac{1}{2^{16}}.$$

In unserem Fall beträgt diese Wahrscheinlichkeit

$$1 - \frac{1}{2^{16}}(1 + 16 + 120 + 560 + 1820 + 4368 + 8008 + 11440 + 12870) = 0.41,$$

d.h., die Zahl der positiven Differenzen ist nicht signifikant. Betrachtet man jedoch auch 0 als ein dem Trend nicht widersprechendes Anzeichen, dann hat man 13 nichtnegative Vorzeichen, und $P(X \geq 13) = 0.003$, was einen nichtfallenden Trend andeutet.

### 2.11.5. Bortz – Lienert – Boehnkes Verfahren

Die Hypothese eines monotonen Trends kann man auch mit Hilfe einer Vierfeldertafel testen, wie es die oben genannten Autoren (1990: 586) vorschlagen. Trägt man die einzelnen Verslängen positionsgetreu in ein Koordinatensystem ein, so erhält man die Abbildung 2.8, in der die zwei Gedichtshälften durch eine vertikale Linie und die Punkte durch eine horizontale Linie, die den Mittelwert darstellt, getrennt wurden.

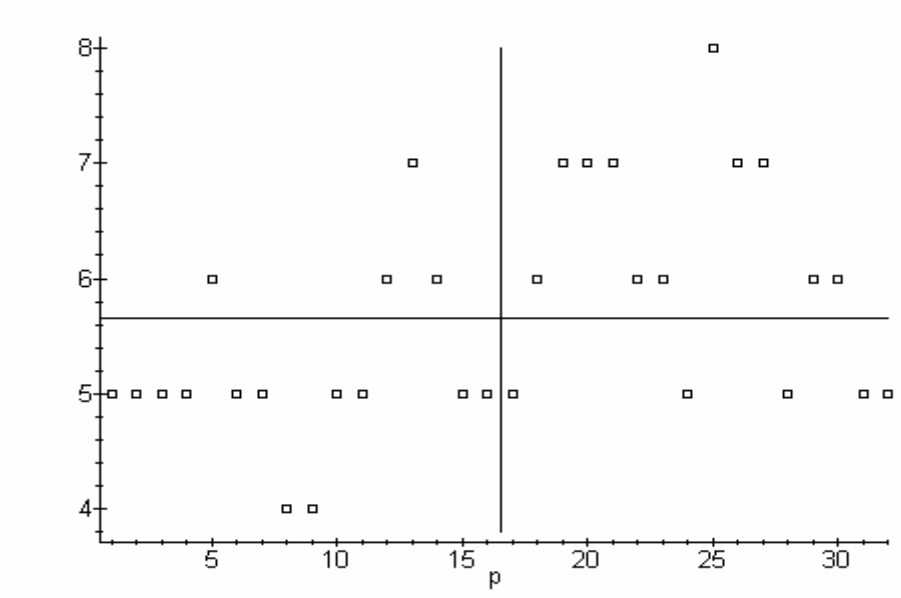


Abbildung 2.8. Musterlängen in den Hälften des Gedichts

Wie man sieht, liegen die Punkte der ersten Hälfte meistens unter dem Durchschnitt, die der zweiten darüber. Die beiden Linien teilen das ganze Feld in 4 Felder auf. Die Zahl der Punkte in den einzelnen Feldern tragen wir in eine Vierfeldertafel (s. Tabelle 2.16) ein

Tabelle 2.16

| | | |
|---|---|---|
| 4 | 10 | 14 |
| 12 | 6 | 18 |
| 16 | 16 | 32 |

und berechnen daraus den Chiquadrat-Wert als

$$X^2 = \frac{32(|4(6)-12(10)|-16)^2}{16(16)14(18)} = 3.17,$$

was bei 1 Freiheitsgrad der Existenz eines Trends widerspricht. Jedoch ergäbe sich ohne Korrektur für Kontinuität $X^2$ = 4.57, was signifikant wäre. Man sieht also, dass man sich hier gerade an der Signifikanzgrenze bewegt.

Noch einfacher kann man diese Hypothese testen, indem man den Mittelwert der ersten ($\overline{x}_1$) und den der zweiten ($\overline{x}_2$) Hälfte berechnet und sie nach Cochran (1954) in die Formel

$$(2.45)\quad u = (\overline{x}_1 - \overline{x}_2)\sqrt{\frac{n_1 n_2}{n}}$$

einsetzt. In unserem Fall ist der Durchschnitt der ersten 16 Muster 5.1875 und der der zweiten 6.0625, so dass

$$u = (5.1875 - 6.0625)\sqrt{16(16)/32} = -2.47\,.$$

Dieser Wert unterstützt die Annahme, dass es im Gedichtverlauf einen Trend gibt, den Vers in der zweiten Hälfte des Gedichts zu verlängern.

### 2.11.6. Tests für die Homogenität der Strophen

In Schema (IV) addieren wir jeweils 4 Zahlen, d.h., wir berechnen die Länge der einzelnen Strophen und erhalten

(IX) 20, 20, 20, 23, 25, 23, 27, 22.

Die Zahl der Strophen ist $k$ = 8, die Summe der Längen ist $N$ = 180, und die durchschnittliche Länge der Strophe ist $\overline{x} = N/k$ = 22.5. Die Homogenität kann man auf verschiedene Arten testen, hier zeigen wir den Chiquadrat-Test für Homogenität. Sind die Strophen längenmäßig homogen, dann ist die erwartete Länge gleich dem Durchschnitt, so dass man Fishers Dispersionstest anwenden kann

$$(2.46)\quad X^2 = \sum_{i=1}^{k} \frac{(x_i - \overline{x})^2}{\overline{x}} = \frac{k}{N} \sum_{i=1}^{k} x_i^2 - N,$$

der in unserem Fall

$$X^2 = (8/180)(\, 20^2 + 20^2 + 20^2 + 23^2 + 25^2 + 23^2 + 27^2 + 22^2) - 180 = 2.04$$

ergibt. Dieser Wert ist mit FG = $k - 1 = 8 - 1 = 7$ nicht signifikant, so dass wir die Homogenität akzeptieren können.

Eine andere, äquivalente Möglichkeit ergibt sich mit Hilfe der Informationsstatistik, in diesem Fall als

$$(2.47)\quad 2I = 2\sum_{i=1}^{k} x_i \ln \frac{x_i}{\overline{x}} = 2\sum_{i=1}^{k} x_i \ln x_i - 2N \ln \overline{x},$$

was in unserem Fall

$$2\mathrm{I} = 2[3(20)\, ln\ 20 + (2)23\, ln\ 23 + 25\, ln\ 25 + 27\, ln\ 27 + 22\, ln\ 22] - \\ - 2(180)\, ln\ 22.5 = 2.01$$

ergibt. Auch $2I$ ist wie ein Chiquadrat mit $k - 1$ Freiheitsgraden verteilt, und man sieht, dass die beiden Werte fast identisch sind. Da beide gleichwertig sind, reicht es wieder, wenn man eine der beiden Methoden anwendet.

### 2.11.7. Linearer Trend

Cochran (1954) hat einen Test für Linearität vorgeschlagen, den man nach

$$(2.48)\quad u = \frac{\sum_{i=1}^{k} (x_i - \overline{x}) \left( i - \frac{k+1}{2} \right)}{\sqrt{n(k^2 - 1)/12}}$$

berechnen kann. In unserem Fall, wo $i$ die Strophenposition bedeutet, $k$ = 8 und $\overline{x}$ = 22.5, erhalten wir für den Zähler

$u$ = (20 – 22.5)(1 – 92) + (20 – 22.5)(2 – 4.5) + (20 – 22.5)(3 – 4.5) +
+ (23 – 22.5)(4 – 4.5) + (23 – 22.5)(4 – 4.5) + (25 – 22.5)(5 – 4.5) +
+ (24 – 22.5)(6 – 4.5) + (27 – 22.5)(7 – 4.5) + (22 – 22.5)(8 – 4.5) =

= 8.75 + 6.25 + 3.75 – 0.25 + 1.25 + 2.25 + 11.5 – 1.75 = 31.75.

Der Nenner ergibt

$$\sqrt{180(8^2 - 1)/12} = 30.7408,$$

so dass

$$u = \frac{31.75}{30.7408} = 1.03,$$

was bedeutet, dass das lineare Anwachsen der Strophenlängen nicht nachgewiesen werden kann.

### 2.11.8. Sprünge im Rhythmus

In einem Gedicht kann es – aus welchen Gründen auch immer – geschehen, dass an einer Stelle eine plötzliche Veränderung des Rhythmus stattfindet. Aus der Musik ist dieses Phänomen hinreichend bekannt. Um aber von einem Sprung reden zu dürfen, muss man diesen erst nachweisen. In Schema (IX), Abschnitt 2.11.6 kann man solche Sprünge in der 7 oder in der 8. Strophe vermuten, d.h. am Ende des Gedichtes.

Um diese „Vermutung“ zu testen, kann man den Chiquadrat-Test für die gegebene Strophe durchführen, und zwar aufgrund eines Vergleichs mit allen vorangehenden Strophen. Die Formel lautet

$$(2.49)\quad X^2 = \frac{(x_1 + x_2 + ... + x_r - rx_{r+1})^2}{r(r+1)\overline{x}}.$$

In Strophe 7, wo sich ein Längensprung von 23 (der 6. Strophe) auf 27 befindet, erhalten wir

$$X^2 = \frac{[20 + 20 + 20 + 23 + 25 + 23 - 6(27)]^2}{6(7)22.5} = 1.02.$$

Da diese Zahl mit 1 Freiheitsgrad nicht signifikant groß ist, kann man die Hypothese eines Sprungs in dieser Strophe ablehnen.

Für die 8. Strophe ist der Sprung im Vergleich zum vorherigen Verlauf sehr gering, wir bekommen $X^2 = 0.02$, d.h., hier ist kein Sprung „nach unten“ zu sehen.

### 2.11.9. Spannung und Streuung

Man kann nicht nur direkt die Veränderung der Musterlänge in einem Gedicht, sondern auch gewisse Funktionen der Länge beobachten, denn der Inhalt oder die Spannung können sich beispielsweise auch in der Abwechslung der Längen, d.h. in deren Streuung widerspiegeln. Andere Funktionen sind ebenso gut denkbar, wie wir schon oben gesehen haben.

Um die Streuung zu untersuchen, beziehen wir uns wieder auf Schema (IV) und berechnen schrittweise den Längendurchschnitt von $j$ Versen ($j$ = 1,2,…,32) und die Streuung um diesen Durchschnitt, d.h.

$$(2.50)\quad \overline{x}_j = \frac{1}{j}\sum_{i=1}^{j} x_i$$

und

$$(2.51)\quad \sigma_j^2 = \frac{1}{j}\sum_{i=1}^{j}(x_i - \overline{x}_j)^2 = \frac{1}{j}\sum_{i=1}^{j} x_i^2 \; - \; \overline{x}_j^2 .$$

So ist z.B. für $j = 5$

$$\sum_{i=1}^{5} x_i = 5+5+5+5+6 = 26$$
$$\overline{x}_5 = 26/5 = 5.2$$

$$\sum_{i=1}^{5} x_i^2 = 5^2+5^2+5^2+5^2+6^2 = 136$$
$$\sigma_5^2 = 136/5 - 5.2^2 = 0.16 .$$

Die Berechnungen sind ausführlich in Tabelle 2.17 angegeben

Wie man in Abbildung 2.9 sieht, wächst die Varianz quasi linear bis zu einem Höhepunkt in der vorletzten Strophe, wonach sich die Spannung – wie aus inhaltlichen Interpretation bekannt – auflöst und auch die Varianz sinkt. Würde man die Spannung mit Hilfe von Versuchspersonen messen, so könnte man even-

tuell feststellen, dass sie mit der Musterlänge korreliert. Mit anderen Worten, die Spannung lädt sich nicht nur inhaltlich, sondern auch rhythmisch schrittweise auf, um sich dann auf beiden Ebenen aufzulösen. Man könnte auch von einer Oszillation ausgehen und eine Fourier-Reihe anpassen, in deren Koeffizienten sich ein inhaltlicher Aspekt widerspiegeln könnte. Sehr interessant wäre auch ein Vergleich mit Schuberts Vertonung dieses Gedichts, wo man das Anwachsen der Spannung mit anderen Mitteln herbeiführen muss. Das Verfahren würde es eventuell ermöglichen, Parallelitäten zwischen Musik und Poesie zu finden. Man erinnere sich an Schuberts Aussage, dass die Musik bereits in dem Text enthalten gewesen sei und er sie nur habe aufschreiben müssen. Auch die Erforschung der Synästhesie könnte von solchen Untersuchungen profitieren.

Hier genügt es uns zunächst, von einem linearen Zusammenhang auszugehen. In der Tat ergibt sich für die Daten in Tabelle 2.17 (letzte Spalte) eine Gerade:

$$\textit{Varianz}\ \sigma_i^2 = -0.0401 + 0.0351i \quad (i \text{ ist Position}),$$

und der Determinationskoeffizient $D = 0.92$ deutet an, dass die Gerade zunächst ausreicht.

Tabelle 2.17
Verlauf der Längenstreuung im „Erlkönig“

| $i$ | $x_i$ | $\sum_{k=1}^{i} x_k$ | $\overline{x}_i$ | $x_i^2$ | $\sum_{k=1}^{i} x_k^2$ | $\sigma_i^2$ |
|---|---|---|---|---|---|---|
| 1 | 5 | 5 | 5.00 | 25 | 25 | 0.00 |
| 2 | 5 | 10 | 5.00 | 25 | 50 | 0.00 |
| 3 | 5 | 15 | 5.00 | 25 | 75 | 0.00 |
| 4 | 5 | 20 | 5.00 | 25 | 100 | 0.00 |
| 5 | 6 | 26 | 5.20 | 36 | 136 | 0.16 |
| 6 | 5 | 31 | 5.17 | 25 | 161 | 0.10 |
| 7 | 5 | 36 | 5.14 | 25 | 186 | 0.15 |
| 8 | 4 | 40 | 5.00 | 16 | 202 | 0.25 |
| 9 | 4 | 44 | 4.89 | 16 | 218 | 0.31 |
| 10 | 5 | 49 | 4.90 | 25 | 243 | 0.29 |
| 11 | 5 | 54 | 4.91 | 25 | 268 | 0.26 |
| 12 | 6 | 60 | 5.00 | 36 | 304 | 0.33 |
| 13 | 7 | 67 | 5.15 | 49 | 353 | 0.63 |
| 14 | 6 | 73 | 5.21 | 36 | 389 | 0.64 |
| 15 | 5 | 78 | 5.20 | 25 | 414 | 0.56 |
| 16 | 5 | 83 | 5.19 | 25 | 439 | 0.50 |
| 17 | 5 | 88 | 5.18 | 25 | 464 | 0.46 |
| 18 | 6 | 94 | 5.22 | 36 | 500 | 0.53 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 19 | 7 | 101 | 5.32 | 49 | 549 | 0.59 |
| 20 | 7 | 108 | 5.40 | 49 | 598 | 0.74 |
| 21 | 7 | 115 | 5.48 | 49 | 647 | 0.78 |
| 22 | 5 | 120 | 5.45 | 25 | 672 | 0.84 |
| 23 | 5 | 125 | 5.43 | 25 | 697 | 0.82 |
| 24 | 6 | 131 | 5.46 | 36 | 733 | 0.73 |
| 25 | 8 | 139 | 5.56 | 64 | 797 | 0.97 |
| 26 | 7 | 146 | 5.62 | 49 | 846 | 0.95 |
| 27 | 7 | 153 | 5.67 | 49 | 895 | 1.00 |
| 28 | 5 | 158 | 5.64 | 25 | 920 | 1.05 |
| 29 | 6 | 164 | 5.66 | 36 | 956 | 0.93 |
| 30 | 6 | 170 | 5.67 | 36 | 992 | 0.92 |
| 31 | 5 | 175 | 5.65 | 25 | 1017 | 0.88 |
| 32 | 5 | 180 | 5.63 | 25 | 1042 | 0.87 |

Dass hier eher eine Kurve geeigneter wäre, ist offenbar, weil die Gerade in Position 1 einen negativen Wert annimmt, was in der Empirie unmöglich ist. Als erste Approximation reicht aber die Gerade.

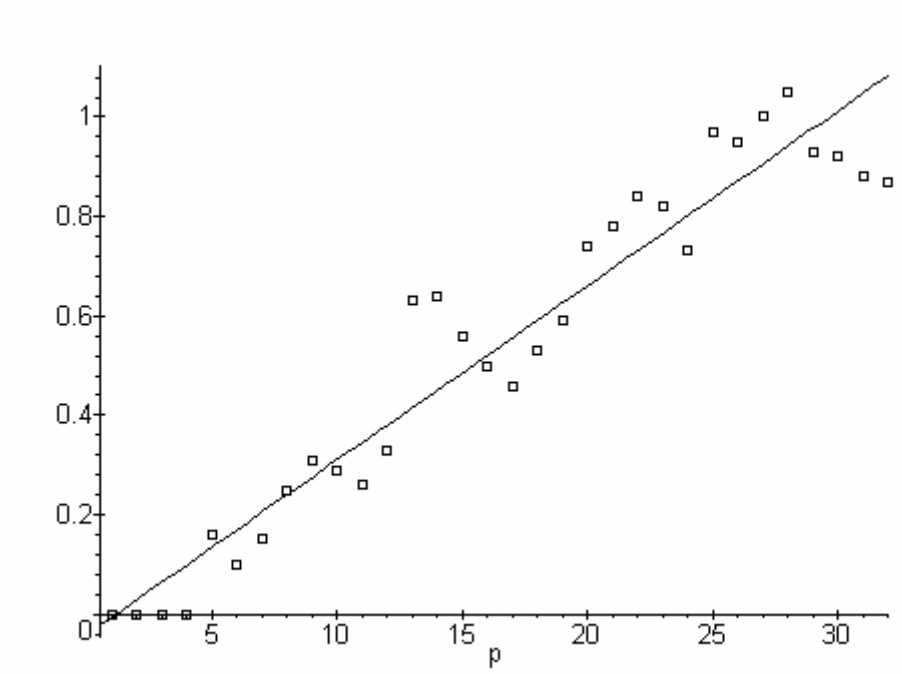


Abbildung 2.9. Verlauf der Längenvarianz im Gedicht

### 2.11.10. Hřebíčeks Verfahren

Offensichtlich sind unterschiedliche Funktionen der Länge bzw. anderer Eigenschaften auf verschiedene Weisen in den Verlauf des Gedichts eingebunden, jedoch oft nur latent. Das Problem liegt darin, messbare Variablen zu finden, die

die gegebene Eigenschaft (hier z.B. die Länge) zumindest latent begleiten, und sie manifest zu machen. In den vorigen Abschnitten haben wir gesehen, dass der einfache Durchschnitt nicht besonders markant mit dem Verlauf bzw. mit der Position korreliert, dafür aber die Varianz, die auch für den Leser des Gedichts kaum wahrnehmbar ist. Die Aufgabe der Textanalyse besteht gerade darin, auch latente Bewegungen und Strukturierungen in Text zu erfassen. Hřebíček (1993, 1995, 1997, 1997a, 2000) hat eine Reihe von derartigen heuristischen Möglichkeiten entwickelt und sie mit der Dynamik des Textes in Zusammenhang gebracht.

(a) ***Die R-Kurve***

Betrachten wir nochmals Tabelle 2.17, in deren vierten Spalte wir die Mittelwerte aus den ersten *i* Versen berechnet haben. Hřebíček definiert

(2.52) $r_i = \overline{x}_i - \overline{x}_{i-1}, \qquad i = 2,3,\ldots,k$

als die Unterschiede der benachbarten Mittelwerte. So ist z.B.

$r_5 = 5.20 - 5.00 = 0.20$
$r_6 = 5.17 - 5.20 = -0.03$
$r_7 = 5.14 - 5.17 = -0.04$
$r_8 = 5.00 - 5.14 = -0.14$

usw. Man findet alle $r_i$-Werte in Tabelle 2.18. Die kumulativen Summen dieser Differenzen ergeben *Hřebíčeks dynamische Charakteristik*

(2.53) $R_i = \sum_{j=2}^{i} r_j$

die man in der vierten Spalte von Tabelle 2.18 findet. Die $R_i$-Werte ergeben dann eine für den Rhythmus des Gedichtes charakteristische Kurve, die man in Abbildung 2.10 sehen kann. Hier ist die Nichtlinearität bereits offensichtlich, und ohne einen theoretischen Ansatz könnte man diese Bewegung nicht mehr erfassen. Im ersten Schritt reicht die Erkenntnis, dass die Strukturierung der Längen nicht so einfach ist, wie man vermutet.

Tabelle 2.18
Hřebíčeks dynamische Maße

| Vers Nr. | $\overline{x}_i$ | $r_i$ | $R_i$ |
|---|---|---|---|
| 1 | 5.00 | - | - |
| 2 | 5.00 | 0.00 | 0.00 |
| 3 | 5.00 | 0.00 | 0.00 |
| 4 | 5.00 | 0.00 | 0.00 |
| 5 | 5.20 | 0.20 | 0.20 |
| 6 | 5.17 | -0.03 | 0.17 |
| 7 | 5.14 | -0.03 | 0.14 |
| 8 | 5.00 | -0.14 | 0.00 |
| 9 | 4.89 | -0.11 | -0.11 |
| 10 | 4.90 | 0.01 | -0.10 |
| 11 | 4.91 | 0.01 | -0.09 |
| 12 | 5.00 | 0.09 | 0.00 |
| 13 | 5.15 | 0.15 | 0.15 |
| 14 | 5.21 | 0.06 | 0.21 |
| 15 | 5.20 | -0.01 | 0.20 |
| 16 | 5.19 | -0.01 | 0.19 |
| 17 | 5.18 | -0.01 | 0.18 |
| 18 | 5.22 | 0.04 | 0.22 |
| 19 | 5.32 | 0.10 | 0.32 |
| 20 | 5.40 | 0.08 | 0.40 |
| 21 | 5.48 | 0.08 | 0.48 |
| 22 | 5.45 | -0.03 | 0.45 |
| 23 | 5.43 | -0.02 | 0.43 |
| 24 | 6.46 | 0.03 | 0.46 |
| 25 | 5.56 | 0.10 | 0.56 |
| 26 | 5.62 | 0.06 | 0.62 |
| 27 | 5.67 | 0.05 | 0.67 |
| 28 | 5.64 | -0.03 | 0.64 |
| 29 | 5.66 | 0.02 | 0.66 |
| 30 | 5.67 | 0.01 | 0.67 |
| 31 | 5.65 | -0.02 | 0.65 |
| 32 | 5.63 | -0.02 | 0.63 |

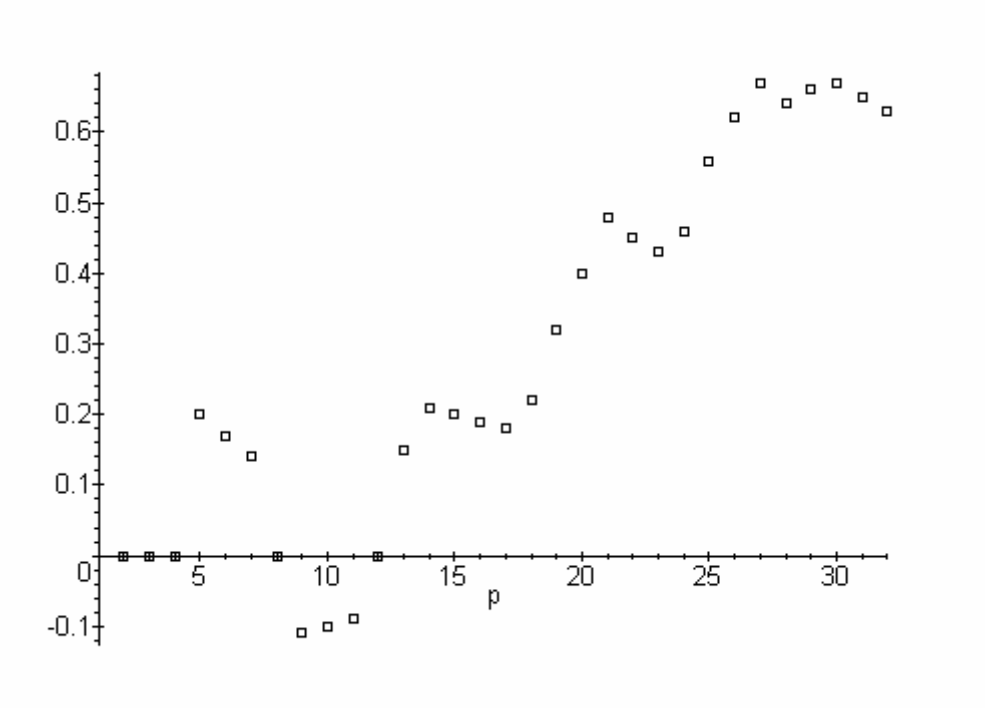


Abbildung 2.10. Verlauf der *R*-Funktion von Hřebíček

(b) **Die *S*-Kurve**

Auf ähnliche Weise kann man die Differenz zwischen dem sequentiellen Mittelwert $\overline{x}_i$ und dem Gesamtmittelwert $\overline{x}$ bilden als

(2.54) $D_i = \overline{x}_i - \overline{x}$

und die kumulativen Werte

(2.55) $S_i = \sum_{j=1}^{i} D_j$

als eine charakteristische Kurve betrachten. Vorläufig weiß man aber nicht, wie es in anderen Gedichten aussieht, womit ein bestimmter Verlauf korreliert und welche Kurve mit welchen nichtrhythmischen Größen in Zusammenhang steht. Es ist ein umfangreiches Programm für künftige Forschung. Hřebíček (1997: 132) hat die Vermutung aufgestellt, dass Längen (in seiner Untersuchung sind es Satzlängen) „…are affected by the non-explicit semantic order of a text so that these characteristics obtain some ambiguous shape: the results of testing vary on the margin of a respective test criterion“, was wir bei verschiedenen Tests bereits festgestellt haben, und er bemerkt, dass „…while in group of analyzed texts posi-

tive results are obtained, in their part positive results are hidden somewhere within the random behaviour of the system."

Daher ist die Suche nach konstanteren Indikatoren angezeigt, und Hřebíček verwendet zu diesem Zweck den sogenannten *Hurst´s adjusted rescaled range*, *R/S*, wo *S* die Standardabweichung der gesamten Sequenz ist, d.h. die Wurzel aus

$$(2.56)\quad S^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2$$

die man in der 32. Zeile der letzten Spalte von Tabelle 2.17, nämlich

$$S = (0.87)^{1/2} = 0.93,$$

und *R* in der letzten Zeile und der letzten Spalte von Tabelle 2.18 findet, so dass *R/S* = 0.63/0.93 = 0.68.

Mit Maßen dieser Art kann man zeigen, dass Texte und Naturphänomene in bestimmter Hinsicht ähnliches Verhalten an den Tag legen. Eine breit gefächerte Untersuchung findet man in Arbeiten von Hřebíček, wo zahlreiche neue Aspekte der Texte entdeckt worden sind.

## 2.12. Zusammenfassung

Rhythmus ist nicht nur eine Abwechslung von betonten und unbetonten Silben, sondern eine multidimensionale dynamische Bewegung aller formalen Bestandteile eines Textes. Oft – aber nicht unbedingt immer – begleitet diese Bewegung der formalen Bestandteile des Textes dessen immaterielle Komponenten, wie Inhalt, Spannung, emotionale Färbung, Bildhaftigkeit und viele andere, die man insbesondere in der Psycholinguistik bereits gefunden hat. Die in diesem Abschnitt dargestellten Methoden und Zusammenhänge sind weder die einzig möglichen noch immer die adäquatesten für das Auffinden von Zusammenhängen. Theoretische Ansätze sind äußerst schwierig, verlangen umfangreiche Überprüfung an vielen Texten und später wahrscheinlich etwas kompliziertere Mathematik.

Wie wir gesehen haben, kann schon die Darstellung des Verses unterschiedlich sein und aufgrund einer speziellen Darstellung bekommt man auch unterschiedliche Resultate. Die Verfasser haben nur einen beschränkten Überblick über die Gestaltung ihrer Texte, sie kontrollieren nur den Teil der Dynamik, der oberflächlich vorgeschrieben ist, z.B. „schreibe Verse mit vier Akzenten und Strophen mit vier Versen", alles andere ist Sache der Semantik, des Themas, des Wortschatzes, der Intention, der Wirkung, die erzielt werden soll usw.

Der fertige Text selbst ist zunächst ein geschriebener Text, der durch den Leser zu einem interpretierten Text wird (Hřebíček 1997), und dieser kann von Leser zu Leser unterschiedlich sein. Die Forschung ist nicht allzu sehr entwickelt, auch wenn man über Poetik bereits ganze Bibliotheken geschrieben hat. Will man einen Schritt tiefergehen, muss man zumindest zu elementaren Methoden der Statistik greifen.

# 3. Phonik

Nach Jakobson tritt in der Poesie vor allem die Form in den Vordergrund, die hauptsächlich den Rhythmus und die Phonik umfasst. Die Hervorhebung der Form ist sicherlich nur eine Sache des Grades, denn mal werden phonische, mal rhythmische Eigenschaften betont, mal wird beides vernachlässigt, mal ist eine bestimmte Strukturierung latent und muss mit Hilfe der Statistik manifest gemacht werden.

Für die Untersuchung der Phonik ist bereits ein umfangreiches Begriffsinventar entwickelt worden, das es uns erlaubt, in Übereinstimmung mit der traditionellen Verslehre unsere Aufmerksamkeit Erscheinungen zu widmen, die eher auf der Oberfläche liegen, d.h. „gut sichtbar“ und fast oder ganz deterministisch sind, wie. z.B. Reim, auffällige Alliteration u.a. Mit kategorialen Begriffen kommt man jedoch nur bis zu einer bestimmten Tiefe in den Untersuchungsgegenstand hinein und macht vor einem noch umfangreicheren Bereich halt, dessen Erforschung nach feineren Methoden verlangt. Man kann zu Anfang nicht wissen, was einen in den tieferen phonischen Schichten erwartet, so dass man manchmal recht kompliziertere Verfahren verwenden muss, um Strukturierungen zu finden und zu erkennen. Das Auffinden einer Strukturierung bedeutet, dass man eine Erscheinung nicht als chaotisch, sondern als zumindest durch Zufallsgesetze entstanden – so paradox dies auch scheinen mag – oder als mit Absicht des Autors herbeigeführt, erkennt. Wie der Hintergrund auch immer beschaffen sein mag, im ersten Fall muss man berechnen, wie groß die Wahrscheinlichkeit des gegebenen (oder eines noch extremeren) Zustandes ist. Ist die Wahrscheinlichkeit, dass die fragliche Erscheinung durch reinen Zufall entstanden ist, zu klein, dann betrachten wir diese Erscheinung als signifikant, strukturiert, in Form einer latenten Tendenz existierend. Dass an dieser Stelle heftige Diskussionen entstehen können, steht außer Zweifel, denn ein klassischer Textanalytiker erkennt im Text nur das als real an, was er mit seiner Intuition erfassen kann. An diesem Verfahren ist nichts Falsches, es ist der normale Entdeckungsweg. Der Statistiker geht lediglich einen Schritt weiter: er berechnet, wie wahrscheinlich es ist, dass die betreffende Entität im Text vorhanden ist. Seine Schlüsse können die Intuition bestärken oder widerlegen, denn gerade dazu sind seine Methoden da. Man sollte nicht vergessen, dass die Hypothesen immer vom Textanalytiker kommen, und, nachdem sie den Apparat der Statistik durchgelaufen haben, wieder zum Textanalytiker zurückkommen müssen, der sie im Lichte seiner Intuition interpretiert.

Es gibt sicherlich auch Hypothesen, wie sie der Textanalytiker nicht einmal aufstellen kann, falls er nicht weiß, was die Statistik zu leisten vermag. Daher ist eine elementare Ausbildung in statistischen Methoden heutzutage für einen Textanalytiker bereits eine unerlässliche Voraussetzung für erfolgreiches Arbeiten. Aber auch unter sehr günstigen Umständen ist es nicht möglich, alle Aspekte eines Textes zu erforschen, weil deren Auswahl und Bestimmung von der Entwicklung unseres Begriffsapparates abhängen. Heute sehen wir vieles

noch nicht, was vermutlich später einmal sichtbar werden wird. Es ist aber zumindest möglich, einige Methoden anzugeben, die allgemein genug sind, um auch bei der Lösung künftiger Problemen behilflich sein zu können.

Die Lage in der Erforschung der Phonik von Gedichten ist heute recht kompliziert geworden (vgl. z.B. Gumenyuk et al. 2004), und es scheint, dass diese Entwicklung weiter gehen wird. Im folgenden werden wieder nur einfache Methoden mit vollständigen Rechnungen eingeführt, um weitere Analysen zu ermöglichen.

### 3.1. Die vokalische Struktur

Wir beschränken uns hier auf die Vokale des uns interessierenden Gedichts und schreiben sie so auf, wie sie hintereinander vorkommen. Die Transkription in Schema (I) wurde von K.-H. Best – dem wir zu verbindlichen Dank verpflichtet sind – durchgeführt, der sie als Germanist als eine von mehreren möglichen bezeichnet. Die Transkription ist eher phonetisch und entspricht K.-H. Bests muttersprachlicher Lesung. Die unterstrichenen Doppellaute sind Diphthonge. Man könnte einfachheitshalber die Vokallängen auch auslassen und sonst gleichartige kurze und lange Vokale zusammenfassen, wodurch man ein kleineres Vokalinventar mit vielleicht deutlicher ausgeprägten Mustern bekäme. Die Untersuchung bei unterschiedlichen Definitionen/Identifikationen der Vokale ist ein mögliches Forschungsproblem, dessen Lösung dem Leser überlassen sei.

(I)

| | | | |
|---|---|---|---|
| 1. | e: ai e o: ä: u a u i | 17. | i ai e a: e u: i i: e: |
| 2. | e i e: a: e i ai e i | 18. | ai e ö e o e i a e ö: |
| 3 | e: a e: a: e o: i e: a | 19. | ai e ö e ü: e e: e i e ai |
| 4. | e: a i: i e e e i: a: | 20. | u i: e u a e u i e i ai |
| 5. | ai o: a i u: o: a ai e i | 21. | ai a: e ai a: e u i: u: i o |
| 6. | i: a: e u: e: e o: i i | 22. | e ö: i o: e a ü: e o |
| 7. | e: e e ö: i i o: u ai | 23. | ai o: ai o: i e: e e au |
| 8. | ai o: e i ai e: e ai | 24. | e ai i: a e ai e o: au |
| 9. | u: i: e i o e: i i: | 25. | i i: e i i ai ai e ö: e e a |
| 10. | a: ö: e i: e i: i i i: | 26. | u i u: i i i o: au i e a |
| 11. | a u e u: e i a e: a | 27. | ai a: e ai a: e e a e: i a |
| 12. | ai e u e a a ü e e a | 28. | e ö: i a i: ai ai e a: |
| 13. | ai a: e ai a e u ö: e u: i | 29. | e: a: e au e e: ai e e i |
| 14. | a e e ö: i i: ai e e i | 30. | e: e i a e ae e e i |
| 15. | ai u: i ai e u: i ai i | 31. | e ai e: o: i ü: e u o: |
| 16. | i ü e e e oi e e: i | 32. | i ai e a e a i a: o: |

Eine derart dargestellte vokalische Struktur kann zwar zahlreiche *intuitive* unsystematische Beobachtungen ermöglichen, kann *auffällige* Klumpungen von Vokalen, *mögliche* positionale Strukturierungen u.ä. erkennen lassen, aber darum geht es nicht. Es geht um eine systematische Charakterisierung dieses Feldes und um eine formale Erfassung seines latentes Hintergrundes, der „mit bloßem Auge" nicht sichtbar ist.

Es ist genauso gut möglich, die konsonantische Struktur auf die gleiche Weise zu untersuchen oder das Gedicht als eine Folge von distinktiven Merkmalen zu kodieren und es als Zeitreihe, als Markovkette zu untersuchen. Die Zahl der Möglichkeiten hängt nur von unserer Sichtweise ab.

## 3.2. Häufigkeitscharakteristika

Die allgemeinen Häufigkeitscharakteristika sind bereits in Kapitel 2 dargestellt worden und werden hier auf die Vokale angewendet. Die Häufigkeiten der einzelnen, nach ihrer Frequenz geordneten Vokale findet man in Tabelle 3.1.

(a) Die *Wiederholungsrate* ergibt sich nach (2.5) als

$$R = (83^2 + 51^2 + \ldots + 1^2)/308^2 = 0.1360.$$

Dieser Wert liegt etwas über dem erwarteten Wert von $R_t = 2/18 = 0.1111$ (mit $K = 18$ Lauten im Inventar).

(b) Die *Entropie* ergibt sich nach (2.7) als

$$H = \text{ld } 308 - (83 \text{ ld } 83 + 51 \text{ ld } 51 + \ldots + 1 \text{ ld } 1)/308 = 3.3501,$$

und dieser Wert liegt leicht unter der theoretischen Kurve (2.8), deren Wert mit $K = 18$ $H_t = 3.4576$ beträgt.

Man sieht, dass beide Werte keine besondere Abweichung von der theoretischen Kurve aufweisen, was zusätzlich die Annahme bestätigt, dass Wiederholungsrate und Entropie vom Inventarumfang abhängen. Sollte einer der Werte aber von der Kurve in Abb. 2.5 bzw. 2.6 „optisch" stärker abweichen, so empfiehlt es sich, die Differenz auf Signifikanz zu prüfen (vgl. Altmann, Lehfeldt 1980: 162, 175).

Tabelle 3.1
Häufigkeiten der Vokale im „Erlkönig“

| Vokal | Rang | Häufigkeit | Zipf-Mandelbrot-Verteilung | Zipf-Alekseev Verteilung |
|---|---|---|---|---|
| e | 1 | 83 | 83.95 | 83.00 |
| i | 2 | 51 | 52.54 | 50.43 |
| ai | 3 | 35 | 35.99 | 36.94 |
| a | 4 | 28 | 26.19 | 27.62 |
| e: | 5 | 21 | 19.92 | 21.19 |
| i: | 6 | 15 | 15.66 | 16.62 |
| o: | 7 | 14 | 12.64 | 13.30 |
| a: | 8 | 13 | 10.42 | 10.82 |
| u | 9 | 12 | 8.74 | 8.92 |
| u: | 10 | 10 | 7.44 | 7.45 |
| ö: | 11 | 9 | 6.40 | 6.28 |
| o | 12 | 4 | 5.57 | 5.35 |
| au | 13 | 4 | 4.89 | 4.59 |
| ü: | 14 | 3 | 4.33 | 3.97 |
| ö | 15 | 2 | 3.86 | 3.46 |
| ü | 16 | 2 | 3.46 | 3.03 |
| ä: | 17 | 1 | 3.12 | 2.67 |
| oi | 18 | 1 | 2.24 | 2.36 |
| | | 308 | a = 1.9783<br>b = 2.7421<br>n = 18<br>$X^2_{14}$ = 9.40<br>P = 0.80 | a = 0.1419<br>b = 0.3493<br>n = 18<br>α = 0.2695<br>$X^2_{13}$ = 7.31<br>P = 0.89 |

(c) Die *Ranghäufigkeitsverteilung* folgt erwartungsgemäß einer „Zipf-schen“ Verteilung. Zwei davon haben wir ausgewählt, nämlich die Zipf-Mandelbrotsche Verteilung, deren Werte in der vierten Spalte von Tabelle 3.1 und in Abbildung 3.1, und die Zipf-Alekseev-Verteilung, deren Werte in der fünften Spalte von Tabelle 3.1 und in Abbildung 3.2 zu sehen sind.

Die Zipf-Mandelbrotsche Verteilung berechnet sich als

(3.1) $P_x = C(b+x)^{-a}, \quad x = 1,2,3,...,n,$

wobei C die Normierungskonstante $C^{-1} = \sum_{i=1}^{n} (b+i)^{-a}$ ist, und die Zipf-Alekseev-Verteilung in modifizierter Form als

$$(3.2) \quad P_x = \begin{cases} \alpha & x = 1 \\ \dfrac{(1-\alpha)x^{-(a+b\ln x)}}{T}, & x = 2,3,...,n, \end{cases}$$

wobei $T = \sum_{j=2}^{n} j^{-(a+b\ln j)}$ bedeutet (s. Altmann-Fitter 1997). Auch die geometrische Verteilung mit $P = 0.62$ ist sehr gut geeignet, und da Wiederholungsrate und Entropie dem aus der geometrischen Verteilung abgeleiteten theoretischen Wert entsprechen, kann man hier von einer elementaren Strukturierung sprechen, die sich von den sonstigen nichtpoetischen Texten nicht unterscheidet. Das heißt, die einfache Häufigkeit der Vokale im „Erlkönig" weist keine besondere Struktur auf, sondern eine solche, die „normalen" Texten entspricht. Dennoch ist die Anpassung dieser Modelle wichtig, weil sie die Theorie der Ranghäufigkeitsverteilung bekräftigten. Für weitere Literatur s. Hřebíček (1997), Chitashvili, Baayen (1993), Baayen (2001).

Auch wenn wir an dieser Stelle die Anwendung von Wahrscheinlichkeitsverteilungen sehr kurz gehalten haben, sollte man deren Bedeutung für die Textanalyse nicht unterschätzen. Sie zeigen nicht nur, dass hinter einer scheinbaren Willkür mit der der Autor den Text bezüglich einiger Spracheigenschaften gestaltet hat, latente Mechanismen stecken, derer er sich entweder nicht bewusst ist oder die er nicht zu steuern vermag, wenn er spontan schreibt, weil man beim spontanen Schreiben nicht alles kontrollieren kann. Solche Mechanismen erfasst man zunächst mit Hypothesen, die den Verlauf des Textes und die Anteile von bestimmten Entitäten im Text vorhersagen. Handelt es sich bei diesen Hypothesen um Wahrscheinlichkeitsverteilungen, so besteht die Möglichkeit, „rückwärts" zur Genese einer solchen Verteilung zu gehen und zu prüfen, ob die Bedingungen, unter denen diese entstanden ist, auch für die Textdaten gelten. Die Bedingungen können sehr allgemein sein, im Grunde entfernt man sich vom Text und betritt eine allgemeinere Ebene, die auch anderen Disziplinen übergeordnet ist. Auf diese Weise kann man - mutatis mutandis - Analogien mit anderen Erscheinungen der Welt finden. Man sollte nicht von der Behauptung ausgehen, Texte seien eine Erscheinung sui generis, die mit der übrigen Welt nichts zu tun hätten, denn Texte werden vom demselben Organ hervorgebracht, mit dem wir wahrnehmen, uns erinnern, wollen, reagieren, die Welt kreieren. Je besser es uns gelingt, Texte auf eine gemeinsame Grundlage mit anderen Phänomenen zu stellen, desto mehr beschleunigt sich der Fortschritt in der Textwissenschaft.

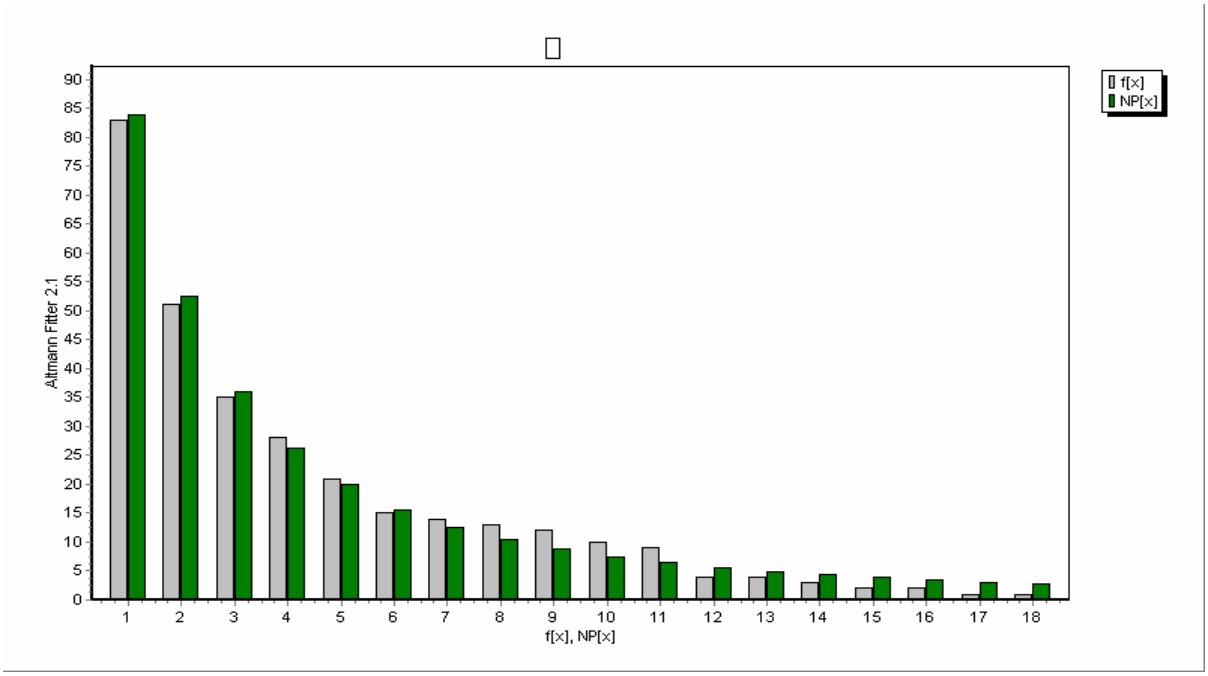


Abbildung 3.1. Rangverteilung der Vokale: Zipf-Mandelbrot-Verteilung

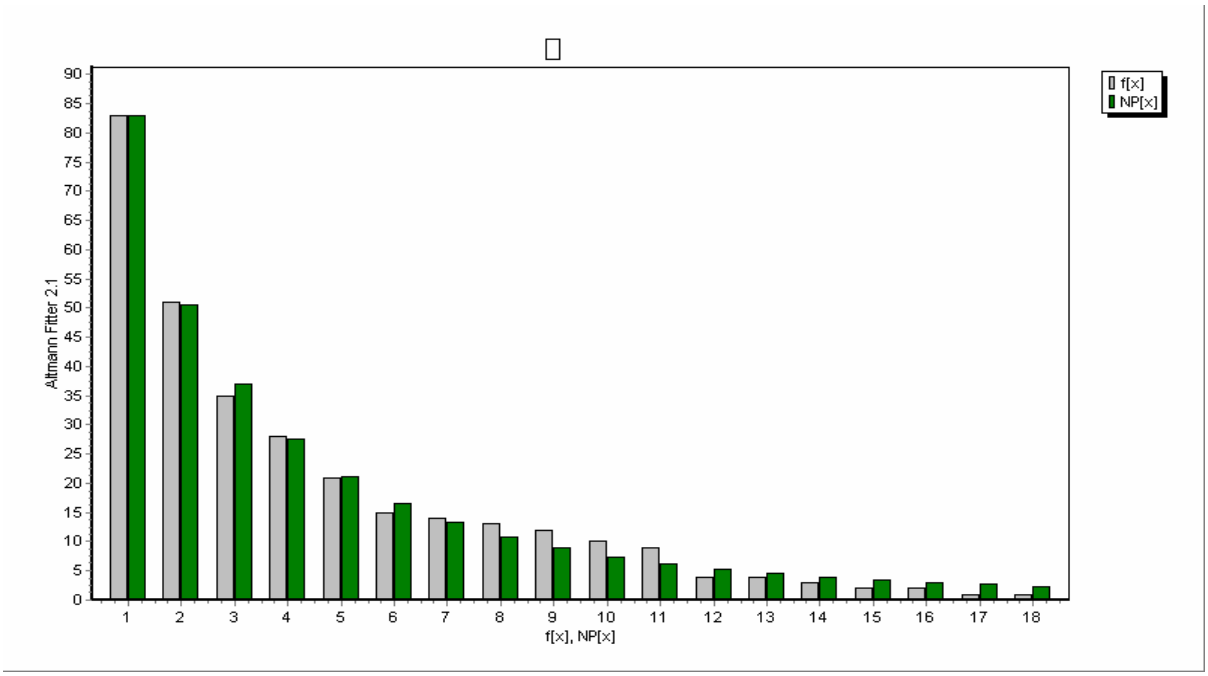


Abbildung 3.2. Ranghäufigkeitsverteilung der Vokale: Zipf-Alekseev-Verteilung

## 3.3. Assonanz

Da die einfache Frequenzen von Vokalen im „Erlkönig“ keine besonderen Tendenzen aufweisen, sich durch keinerlei Idiosynkrasien auszeichnen – sie verhalten sich im Rahmen der üblichen Modelle –, kann man sich fragen, ob es viel-

leicht Paare, Triaden usw. von Vokalen gibt, die sich öfters als erwartet wiederholen. In Schema (I) kann man solche Strukturen stellenweise mit bloßem Auge sehen – z.B. findet man in Zeile 2 und 3 untereinander die Folge [*e:, a:, e*] –, aber solche Beobachtungen sagen an sich noch nichts aus, weil sie rein zufällig entstanden sein können. Man muss eben ihre Nichtzufälligkeit nachweisen. Hierfür gibt es eine Reihe von Möglichkeiten, von denen wir nur einige zeigen werden.

### 3.3.1. Vokalpaare

Vokalische Assonanzen verwirklichen sich dadurch, dass eine bestimmte Folge von Vokalen entweder öfters als erwartet vorkommt oder dass bestimmte Folgen eine bestimmte parallele Position einnehmen. Der zweite Fall ist in Reimgedichten ganz üblich, denn dort ist die Wiederholung einer Sequenz von zwei Vokalen oder von mehreren Konsonanten die Regel; daneben aber gibt es auch Poesien, die Vokalfolgenassonanzen auch innerhalb des Verses signifikant oft benutzen, z.B. die malaiische Volkspoesie (vgl. Altmann 1963). An dieser Stelle soll es uns nur darum gehen, Folgen von Vokalen ohne Rücksicht auf ihre Position zu untersuchen. Der „Erlkönig“ ist etwas kurz, als dass auffällige Folgen vorkommen könnten, daher werden wir nur die Methode zeigen.

In Tabelle 3.2 findet man die Übergänge zwischen den Vokalen, wobei die Übergänge von einem Vers zum anderen nicht in Betracht gezogen wurde. Auf ähnliche Weise lassen sich auch Triaden (Trigramme) usw. ermitteln.

Das Testen der ganzen Tabelle auf Unabhängigkeit ist bei dem gegebenen Umfang nicht besonders ergiebig. Interessanter ist das Testen einzelner Felder, um „Assonanzmotive“ zu erkennen. Als Assonanzmotiv kann eine sich signifikant oft wiederholende Sequenz bezeichnet werden. Der Test für eine Zelle dieser Tabelle lautet

$$(3.3) \quad u = \frac{n_{ij} - \frac{n_i.n._j}{n}}{\sqrt{\frac{n_i.n._j(n-n_i.)(n-n._j)}{n^2(n-1)}}},$$

wobei $n_i.$ die rechte Randsumme, $n._j$ die untere Randsumme und $n$ die Gesamtsumme (hier $n = 275$) sind. So erhalten wir z.B. für die Folge [*a:, e*] mit den Ausgangsdaten aus der Tabelle

$$n_{a:,e} = 10, \quad n_{a:}. = 12, \quad n._e = 78, \quad n = 275,$$

Tabelle 3.2
Vokalpaare innerhalb des Verses

| | a | a: | ä: | e | e: | i | i: | o | o: | u | u: | ö | ö: | ü | ü: | ai | au | oi | $n_{i.}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 1 | - | - | 8 | 3 | 2 | 2 | - | - | 2 | - | - | - | 1 | 1 | 1 | - | - | 21 |
| a: | - | - | - | 10 | - | - | - | - | 1 | - | - | - | 1 | - | - | - | - | - | 12 |
| ä: | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | 1 |
| e | 9 | 2 | - | 12 | 4 | 15 | 3 | 2 | 3 | 6 | 5 | 2 | 7 | - | 1 | 8 | 2 | 1 | 83 |
| e: | 4 | 3 | - | 7 | - | 3 | - | - | 1 | - | - | - | - | - | - | 2 | - | - | 20 |
| i | 5 | 1 | - | 4 | 3 | 6 | 5 | 2 | 3 | - | 2 | - | - | 1 | 1 | 8 | - | - | 41 |
| i: | 2 | 1 | - | 4 | 1 | 2 | - | - | - | - | 1 | - | - | - | - | 2 | - | - | 13 |
| o | - | - | - | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 |
| o: | 2 | - | 1 | 2 | - | 3 | - | - | - | 1 | - | - | - | - | - | 1 | 2 | 1 | 12 |
| u | 2 | - | - | 2 | - | 3 | 2 | - | 1 | - | - | - | 1 | - | - | 1 | - | - | 12 |
| u: | - | - | - | 1 | 1 | 6 | 1 | - | 1 | - | - | - | - | - | - | - | - | - | 10 |
| ö | - | - | - | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 |
| ö: | - | - | - | 3 | - | 5 | - | - | - | - | - | - | - | - | - | - | - | - | 8 |
| ü | - | - | - | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 |
| ü: | - | - | - | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 3 |
| ai | - | - | - | 14 | 2 | 1 | 1 | - | 4 | - | 1 | - | - | - | - | 2 | - | - | 31 |
| au | - | - | - | 1 | - | 1 | - | - | - | - | - | - | - | - | - | - | - | - | 2 |
| oi | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 |
| $n_{.j}$ | 26 | 12 | 1 | 78 | 15 | 47 | 14 | 4 | 14 | 10 | 9 | 2 | 9 | 2 | 3 | 25 | 4 | 1 | 275 |

$$u = \frac{10 - \dfrac{12(78)}{275}}{\sqrt{\dfrac{12(78)(275-12)(275-78)}{275^2(247)}}} = 4.31.$$

Da dieser Wert größer als 1.96 ist, betrachten wir die Zelle als signifikant stark belegt, d.h., die Sequenz [a:, e] können wir als bevorzugt betrachten. In der Tat findet sich diese Folge in solchen Schlüsselwörtern wie „Vater“, „Knabe“. Weitere Folgen wie [*e, ö*:] und [*ö:, i*], die im Wort „Erlkönig“ vorkommen, ergeben $u = -0.73$ (nicht signifikant) bzw. $u = 3.46$ und [*ai, o*:] (in „mein Sohn“) ergibt $u = 2.10$.

Auf die beschriebene Weise lassen sich Assoziationsmotive mechanisch erkennen. In unserem Fall erleben wir keine Überraschung, weil wir die Schlüsselwörter bereits kennen. Es mag aber auch Fälle geben, wo in einem Gedicht nur ein „Echo“ des Schlüsselwortes in anderen Wörtern erscheint, selbst aber latent bleibt. Daher ist diese Methode als Instrument für Auffindung latenter Schlüssel-

wörter geeignet. Weitere umfangreiche Forschung ist nötig, um die Nützlichkeit dieses Verfahrens auszuloten.

Es gibt zahlreiche andere Möglichkeiten, Tabellen der hier zugrunde gelegten Art zu untersuchen (vgl. Schulz, Altmann 1988). Hier werden wir nur noch zeigen, wie man längere Folgen untersuchen kann.

### 3.3.2. Vokalfolgen

Das Erstellen einer dreidimensionalen Tabelle analog zu Tabelle 3.1 ist mit großen Schwierigkeiten verbunden, nicht nur wegen der unpraktischen Darstellung, die auch im Computer nicht besonders gut aussieht, sondern in unserem Fall auch wegen der allzu kleinen Zahlen. Ein Würfel mit $18^3 = 5832$ Zellen, auf dem sich nur 274 Eintragungen befinden, würde recht leer aussehen. Um aber trotzdem herausfinden zu können, ob eine längere Sequenz (d.h. eine solche von 3 oder mehr Vokalen) signifikant häufig vorkommt – wobei die fragliche Sequenz mindestens 2 mal vorkommen muss –, führen wir mit Hilfe der Poisson-Verteilung einen Test durch.

Betrachten wir eine Folge von drei Vokalen $V_1V_2V_3$, dann können wir die Wahrscheinlichkeit, dass die Vokale in dieser Reihenfolge vorkommen, mit Hilfe ihrer relativen Häufigkeiten schätzen, d.h.

$$(3.4) \quad P(V_1V_2V_3) = \frac{n_{V_1}}{n} \cdot \frac{n_{V_2}}{n} \cdot \frac{n_{V_3}}{n}.$$

Die erwartete Häufigkeit dieser Folge ergibt sich als

$$(3.5) \quad E(V_1V_2V_3) = nP(V_1V_2V_3) = \frac{n_{V_1} n_{V_2} n_{V_3}}{n^2} = \lambda \ .$$

Da es sich um seltene Ereignisse handelt (wie man leicht sehen könnte, wenn sich eine dreidimensionale Tabelle erstellen ließe), betrachten wir λ als den Parameter der Poisson-Verteilung. Um nun festzustellen, ob die beobachtete Häufigkeit $x_c$ der Folge $V_1V_2V_3$ signifikant groß ist, berechnen wir die Wahrscheinlichkeit

$$(3.6) \quad P(X \geq x_c) = \sum_{x=x_c}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = 1 - \sum_{x=1}^{x_c-1} \frac{e^{-\lambda} \lambda^x}{x!}.$$

Ist das berechnete $P \leq 0.05$ und wurde die Folge mindestens zweimal beobachtet, dann bezeichnen wir die Sequenz als signifikant häufig vorkommend.

Betrachten wir die Folge [*e:*, *a:*, *e*], die wir in der ersten Strophe zweimal und im 29. Vers noch einmal finden. Sie kommt also dreimal vor. Die Häufigkeiten der Laute entnehmen wir der Tabelle 3.1: $n_{e:} = 21$, $n_{a:} = 13$, $n_e = 83$, und die Zahl aller Vokale ist $n = 308$. Daraus erhalten wir

$$\lambda = 21(13)83/308^2 = 0.2389.$$

Gemäß (3.6) müssen wir

$$P(X \geq 3) = 1 - (P_0 + P_1 + P_2)$$

berechnen. Setzen wir den Wert des Parameters λ in Formel (3.6) ein, so erhalten wir zunächst

$$P_0 = e^{-\lambda} = 2.7183^{-0.2389} = 0.7875.$$

Die restlichen Werte berechnen wir rekursiv aus

$$P_x = \frac{\lambda}{x} P_{x-1},$$

d.h.

$$P_1 = \frac{0.2389}{1} 0.7875 = 0.1881$$

$$P_2 = \frac{0.2389}{2} 0.181 = 0.0225.$$

Setzen wir diese Zahlen in die Formel ein, so erhalten wir

$$P(X \geq 3) = 1 - (0.7875 + 0.1881 + 0.0225) = 0.0019.$$

Dies ist ein hoch signifikantes Resultat und besagt, dass die Folge [*e:*, *a:*, *e*] ein verstecktes Assonanzmotiv mit sich trägt. Man kann leicht feststellen, dass dieses Motiv lediglich in „der Vater“, „dem Knaben“ und „dem Vater“ vorkommt, d.h. bereits bekannt ist.

Was untersuchen wir eigentlich bei den Vokalfolgen? Statistisch gesehen, beurteilen wir lediglich, ob die gegebene Folge als unabhängiges Vorkommen bestimmter Vokale betrachtet werden kann oder ob ihr Vorkommen durch die gegebenen Positionen bedingt ist. Eine niedrige Wahrscheinlichkeit des unab-

hängigen Vorkommens besagt, dass die untersuchte Folge ein Assonanzmotiv darstellt.

### 3.4. Alliteration

Zwar kommen im „Erlkönig“ Vokale seltener als Konsonanten am Wortanfang vor, aber es interessiert uns, ob in speziellen Positionen, hier in der ersten Silbe der Verse, eine Präferenz für bestimmte Vokale vorherrscht. Mit ähnlichen Problemen haben sich bereits Krappe (1921), Shewan (1925), Skinner (1939) beschäftigt, es ist eines der häufig untersuchten Erscheinungen im Beowulf, systematisch wird es aber nicht vorangetrieben (vgl. auch Wimmer et al. 2003).

In Schema (I) dieses Kapitels kann man feststellen, dass an $n = 32$ Versanfängen, d.h. als erste Vokale des Verses, folgende Laute vorkommen ($x_i$ = Vorkommenshäufigkeit):

| Vokal *i* | $x_i$ |
|---|---|
| <u>ai</u> | 10 |
| e: | 6 |
| e | 5 |
| i | 4 |
| a | 2 |
| u | 2 |
| i: | 1 |
| a: | 1 |
| | 32 |

Um zu testen, ob eine gegebene oder eine noch extremere Anzahl eines Vokals am Versanfang als signifikant häufig zu betrachten ist, reicht es wieder, die entsprechende Wahrscheinlichkeit zu berechnen. Da ein Laut an dieser Stelle entweder vorkommt oder nicht, handelt es sich um eine Binomialsituation, in der wir die entsprechende Wahrscheinlichkeit als

$$P_x = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.7}$$

berechnen. In unserem Fall ist $n = 32$, d.h. die Zahl der Positionen (Verse) im ganzen Gedicht, und $p$ schätzt man ab als die Proportion des Vokals *i* im ganzen Gedicht. Diese Proportion berechnen wir wieder aus Tabelle 3.1, indem wir die entsprechende Häufigkeit durch 308 dividieren. So ergibt sich beispielsweise für [<u>*ai*</u>]

$$p_{\underline{ai}} = 35/308 = 0.1136\,.$$

Um die entsprechende Wahrscheinlichkeit zu erhalten, berechnen wir

$$(3.8)\quad P(X \geq x_i) = \sum_{x=x_i}^{n} \binom{n}{x} p_i^x (1-p_i)^{n-x} = 1 - \sum_{x=0}^{x_i-1} \binom{n}{x} p_i^x (1-p_i)^{n-x} \quad .$$

In unserem Fall, in dem wir [*ai*] 10-mal am Versanfang gefunden haben, ist

$$P(X \geq 10) = 1 - \sum_{x=0}^{9} \binom{32}{x} 0.1136^x (0.8864)^{32-x}\,.$$

Man fängt die Rechnung nach (3.7) mit $P_0$ an, erhält so

$$P_0 = (1-p)^n = 0.8863^{32} = 0.0211$$

und rechnet rekursiv mit Hilfe von

$$(3.9)\quad P_x = \frac{n-x+1}{x}\frac{p}{1-p}P_{x-1}$$

weiter. So erhält man der Reihe nach

$$P_1 = \frac{32-1+1}{1}\frac{0.1136}{0.8864}0.0211 = 0.0865$$

$$P_2 = \frac{32-2+1}{2}\frac{0.1136}{0.8864}0.0865 = 0.1718$$

$$P_3 = \frac{30}{3}0.1282(0.1718) = 0.2202$$

$$P_4 = \frac{29}{4}0.1282(0.2202) = 0.2046$$

$$P_5 = \frac{28}{5}0.1282(0.2046) = 0.1469$$

$$P_6 = \frac{27}{6}0.1282(0.1469) = 0.0849$$

$$P_7 = \frac{26}{7}0.1282(0.0849) = 0.0403$$

$$P_8 = \frac{25}{8}0.1282(0.0403) = 0.0161$$

$$P_9 = \frac{24}{9}0.1282(0.1061) = 0.0055.$$

Addiert man diese Zahlen, so erhält man

$$\begin{aligned} P(X \geq 10) \;=\; & 1 - (0.0211 + 0.0865 + 0.1718 + 0.2202 + 0.2046 + \\ & + 0.1469 + 0.0849 + 0.0403 + 0.0161 + 0.0055) = 0.0021. \end{aligned}$$

Da diese Wahrscheinlichkeit sehr klein ist (< 0.05), schließen wir, dass [*ai*] am Versanfang (erste Silbe) signifikant oft vorkommt.

Betrachtet man für [*e*:] mit $p_{e:} = 21/308 = 0.0682$ die Summe

$$P(X \geq 6) = 1 - \sum_{x=0}^{5} \binom{32}{x} 0.0682^x (0.9318)^{32-x} = 0.019,$$

die kleiner als 0.05 ist, so kann man auch noch recht zuverlässig von einer Tendenz sprechen, an den Versanfang ein [*e*:] zu stellen. Für die anderen Vokale ergibt sich keine derartige Tendenz. Auf die gleiche Weise kann man natürlich auch die Konsonanten testen.

Es ergeben sich für die weitere Forschung sofort zahlreiche Fragen:

(a) Sind diese zwei Tendenzen charakteristisch nur für den „Erlkönig“ oder für alle Gedichte von Goethe, oder

(b) gelten sie für deutsche Gedichte allgemein? Denn die Vokale [*ai*] und [*e*:] kommen in allen unbestimmten Artikeln und in einigen Formen des bestimmten Artikels (*der, dem, den*) vor. Die Frage lässt sich beantworten, wenn man die gleiche Zählung ohne diese Artikel am Versanfang durchführt bzw. wenn man viele Gedichte analysiert.

(c) In vielen Sprachen gibt es Gedichte, in denen *innerhalb* des Verses nachweisbare Alliteration vorhanden ist. Wimmer et al. (2003) haben einen Alliterationskoeffizienten für jeden Vers entwickelt und den Durchschnitt aller Werte dieses Koeffizienten betrachteten sie als Alliterationskoeffizienten eines Ge-

dichts. Die Rechnung ist etwas umständlicher, weil hier die Multinomialverteilung benutzt werden muss.

(d) Falls man eine bestimmte Art von Alliteration in einem Gedicht entdeckt hat, ist diese mit der Semantik des Textes verbunden?

(e) In vielen Sprachen gibt es Gedichte, in denen die betonten Vokale eine bestimmte Färbung haben − z.B. alle sind Hintervokale −, wobei diese Färbung von Vers zu Vers unterschiedlich sein kann. Ein slowakisches Gedicht (Chalupka, Morho!) fängt mit dem Vers an: „Duní Dunaj a luna za lunou sa valí“ („Es dröhnt die Donau, und Welle auf Welle rollt“), wobei sogar das Wort „vlna“ (Welle) als „luna“ dargestellt wurde, um eine bestimmte Atmosphäre hervorzurufen, obwohl die Donau nie dröhnt und die Wellen höchstens 5 cm hoch sind – wenn es weht.

(f) Es gibt poetische Verfahren, bei denen die Lage eines Vokal im Vers fest vorgeschrieben ist (Javanisch) und dies als eine Regel gilt.

Der „Erlkönig“ konfrontiert uns nicht mit all diesen Erscheinungen, so dass wir uns methodisch etwas einschränken müssen, aber bei einer Erweiterung der Forschung kann man diese Probleme angehen.

### 3.5. Reim

Für den Reim ergibt sich die gleiche Situation wie für die Alliteration. Stellt man die Häufigkeit der Vokale fest, die sich in der letzten Position des Verses befinden, so erhält man für den „Erlkönig“ folgende Werte:

| Vokal *i* | $x_i$ |
|---|---|
| i | 10 |
| a | 6 |
| ai | 4 |
| a: | 2 |
| i: | 2 |
| o | 2 |
| o: | 2 |
| au | 2 |
| e: | 1 |
| ö: | 1 |
| | 32 |

Die Rechnung gemäß Formel (3.8) ergibt, dass nur [*i*] signifikant häufig in dieser Position vorkommt; es ist nämlich $P(X \geq 10) = 0.0297$, alle anderen Vokale haben ein $P > 0.05$.

Die qualitative Seite des Reims kann für die Geschichte des Reimbildung eine Rolle spielen. Häufig verwendete Reime nutzen sich ab, und mit der Zeit werden sie gemieden (vgl. Štukovský, Altmann 1965, 1966), und auch die Reimtechnik kann sich ändern. Es wäre daher interessant zu untersuchen, wie sich die Vertretung der Vokale in den Reimen deutscher Gedichte entwickelt, ob z.B. [*i*] abnimmt, ob es zyklische Bewegungen gibt usw. Aber auch in bezug auf einen einzigen Dichter kann eine solche Untersuchung zu wichtigen Schlussfolgerungen führen. Eine andere Frage ist die Verknüpfung der Gesamtstimmung eines Gedichtes mit der Vokalklasse im Reim. Zahlreiche andere Probleme können untersucht werden, z.B. die Frage, welche Wortarten im Reim benutzt werden, aus wie vielen Lauten der Reim besteht, ob die Reimwörter auf einen Vokal oder auf einen Konsonanten auslauten, wie lang die Reimwörter sind usw. Viele dieser Probleme sind sprach- und grammatikgebunden. Allgemeine Schlüsse wird man erst dann ziehen können, wenn viele Daten zumindest aus einer Sprache zur Verfügung stehen.

### 3.6. Distanzen

Sind die Abstände zwischen gleichen Vokalen rein zufällig, oder gibt es eine Strukturierung der Abstände? Diese Frage kann am einfachsten so beantwortet werden, dass man die Abstände für einzelne Vokale zuerst separat berechnet und dann die Verteilungen der Distanzen mit dem Zufallsmodell von Zörnig (s. Kap. 2.7) vergleicht. Bei zufälliger Platzierung folgen die Distanzen diesem Modell. Wenn es aber eine Strukturierung gibt, dann müssen sie für einzelne Vokale der negativen Binomialverteilung gehorchen, da diese einen „Klumpungstrend“ ausdrückt, der aus einem Poisson-Prozess folgt (vgl. Strauss et al. 1984; Altmann 1988: 151 ff.). Mischt man aber die einzelnen Distanzverteilungen zusammen, dann müsste sich eine gemischte negative Binomialverteilung ergeben, wobei man die Zahl der Komponenten dieser Verteilung nach Bedarf verringern oder vergrößern kann. Es ist zu erwarten, dass in den meisten Fällen zwei Komponenten ausreichen werden.

Bei der Überprüfung verfahren wir folgendermaßen. Wir betrachten den ersten Laut [*e*:] in Schema (I) und zählen den Abstand zum nächsten [*e*:], von diesem denjenigen zum nächsten [*e*:] usw., bis zum Ende des Schemas. So erhalten wir hintereinander:

10,6,1,4,1,4,17,4,13,7,18,48,9,16,40,44,11,4,4,11

Auf ähnliche Weise erhalten wir die Abstände auch für die anderen Vokale. Man fängt immer beim ersten Vorkommen an und endet beim letzten. *Abstand* bedeutet die Zahl der Vokale zwischen zwei gleichen Vokalen. Möglicherweise ergeben sich bestimmte sequentielle Muster auch für einzelne Vokale, aber dies kann man mit größerer Konfidenz nur bei längeren Texten ermitteln.

Fasst man alle Abstände zusammen, so erhält man die empirische Verteilung, die in Tabelle 3.3, Spalte 2, dargestellt ist. Aber schon die Berechnung des nullten Abstandes nach Zörnigs Formel, wo man $NP_0 = 41$ erwartet, zeigt, dass man weit von reiner Zufälligkeit entfernt ist. Daher passt man die aus zwei Komponenten bestehende gemischte negative Binomialverteilung an und bekommt die Resultate in der dritten Spalte von Tabelle 3.3. Die Formel ist

$$(3.10)\quad P_x = \alpha \binom{k_1 + x - 1}{x} p_1^{k_1} q_1^x + (1-\alpha) \binom{k_2 + x - 1}{x} p_2^{k_2} q_2^x, \quad x = 0,1,2,...$$

Wie man sieht, ist diese Anpassung mit entsprechenden Zusammenfassungen der Häufigkeitsklassen (so, dass der theoretische Wert immer > 1 war) sehr gut und zeigt, dass hier eine Strukturierung stattgefunden hat. Mit „bloßem Auge“ kann man sie aber nicht entdecken, und der Autor war sich vermutlich nicht einmal dessen bewusst, dass er sie erzeugte. In Abbildung 3.3, die diesmal auf eine andere Weise gezeichnet wurde, sieht man den Verlauf der empirischen und der theoretischen Werte.

Man kann die Untersuchung auch für andere Konfigurationen durchführen, z.B. so, dass man kurze und lange Vokale zu einer Klasse zusammenfasst oder indem man die Vokale nach der Zungenlage ordnet usw. Man kann damit solange experimentieren, bis man sehr ausgeprägte Strukturen findet – falls es sie gibt.

Man kann in Schema (I) der Reihe nach auch die Distanz zum nächsten gleichen Vokal vom Anfang bis zum Ende des Gedichtes ermitteln, wodurch eine Zeitreihe entsteht, die möglicherweise interessante Eigenschaften aufweist (vgl. Hřebíček 2000). Das Gleiche kann man auch für Konsonanten separat oder zusammen mit Vokalen durchführen.

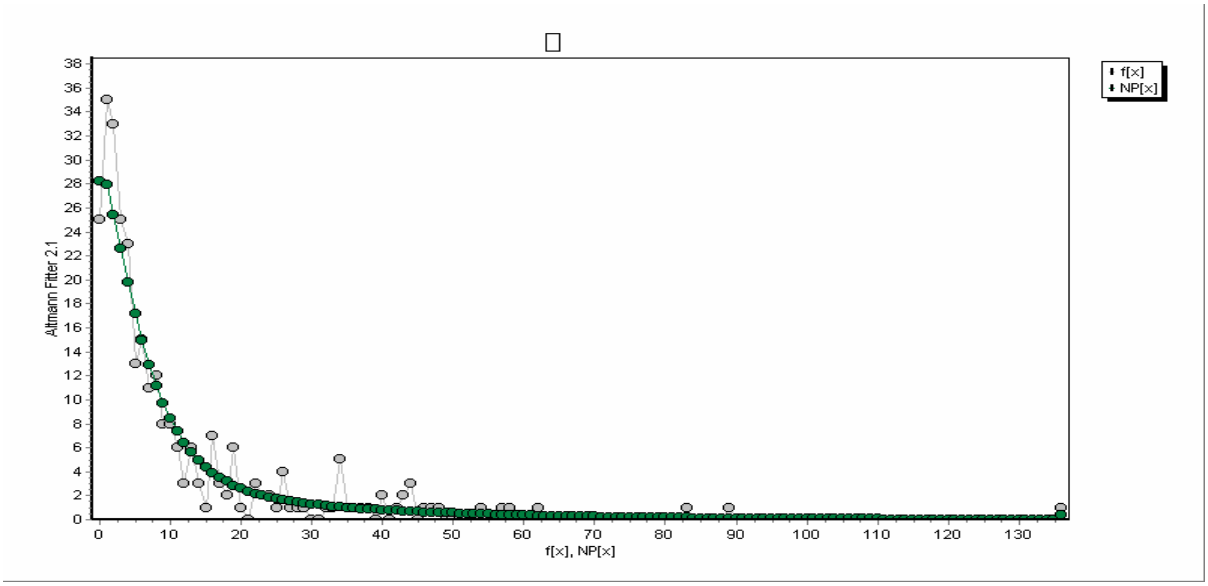


Abbildung 3.3. Verteilung der Distanzen zwischen gleichen Vokalen

Tabelle 3.3
Verteilung der Distanzen zwischen Vokalen

| x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 25 | 28.24 | 46 | 1 | 0.63 | 92 | 0 | 0.10 |
| 1 | 35 | 27.93 | 47 | 1 | 0.61 | 93 | 0 | 0.10 |
| 2 | 33 | 25.45 | 48 | 1 | 0.58 | 94 | 0 | 0.09 |
| 3 | 25 | 22.57 | 49 | 0 | 0.56 | 95 | 0 | 0.09 |
| 4 | 23 | 19.76 | 50 | 0 | 0.54 | 96 | 0 | 0.09 |
| 5 | 13 | 17.19 | 51 | 0 | 0.52 | 97 | 0 | 0.08 |
| 6 | 15 | 14.90 | 52 | 0 | 0.50 | 98 | 0 | 0.08 |
| 7 | 11 | 12.90 | 53 | 0 | 0.48 | 99 | 0 | 0.08 |
| 8 | 12 | 11.17 | 54 | 1 | 0.46 | 100 | 0 | 0.07 |
| 9 | 8 | 9.69 | 55 | 0 | 0.44 | 101 | 0 | 0.07 |
| 10 | 8 | 8.42 | 56 | 0 | 0.42 | 102 | 0 | 0.07 |
| 11 | 6 | 7.33 | 57 | 1 | 0.41 | 103 | 0 | 0.07 |
| 12 | 3 | 6.41 | 58 | 1 | 0.39 | 104 | 0 | 0.06 |
| 13 | 6 | 5.63 | 59 | 0 | 0.38 | 105 | 0 | 0.06 |
| 14 | 3 | 4.96 | 60 | 0 | 0.36 | 106 | 0 | 0.06 |
| 15 | 1 | 4.40 | 61 | 0 | 0.35 | 107 | 0 | 0.06 |
| 16 | 7 | 3.92 | 62 | 1 | 0.34 | 108 | 0 | 0.05 |
| 17 | 3 | 3.51 | 63 | 0 | 0.32 | 109 | 0 | 0.05 |
| 18 | 2 | 3.15 | 64 | 0 | 0.31 | 110 | 0 | 0.05 |
| 19 | 6 | 2.85 | 65 | 0 | 0.30 | 111 | 0 | 0.05 |
| 20 | 1 | 2.60 | 66 | 0 | 0.29 | 112 | 0 | 0.05 |
| 21 | 0 | 2.37 | 67 | 0 | 0.28 | 113 | 0 | 0.04 |
| 22 | 3 | 2.18 | 68 | 0 | 0.26 | 114 | 0 | 0.04 |
| 23 | 2 | 2.01 | 69 | 0 | 0.25 | 115 | 0 | 0.04 |
| 24 | 2 | 1.87 | 70 | 0 | 0.24 | 116 | 0 | 0.04 |
| 25 | 1 | 1.74 | 71 | 0 | 0.24 | 117 | 0 | 0.04 |
| 26 | 4 | 1.63 | 72 | 0 | 0.23 | 118 | 0 | 0.04 |
| 27 | 1 | 1.52 | 73 | 0 | 0.22 | 119 | 0 | 0.03 |
| 28 | 1 | 1.43 | 74 | 0 | 0.21 | 120 | 0 | 0.03 |
| 29 | 1 | 1.35 | 75 | 0 | 0.20 | 121 | 0 | 0.03 |
| 30 | 0 | 1.28 | 76 | 0 | 0.19 | 122 | 0 | 0.03 |
| 31 | 0 | 1.21 | 77 | 0 | 0.19 | 123 | 0 | 0.03 |
| 32 | 1 | 1.15 | 78 | 0 | 0.18 | 124 | 0 | 0.03 |
| 33 | 1 | 1.10 | 79 | 0 | 0.17 | 125 | 0 | 0.03 |
| 34 | 5 | 1.05 | 80 | 0 | 0.16 | 126 | 0 | 0.03 |
| 35 | 1 | 1.00 | 81 | 0 | 0.16 | 127 | 0 | 0.02 |
| 36 | 1 | 0.96 | 82 | 0 | 0.15 | 128 | 0 | 0.02 |
| 37 | 1 | 0.91 | 83 | 1 | 0.15 | 129 | 0 | 0.02 |
| 38 | 1 | 0.88 | 84 | 0 | 0.14 | 130 | 0 | 0.02 |
| 39 | 0 | 0.84 | 85 | 0 | 0.13 | 131 | 0 | 0.02 |
| 40 | 2 | 0.80 | 86 | 0 | 0.13 | 132 | 0 | 0.02 |
| 41 | 0 | 0.77 | 87 | 0 | 0.12 | 133 | 0 | 0.02 |
| 42 | 1 | 0.74 | 88 | 0 | 0.12 | 134 | 0 | 0.02 |
| 43 | 2 | 0.71 | 89 | 1 | 0.11 | 135 | 0 | 0.02 |
| 44 | 3 | 0.68 | 90 | 0 | 0.11 | 136 | 1 | 0.42 |
| 45 | 0 | 0.66 | 91 | 0 | 0.11 | | | |

$k$ = 1.1899, $p_1$ = 0.0414, $p_2$ = 0.1785, $a$ = 0.2949, FG = 49, $X^2$ = 47.1884, $P$ = 0.5468

## 3.7. Euphonie im allgemeinen

Untersucht man die phonische Seite von Gedichten, so hat man es immer mit der Euphonie zu tun. Diese stellt zwar einen Begriff dar, aber von einer operationalen Definition dieses Begriffs sind wir noch weit entfernt. Sie umfasst alle hier behandelten Phänomene und weist zwei Erscheinungsformen auf: (1) die allgemeine Häufigkeit der phonischen Elemente und (2) ihre Platzierung im Vers. Wie immer man eine euphonische Hypothese auch formuliert, in allen Fällen muss gezeigt werden, dass entweder die Häufigkeit oder die Platzierung der Elemente nicht als zufällig zu betrachten ist. Es gibt selbstverständlich auch Erscheinungen dieser Art, die zwar ein signifikant häufiges oder platziertes Vorkommen aufweisen, jedoch nicht als euphonisch zu betrachten sind, z.B. Kakophonien, Zungenbrecher u.ä. Sie gehören aber trotzdem in diese Sparte, die sehr breit ist.

Der schwache Entwicklunggrad der Forschung auf diesem Gebiet lässt sich dadurch erklären, dass die meisten Erscheinungen textgebunden sind, kaum mit quantitativen Methoden untersucht wurden, und sich, wenn dies doch geschah, inkommensurable Resultate ergaben. Die Möglichkeit der Theoriebildung war bis zum Erscheinen der Arbeiten von Hřebíček nicht einmal in Betracht gezogen worden, die größte Aufmerksamkeit wurde der Auszählung von Elementen gewidmet, ohne dass entsprechende Hypothesen aufgestellt wurden.

Sicher ist, dass auch in den Fällen, in denen man eine vermutete euphonische Struktur nachweisen kann, ohne Hilfe der Psychologie oder der Psycholinguistik keine aussagekräftige Interpretation des Phänomens möglich ist. Man kann zwar Regularitäten finden, kann sie sogar deduktiv ableiten, aber aus ihnen lassen sich vorläufig noch keine Rückschlüsse auf die Reaktionen des Gehirns ziehen. Warum betrachtet das Gehirn etwas als wohlklingend? Vielleicht wird sich die Sprachforschung hier einmal von der Musik inspirieren lassen, aber die Textologie muss auch ohne sie auskommen.

Eine euphonische Absicht kann durch Befragung des Textautors – falls er noch lebt –, durch Befragung von Lesern oder von Hörern oder statistisch ermittelt werden. Die dritte Hinangehensweise ist die sicherste, weil sie völlig objektiv ist. Wenn es in einem Gedicht Euphonie gibt, dann muss sie statistisch ermittelbar sein. Ihre Formen können aber so komplex sein, dass oft sehr fortgeschrittene Methode nötig sind, um sie zu entschlüsseln. Das Resultat braucht weder mit der Ansicht des Autors, der spontan schreibt, übereinzustimmen, noch mit der Ansicht vieler Leser, die intuitiv etwas spüren, dieses aber nicht belegen können,. Wie man das Phänomen benennt, ist eine cura posterior. Es geht um bestimmte Wirkungen, man kann nie sicher sein, ob diese durch Rhythmus, Lautung, Wortwahl oder wodurch auch immer hervorgerufen wurden. Die „schönsten“ Gedichte einer Sprache können für jemanden, der diese Sprache nicht kennt, als reinste Kakophonie klingen. Ist nun die Euphonie auch mit der Bedeutung assoziiert? Wir wissen es nicht. Ein ausschließlich quantitativ arbeitender Textologe kann nur danach trachten, im scheinbaren Chaos des Textes nach Inseln der Stabilität

zu suchen, die weder dem Autor noch dem Leser bewusst sind, deren Existenz aber mit bestimmten Methoden erfassbar ist.

# 4. Wörter

Auf der inhaltlichen Ebene sind Wörter diejenigen Entitäten, die man als erste zu untersuchen versucht. Heutzutage sind jedoch schon so viele Eigenschaften des Wortes definiert worden – phonetische, morphologische, syntaktische, semantische, etymologische, historische, psycholinguistische, ästhetische u.a. –, dass man notgedrungen gezielt auswählen und sich auf einige wenige beschränken muss. Welche Eigenschaften man auswählt, hängt vom Ziel der Untersuchung ab. Bei reinen *Beschreibungen* reicht es, die erfassten Eigenschaften mit einem Index, einem Vektor, einer Kurve oder einer Verteilung zu erfassen, denn in dieser Form können sie anderen Forschern als Vergleichsmaterial dienen. Schon auf dieser Ebene sind Hypothesen möglich, denn man kann beispielsweise testen, ob sich die Mittelwerte einer Eigenschaft in zwei Texten signifikant unterscheiden oder nicht. Auf diese Weise kann man eine Klassifikation von Texten aufstellen, die bereits eher eine *Charakterisierung* der Texte darstellt, wenn wir eine solche von der bloßen Beschreibung unterscheiden wollen. Die Zugehörigkeit eines Textes zu einer Klasse setzt automatisch bestimmte Ausprägungen ausgewählter Eigenschaften voraus.

Wie bekannt, existiert keine Eigenschaft isoliert, sondern hängt mit anderen zusammen. Die Suche nach Zusammenhängen zwischen Eigenschaften versetzt uns auf die *relationale* Stufe der Forschung, eine Vorstufe der Theoriebildung. Texte, ebenso wie Sprachen, sind Gebilde, die man als Systeme auffassen kann, die nicht nur ein sehr komplexes „Innenleben“, sondern auch eine sehr komplexe Umgebung haben. Die Entwirrung aller Zusammenhänge ist eine Aufgabe für Generationen von Forschern, von denen jede nur einen kleinen Beitrag leisten kann. Schon die Entdeckung einer Korrelation, eines Verlaufes, einer stochastischen Verknüpfung zweier Eigenschaften in einem einzigen Text ist ein richtungsweisender Hinweis für andere Forscher, ein Mosaikstein, der später in eine Theorie eingehen kann. Die relationale Betrachtung kann am Anfang rein empirisch, induktiv sein, auch wenn man nicht nur einfache Korrelationen berechnet, sondern die Relation schon mit einer Funktion erfasst. Eine solche empirische − gut passende − Funktion ist allerdings nur eine induktive Hypothese, die zwar etwas mehr verspricht als eine reine Korrelation, aber trotzdem nur einen prototheoretischen Ansatz darstellt.

*Theoriebildende* Ansätze müssen aus Theorien, Gesetzen oder Axiomen abgeleitet werden. Die Erforschung konkreter Texte ist dann nicht das Ziel, sondern eine Instanz zur Überprüfung solcher deduktiv gewonnener Hypothesen. Bei der Betrachtung des Wortes bewegt sich die Textologie gleichzeitig auf verschiedenen Ebenen. Stellenweise, z.B. in der Worthäufigkeitsforschung, ist die Lage recht fortgeschritten, in anderen Bereichen aber noch sehr embryonal. Ein grundlegender Fehler unterläuft besonders häufig Mathematikern und Physikern, die sich mit der Sprache beschäftigen, nämlich die Aufstellung von theoretisch begründeten Modellen, die für bestimmte Daten entwickelt worden sind, für an-

dere Daten aber nicht ohne weiteres adäquat zu sein brauchen. Dabei wird explizit oder implizit die Meinung vertreten, dass ein Modell für alle Texte aller Sprachen gelten müsse, ohne Rücksicht auf die Randbedingungen, die von Mathematikern notgedrungen vernachlässigt werden, da diese im Normalfall keine ausgebildeten Linguisten sind. Auch Analogien werden sehr oft auf Verdacht hin geäußert, etwa wenn eine Kurve aus der Physik für eine Texteigenschaft geeignet zu sein scheint. Beim Aufbau von Theorien suchen wir nach Mechanismen, die im Text wirken, und über diese Mechanismen äußern wir Hypothesen, die wir aus bestimmten Grundannahmen ableiten. Dabei ist es gut möglich, dass wir für eine Erscheinung viele Kurven oder Verteilungen ableiten müssen, je nach den Anfangs- und den Randbedingungen die in einer Sprache oder in einem Text gegeben sind. Und schon die Auffindung dieser Bedingungen ist der schwierigere Teil der theoretischen Arbeit, die nur Textologen leisten können. Der meistversprechende Ansatz ist hier die linguistische Synergetik (vgl. Köhler 1986, 2002), die trotz eines völlig unterschiedlichen Ansatzes zu den gleichen Funktionen führt wie die einheitliche Ableitung von Sprachgesetzen (vgl. Wimmer, Altmann 2005), die man schon oft an Texten überprüft hat (vgl. Best 2001).

In diesem Kapitel werden wir als Hauptziel die Erfassung bestimmter Eigenschaften im „Erlkönig“ vor Augen haben. Wir beschränken uns hierbei auf die Worthäufigkeit und ihre Verarbeitung, die recht fortgeschritten ist.

## 4.1. Worthäufigkeit

Man kann Worthäufigkeit in einem Text auf zwei unterschiedliche Weisen zählen:

(a) Man unterscheidet zwischen den Formen eines Wortes und beobachtet (zählt) jede Form separat, d.h., man untersucht die Wortformhäufigkeit. Bei diesem Verfahren ist keine spezielle Zubereitung des Textes erforderlich. Ein Programm nimmt die Leerstellen zwischen den Wortformen wahr und eliminiert automatisch die Interpunktion. Diese Art der Zählung wird besonders von Nichtlinguisten bevorzugt, weil sie dabei nur mechanische Probleme zu lösen haben. Für bestimmte Zwecke, z.B. für morphologische Untersuchungen wie Kasusproduktivität, Numerus u.a., ist dieses Vorgehen gleichfalls geeignet, nicht aber für die Messung des Informationsflusses in einem Text (type-token-Verhältnis), die Lösung des Textabdeckungsproblems, die Erfassung des lexikalischen Reichtums eines Textes usw., weil in diesem Fall der Text lemmatisiert werden muss. Bevor man zu zählen anfängt, sollte man diesen Umstand gut überlegen. Nichtsdestoweniger kann man Modelle für Zählungen beider Art entwickeln und testen. Sie gelten dann nur für den gegebenen Aspekt und haben ihre Berechtigung.

(b) Man lemmatisiert zuerst und beobachtet (zählt) nur die Lemmata. Zu diesem Zweck muss man Lemmatisierungsprogramme benutzen, deren Output man noch kontrollieren bzw. für die man den ganzen Text im voraus vorbereiten muss, damit Homographen nicht zusammenfallen. In einigen Sprachen ohne

agglutinierende oder flektierende Morphologie erübrigt sich dies alles, da sämtliche Formen gleichzeitig Lemmata sind. In den indoeuropäischen Sprachen ist dies leider nicht der Fall. Die Lemmatisierung ist keineswegs eine kategorische Angelegenheit, bei der immer klar wäre, zu welchem Lemma eine Wortform gehört. Jede Sprache befindet sich in einem ständigen Fließgleichgewicht, bei dem viele Grenzen nur durch willkürliche Kriterien festgelegt werden können. Es gibt keine „natürlichen" Kriterien. Vielmehr werden Kriterien durch Entscheidungen aufgestellt, sie sind nicht in den Daten enthalten, sie helfen nur, die Daten in eine beabsichtigte Ordnung zu bringen. Auch professionelle Linguisten sollten sich nicht der Illusion hingeben, sie hätten irgendwann die „korrekten" Kriterien ausgewählt. Ihre Kriterien sind genauso konventionell wie Regeln, Begriffe usw., und manchmal haben sie sogar den Status der Bedingungen eines mathematischen Satzes, der mit den Worten: „Sei gegeben …, dann gilt …" beginnt, d.h., unsere Resultate sind nur bei der Erfüllung bestimmter Kriterien valide. So ist es sozusagen Ansichtssache, ob man z.B. *ich*, *mein*, *mir* zu einem Lemma oder zwei Lemmata zusammenfasst, ob *ich habe gerufen* zwei oder drei Lemmata sind oder ob *der, die, das* ein oder drei Lemmata bilden usw. In anderen Sprache, z.B. in stark synthetisierenden, können die Probleme noch komplizierter sein, jedoch ist eine Lemmatisierung immer möglich. Auf diese Variante werden wir hier aber verzichten, da alles, was bisher entwickelt wurde, für die Variante (a) getan wurde. Wenn wir von „Type" sprechen, meinen wir damit „Formtype", wenn wir von „Token" sprechen, meinen wir damit „Formtoken".

Bedenkt man, dass Lemmatisierung eigentlich eine Reduktion der „Types" ist − z.B. gibt es im „Erlkönig" 124 Formentypes, aber nur 96 Lemmata −, dann ergibt sich die berechtigte Frage, ob nicht Pronomina zu einem bestimmten Substantiv gehören und nur als dessen Vertreter behandelt werden sollten. Ebenso berechtigt ist die Frage, ob Präpositionen oder Konjunktionen etwas mit dem Wortschatzreichtum zu tun haben, zumal es Sprachen gibt, die keine Präpositionen haben, und andere, die sie durch Nomina ersetzen. Wortwörtlich in verschiedene Sprachen übersetzt, hätte in der „Erlkönig" jeder einen scheinbar anderen Wortschatzreichtum. Ein mechanisches mathematisches Herangehen an sprachliches Material ist daher mit Vorsicht zu betrachten.

Zählt man im Text aus, wie oft die einzelnen Formen vorkommen, so kann man das Resultat auf verschiedene Weisen darstellen:

(i) alphabetisch oder rückläufig alphabetisch;
(ii) rangiert nach abnehmender Häufigkeit.

Geläufige Programme führen diese Aufgaben mechanisch durch und sind seit langem zu Dutzenden vorhanden. An diese beiden Darstellungen kann man weitere Eigenschaften anfügen, wie beispielsweise die Länge der Formen (gemessen in Buchstaben oder in Silben oder in Morphemen), die Zahl der Bedeutungen des entsprechenden Lemmas, die man aus einem Wörterbuch ermitteln kann, die morphologische Komplexität, usw. Dann kann man die einzelnen Eigenschaften

in Bezug zur Häufigkeit setzen und die Stärke des Zusammenhangs untersuchen. Bei diesem Verfahren entfernt man sich von den einzelnen konkreten Wörtern und deren spezifischen Häufigkeiten und geht über auf die Ebene der Eigenschaften und ihrer Relationen. Weiter kann man gleichzeitig die gegenseitige Beziehung mehrerer Eigenschaften untersuchen, wobei man allmählich auf die Ebene der Textsynergetik übergeht, wo man Regelkreise aufstellen kann, deren Aussagekraft für Einzeltexte noch nicht untersucht wurde. Man weiß bereits, wie die Beziehungen verschiedener Eigenschaften zueinander beschaffen sind, jedoch können sich in einem gegebenen Text andere Parameter, eine andere Kurve, ein andere Attraktor, andere Randbedingungen usw. ergeben, die für ihn charakteristisch sein könnten.

(iii) Während man sich bei der Darstellung in Form (ii) im Bereich des Zipfschen Gesetzes bewegt, führt eine Transformation zur Häufigkeitsverteilung der Wörter, bei der die Einzelwörter selbst keine Rolle mehr spielen; es wird lediglich festgestellt, wie viele Wörter es gibt ($f_x$), die genau $x$-mal vorkommen. Diese Darstellung bezeichnet man als *Häufigkeitsverteilung* (im Unterschied zur Ranghäufigkeitsverteilung) oder als *Häufigkeitsspektrum.*

Man sieht leicht, dass die Zahl der Aspekte beliebig erweitert werden kann, und je mehr Eigenschaften man berücksichtigt, desto mehr Kombinationen (Zusammenhänge) entstehen, desto umfangreicher wird der Regelkreis. Diese Tatsache zeigt, dass die quantitative Erforschung der Texte kaum erst angefangen hat, obwohl stellenweise schon anspruchsvolle mathematische Argumentationen vorliegen (vgl. Hřebíček 1997, 2000; Glottometrics 3, 4, 5, 2002; Baayen 2001; Wimmer et al. 2003 usw.).

### 4.1.1. Die alphabetische Wortliste

Eine Liste dieser Art lässt sich bei kurzen Texten „von Hand“ erstellen, sicherer ist es aber, ein Programm zu benutzen, das eine Häufigkeitesliste erstellt und anschließend die Daten alphabetisch ordnet (vgl. z.B. INTEXT von Klein 1999). Für den „Erlkönig“ erhalten wir so eine alphabetische Liste, die man um weitere Worteigenschaften bereichern kann, z.B. hier Häufigkeit, Länge in Silben, Wortart. Die Abkürzungen in der letzten Spalte geben die Wortart an (A = Adjektiv, Adv = Adverb, Art = Artikel, C = Konjunktion, P = Pronomen, Pr = Präposition, S = Substantiv, V = Verb).

Diese Liste kann man nach Bedarf um weitere gemessene Eigenschaften erweitern, um deren Zusammenhänge zu untersuchen. Sie ist der Ausgangspunkt für alle weiteren Untersuchungen.

Schema (I)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ächzende | 1 | 3 | A | Knaben | 1 | 2 | S |
| alten | 1 | 2 | A | komm | 1 | 1 | V |
| am | 1 | 1 | Pr | Kron | 1 | 1 | S |
| an | 2 | 1 | Pr 1,Adv 1 | Leids | 1 | 1 | S |
| Arm | 1 | 1 | S | leise | 1 | 2 | Adv |
| Armen | 2 | 2 | S | liebe | 1 | 2 | V |
| bang | 1 | 1 | Adv | liebes | 1 | 2 | A |
| birgst | 1 | 1 | V | manch | 2 | 1 | P |
| bist | 1 | 1 | V | mein | 11 | 1 | P |
| Blättern | 1 | 2 | S | meine | 3 | 2 | P |
| bleibe | 1 | 2 | V | mich | 2 | 1 | P |
| Blumen | 1 | 2 | S | mir | 4 | 1 | P |
| brauch | 1 | 1 | V | mit | 6 | 1 | Pr |
| bunte | 1 | 2 | A | Mühe | 1 | 2 | S |
| das | 2 | 1 | Art | Mutter | 1 | 2 | S |
| dein | 2 | 1 | P | Nacht | 1 | 1 | S |
| deine | 1 | 2 | P | nächtlichen | 1 | 3 | A |
| dem | 3 | 1 | Art | Nebelstreif | 1 | 3 | S |
| den | 5 | 1 | Art | nicht | 4 | 1 | Adv |
| der | 2 | 1 | Art | Not | 1 | 1 | S |
| dich | 3 | 1 | P | Ort | 1 | 1 | S |
| die | 1 | 1 | Art | Reihn | 1 | 1 | S |
| dir | 1 | 1 | P | reitet | 2 | 2 | V |
| dort | 1 | 1 | Adv | reizt | 1 | 1 | V |
| du | 7 | 1 | P | ruhig | 2 | 2 | A |
| dürren | 1 | 2 | A | säuselt | 1 | 2 | V |
| düstern | 1 | 2 | A | scheinen | 1 | 2 | V |
| durch | 1 | 1 | Pr | schön | 1 | 1 | Adv |
| ein | 3 | 1 | Art 2,Adv 1 | schöne | 2 | 2 | A |
| er | 5 | 1 | P | Schweif | 1 | 1 | S |
| Erlenkönig | 2 | 4 | S | seh | 1 | 1 | V |
| Erlkönig | 2 | 3 | S | sei | 1 | 1 | V |
| Erlkönigs | 1 | 3 | S | seinem | 1 | 2 | P |
| erreicht | 1 | 2 | V | seinen | 1 | 2 | P |
| es | 5 | 1 | P | sicher | 1 | 2 | Adv |
| faßt | 2 | 1 | V | siehst | 2 | 1 | V |
| feiner | 1 | 2 | A | sind | 1 | 1 | V |
| führen | 1 | 2 | V | singen | 1 | 2 | V |
| gar | 1 | 1 | Adv | so | 4 | 1 | Adv |
| geh | 1 | 1 | V | Sohn | 4 | 1 | S |
| gehn | 1 | 1 | V | sollen | 1 | 2 | V |
| genau | 1 | 2 | Adv | spät | 1 | 1 | Adv |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| geschwind | 1 | 2 | Adv | spiel | 1 | 1 | V |
| Gesicht | 1 | 2 | S | Spiele | 1 | 2 | S |
| Gestalt | 1 | 2 | S | Strand | 1 | 1 | S |
| getan | 1 | 2 | V | tanzen | 1 | 2 | V |
| Gewalt | 1 | 2 | S | Töchter | 3 | 2 | S |
| Gewand | 1 | 2 | S | tot | 1 | 1 | A |
| grau | 1 | 1 | Adv | und | 9 | 1 | C |
| grausets | 1 | 2 | V | Vater | 9 | 2 | S |
| gülden | 1 | 2 | A | verspricht | 1 | 2 | V |
| hält | 2 | 1 | V | war | 1 | 1 | V |
| hat | 3 | 1 | V | warm | 1 | 1 | Adv |
| hörest | 1 | 2 | V | warten | 1 | 1 | V |
| Hof | 1 | 1 | S | was | 2 | 1 | P |
| ich | 4 | 1 | P | Weiden | 1 | 2 | S |
| ihn | 2 | 1 | P | wer | 1 | 1 | P |
| in | 4 | 1 | Pr | wiegen | 1 | 2 | V |
| ist | 2 | 1 | V | willig | 1 | 2 | A |
| jetzt | 1 | 1 | Adv | willst | 1 | 1 | V |
| Kind | 5 | 1 | S | Wind | 2 | 1 | S |
| Knabe | 1 | 2 | S | wohl | 1 | 1 | Adv |

In der Liste finden sich $K$ = 124 Formen (= Vokabular), und die Gesamtzahl aller ihrer Vorkommen (= Textlänge) ist $N$ = 225. Das durchschnittliche Vorkommen der Wörter beträgt $N/K$ = 1.8145. Die letzte Zahl ist gleichzeitig ein Indikator der Formenwiederholung, eine Art elementarer *Wiederholungsrate*. Ihr reziproker Wert, $K/N$ = 0.5511 ist ein selten benutzter Index für das Type-Token-Verhältnis (s.u.). Bei der Interpretation solcher Indizes muss man vorsichtig sein, wenn man sie in einen Zusammenhang mit dem Wortschatzreichtum setzt. Der Index $N/K$ liegt im Intervall $<1,N>$, der größte Wortschatzreichtum wird durch 1, der kleinste druch $N$ signalisiert. Der Index $K/N$ liegt im Intervall $<1/N, 1>$, der größte Vokabularreichtum liegt hier bei 1. Jedoch sind die Werte in diesen Intervallen nicht linear geordnet, daher hat sich die Wortschatzreichtumsforschung von ihnen abgewendet und hat eine ganze Batterie von neuen Ansätzen hervorgebracht (s.u.). Es wird empfohlen, diese einfachen „Inverhältnissetzungen“ nur mit großer Vorsicht zu benutzen.

### 4.1.2. Die Worthäufigkeitsliste

Bei allen Berechnungen ist es vorteilhafter und übersichtlicher, die Wörter nach ihrer Häufigkeit zu rangieren, d.h. aus dem Schema ( I) Schema (II) herzustellen.

Schema (II)

| | | | | | |
|---|---|---|---|---|---|
| mein | 11 | durch | 1 | sollen | 1 |
| und | 9 | Nacht | 1 | warten | 1 |
| Vater | 9 | seinem | 1 | schön | 1 |
| du | 7 | Knaben | 1 | führen | 1 |
| mit | 6 | wohl | 1 | nächtlichen | 1 |
| es | 5 | Arm | 1 | Reihn | 1 |
| Kind | 5 | sicher | 1 | wiegen | 1 |
| er | 5 | warm | 1 | tanzen | 1 |
| den | 5 | birgst | 1 | singen | 1 |
| so | 4 | bang | 1 | dort | 1 |
| in | 4 | dein | 1 | Erlkönigs | 1 |
| Sohn | 4 | Gesicht | 1 | am | 1 |
| nicht | 4 | Kron | 1 | düstern | 1 |
| mir | 4 | Schweif | 1 | Ort | 1 |
| ich | 4 | Nebelstreif | 1 | seh | 1 |
| hat | 3 | liebes | 1 | genau | 1 |
| dem | 3 | komm | 1 | scheinen | 1 |
| ein | 3 | geh | 1 | die | 1 |
| meine | 3 | gar | 1 | alten | 1 |
| Töchter | 3 | Spiele | 1 | Weiden | 1 |
| dich | 3 | spiel | 1 | grau | 1 |
| reitet | 2 | dir | 1 | liebe | 1 |
| Wind | 2 | bunte | 1 | reizt | 1 |
| ist | 2 | Blumen | 1 | deine | 1 |
| der | 2 | sind | 1 | Gestalt | 1 |
| faßt | 2 | Strand | 1 | bist | 1 |
| ihn | 2 | Mutter | 1 | willig | 1 |
| hält | 2 | gülden | 1 | brauch | 1 |
| was | 2 | Gewand | 1 | Gewalt | 1 |
| siehst | 2 | hörest | 1 | jetzt | 1 |
| Erlkönig | 2 | leise | 1 | Leids | 1 |
| Erlenkönig | 2 | verspricht | 1 | getan | 1 |
| schöne | 2 | sei | 1 | grausets | 1 |
| manch | 2 | bleibe | 1 | geschwind | 1 |
| an | 2 | dürren | 1 | ächzende | 1 |
| ruhig | 2 | Blättern | 1 | erreicht | 1 |
| mich | 2 | säuselt | 1 | Hof | 1 |
| Armen | 2 | willst | 1 | Mühe | 1 |
| das | 2 | feiner | 1 | Not | 1 |
| wer | 1 | Knabe | 1 | seinen | 1 |
| spät | 1 | gehn | 1 | war | 1 |
| | | | | tot | 1 |

Der erste Schritt ist immer eine globale Charakterisierung der Verteilung mit bestimmten Maßen. Man kann den Mittelwert, die Varianz, die Schiefe, den Exzess, den Median und andere Maßzahlen benutzen, die ein statistisches Programm automatisch ausgibt. Man sollte sie aber nicht aufführen, wenn man mit ihnen keine weiteren Operationen durchführt. Wir benutzen hier die in Kapitel 2 eingeführten Charakteristika, nämlich die *Wiederholungsrate* nach (2.5), und berechnen aus dem Schema (II)

$$R = [11^2 + 2(9^2) + 7^2 + 6^2 + 4(5^2) + 6(4^2) + 6(3^2) + 18(2^2) + 85(1^1)]/225^2 = 0.0153.$$

Der einfachste theoretische Wert (2.6) ist $R_t = 2/K = 2/124 = 0.0161$, d.h., der empirische Wert liegt etwas unter dem theoretischen. Man kann diese Tatsache auch so interpretieren, dass sich Wörter weniger oft als erwartet wiederholen – eine bekannte Eigenschaft poetischer Texte – so dass der Text einen geringfügig erhöhten Vokabularreichtum hat, was man nur cum grano salis sagen kann.

Ähnlich berechnen wir die *Entropie* nach (2.7) als

$$H = \text{ld } 225 - [11 \text{ ld } 11 + (2)9 \text{ ld } 9 + 7 \text{ ld } 7 + 6 \text{ ld } 6 + (4)5 \text{ ld } 5 + (6)4 \text{ ld } 4 + (6)3 \text{ ld } 3 + (18)2 \text{ ld } 2 + (85)1 \text{ ld } 1]/225 = 7.8138 - 1.2855 = 6.5283.$$

Verglichen mit dem theoretischen Wert, den man mit $K = 124$ aus (2.8) erhält, nämlich $H_t = 6.3968$, ist der empirische Wert etwas größer als der theoretische, was dasselbe sagt wie die Wiederholungsrate. Je größer die Entropie ist, desto schwerer ist es, ein Wort vorauszusagen, weil der Text reicher ist als erwartet. Jedoch setzen wir diese Werte *nicht* in Zusammenhang mit dem Vokabularreichtum, sondern betrachten sie nur als einfache Textcharakteristika.

Diese beiden berechneten Werte sind leider nicht direkt vergleichbar, man kann nicht einmal schätzen, ob 0.0153 bzw. 6.5283 groß oder klein ist. Daher ist es manchmal nützlich, zuerst etwas über ihr Verhalten zu erfahren.

Um die Wiederholungsrate bei Gleichverteilung aller Formen zu berechnen, reicht es zu bedenken, dass dabei alle Häufigkeiten $f_i = N/K$ wären, denn in diesem Falle würde sich jedes Wort (mit Rücksicht auf alle anderen) maximal oft Maße wiederholen. Wir würden

$$(4.1) \quad R_m = \frac{1}{N^2}\sum_{i=1}^{K} f_i^2 = \frac{1}{N^2}\sum_{i=1}^{K}\left(\frac{N}{K}\right)^2 = \frac{1}{N^2} K \frac{N^2}{K^2} = \frac{1}{K},$$

erhalten, was aber das *Minimum von Wiederholungen* aller unterschiedlichen Wörter, die im Text vorkommen, bedeutet[1]. Das Maximum an Wiederholungen

[1] Zu bemerken ist, dass geringe Wiederholbarkeit nicht direkt mit dem Vokabularreichtum zusammenhängt. Für maximalen Vokabularreichtum dürfte sich jedes Wort nur einmal wiederholen, was einen Wert $VR_M = 1/N$ ergeben würde.

wäre erreicht, wenn alle Häufigkeiten in einem Punkt konzentriert wären, d.h., wenn der Text aus der Wiederholung nur eines einzigen Wortes bestünde, so dass

$$(4.2) \quad R_M = \frac{1}{N^2}\sum_{i=1}^{1} f_i^2 = \frac{1}{N^2} N^2 = 1.$$

Daher können wir sagen, dass sich die Wiederholungsrate im Intervall <1/*K*, 1> bewegt. Bei großem *K* (d.h. langen Texten) ist es praktisch (0, 1>. Bei kurzen Texten ist aber 1/*K* als untere Grenze durchaus realistisch.

Die maximale Wiederholbarkeit bedeutet also einen minimalen Wortschatz, weil es ja nur ein einziges Wort gibt, das sich ständig wiederholt. In der Praxis kann man sich so etwas kaum vorstellen. Bei Ranghäufigkeitsverteilungen, die bei Texten meistens einen sehr langen Schweif haben, pflegt die Wiederholungsrate sehr klein zu sein. Daher pflegt man sie entweder zu relativieren (vgl. McIntosh 1967; Altmann 1988) – was bei Ranghäufigkeitsverteilungen auch kaum zum Ziel führt – oder zu normalisieren, was den Vorteil hat, dass man auch ungleich lange Texte miteinander vergleichen kann (vgl. Wimmer et al. 2003). An dieser Stelle zeigen wir nur die Relativierung von McIntosh (1967).

Sind das Minimum und das Maximum der Wiederholungsrate wie oben (Formeln (4.1) und (4.2)) gegeben, dann bilden wir den Index

$$(4.3) \quad R_{rel} = \frac{R_M - \sqrt{R}}{R_M - 1/\sqrt{K}} = \frac{1-\sqrt{R}}{1-1/\sqrt{K}}.$$

Der Sinn dieser Relativierung liegt darin, den Wert von *R*, der bei Rangverteilungen sehr niedrig liegt, etwas „ansehnlicher“ zu machen. Bei entsprechender Behandlung kann man mit ihm alle statistischen Operationen durchführen. Setzen wir die vorhandenen Werte: *K* = 124, *R* = 0.0153 ein, so erhalten wir

$$R_{rel} = \frac{1-\sqrt{0.0153}}{1-1/\sqrt{124}} = \frac{1-0.1237}{1-1/11.1355} = 0.9627.$$

Wir möchten wiederholen, dass man den Wert von $R_{rel}$ nicht in Verbindung mit dem Vokabularreichtum setzen sollte.

Eine andere Relativierung findet man in Wimmer et al. (2003: 130).

Ebenso gut kann man auch die Entropie relativieren, indem man sie durch ihr Maximum dividiert. Das Minimum der Entropie erhalten wir dann, wenn alle Häufigkeiten auf einen einzigen Wert konzentriert sind, d.h.

$$(4.4) \quad H_{\min} = -\sum_{i=1}^{1} p_i \, ld \, p_i = -1 \, ld \, 1 = 0$$

(was dem Maximum der Wiederholungsrate entspricht). Die maximale Entropie ergibt sich dann, wenn alle Wahrscheinlichkeiten gleich sind, nämlich wenn $p_i = 1/K$, d.h.

$$(4.4) \quad H_{\max} = -\sum_{i=1}^{K} \frac{1}{K} ld \frac{1}{K} = ld\ K,$$

so dass als Charakteristikum der Rangverteilung die relative Entropie

$$(4.5) \quad H_{rel} = \frac{H}{H_{\max}}$$

ausreicht. In unserem Fall setzen wir $H = 6.5283$ und $H_{max} = \text{ld } 124 = 6.9542$ und erhalten

$$H_{rel} = 6.5283/6.9542 = 0.9388.$$

Üblicherweise reicht einer dieser Indizes für die Charakterisierung eines Textes aus. Unten werden wir zeigen, wie man mit ihnen beim Textvergleich umgeht. Diese Indizes zeigen bei direktem Vergleich mit ihren erwarteten Werten lediglich, ob der Text der Beziehung zwischen dem Index und der Typezahl (Inventar, Vokabular) folgt oder eine Idiosynkrasie enthält.

### 4.1.3. Die Ranghäufigkeitsverteilung

Ein etwas anspruchsvollerer und umfangreicherer Nachweis des gesetzesartigen Verhaltens der Ranghäufigkeitsverteilung ist die Auffindung der entsprechenden Wahrscheinlichkeitsverteilung. Dieses Gebiet ist der bestentwickelte Bereich der quantitativen Linguistik, nicht nur weil er der älteste ist – er fing schon bei Estoup (1916), Condon (1928) und besonders Zipf (1935) an – sondern auch, weil sich wegen der leichten Zugänglichkeit der Daten auch Mathematiker und Physiker an der Entwicklung beteiligt haben und besonders, weil man in fast allen wissenschaftlichen Disziplinen Analogien zum Zipfschen Gesetz gefunden hat. Bei der Entwicklung der Formeln gerieten die Forscher in zahlreiche Fallen, wie sie uns die Natur immer wieder stellt (wie Einstein sagte: „Der liebe Gott ist nicht bösartig, sondern nur raffiniert.“). Einige Fallen sind z.B. die Annahme, dass ein Modell für alle Texte aller Sprachen gelten müsse, oder der Glaube, dass man Wörter streng in Hilfswörter und Bedeutungswörter trennen könne, oder die Annahme, dass das Wörterbuch einer Sprache endlich oder unendlich sei, dass Ränge nur eine sekundäre Variable darstellten, die aus Häufigkeiten gewonnen worden seien und daher nichts bedeuteten, oder dass es ausreiche, mehrere Mo-

delle einfach miteinander zu multiplizieren und nach Bedarf eine oder mehrere Komponenten gleich 1 zu setzen, usw. Es gibt in dieser Hinsicht merkwürdigerweise weniger Streit zwischen Linguisten und Mathematikern als zwischen Mathematikern untereinander, die die Anfangsbedingungen entweder ignorieren oder der Meinung sind, dass das Englische „die“ Sprache sei.

Auch wenn es heutzutage etwa 50 Modelle für Ranghäufigkeitsverteilungen gibt und zahlreiche von ihnen sich mit einem übergreifenden rekursiven Ansatz erfassen lassen (vgl. Wimmer, Altmann 2005), geht die Entwicklung ungehindert weiter. Es ist nur zu bedauern, dass die Linguisten sich kaum daran beteiligen, die Randbedingungen auszuarbeiten, um den Mathematikern die Arbeit zu erleichtern.

Wir werden uns hier an bewährte Resultate halten und mit Hilfe einer Software die sog. Zipf-Mandelbrot-Verteilung (vgl. Formel (3.1)) anpassen:

$$(4.6) \quad P_x = \frac{K}{(x+b)^a}, \quad x = 1,2,3,...,n$$

wobei *K* die Normierungskonstante und *a, b, n* Parameter sind. Es lässt sich zeigen, dass hier zahlreiche Verteilungen erfolgreich passen, man fängt aber meistens mit der Zipf-Mandelbrot-Verteilung an, die sich in den meisten Fällen bewährt hat. Wir möchten darauf aufmerksam machen, dass wir die Zipf-Mandelbrot-Verteilung auf der rechten Seite gestutzt haben. Die Parameter *a* und *b* sind für den gegebenen Text charakteristisch, aber erst der Vergleich vieler Texte könnte zu einer komparativen Bewertung führen (vgl. Orlov, Boroda, Nadarejšvili 1982). Sicher ist, dass sie in einem noch nicht erforschten Zusammenhang zu anderen Texteigenschaften stehen, aber trotz einfacher Verwendung dieser Verteilung (vgl. Guiter, Arapov 1982; Chitashvili, Baayen 1993; Baayen 2001; Glottometrics 3, 4, 5, 2002) ist man hier nicht weitergekommen. An dieser Stelle reicht der Verweis auf die reichlich vorhandene Literatur; s. auch das umfangreiche Literaturverzeichnis des mathematischen Biologen Wentian Li in http://www.nslij-genetics.org/wli/zipf.

Zu bemerken ist, dass für diese einfachen Daten viele andere Verteilung, sogar solche mit nur einem Parameter, gut passen würden (z.B. Prasad, Johnson-Kotz, geometrische, Zeta s. Wimmer, Altmann 1999a).

Tabelle 4.1
Ranghäufigkeitsverteilung der Wortformen im „Erlkönig“

| x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 9.89 | 32 | 2 | 1.97 | 63 | 1 | 1.20 | 94 | 1 | 0.89 |
| 2 | 9 | 8.48 | 33 | 2 | 1.93 | 64 | 1 | 1.19 | 95 | 1 | 0.88 |
| 3 | 9 | 7.46 | 34 | 2 | 1.89 | 65 | 1 | 1.17 | 96 | 1 | 0.87 |
| 4 | 7 | 6.68 | 35 | 2 | 1.85 | 66 | 1 | 1.16 | 97 | 1 | 0.87 |
| 5 | 6 | 6.06 | 36 | 2 | 1.81 | 67 | 1 | 1.15 | 98 | 1 | 0.86 |
| 6 | 5 | 5.56 | 37 | 2 | 1.78 | 68 | 1 | 1.13 | 99 | 1 | 0.85 |
| 7 | 5 | 5.14 | 38 | 2 | 1.74 | 69 | 1 | 1.12 | 100 | 1 | 0.85 |
| 8 | 5 | 4.79 | 39 | 2 | 1.71 | 70 | 1 | 1.11 | 101 | 1 | 0.84 |
| 9 | 5 | 4.49 | 40 | 1 | 1.68 | 71 | 1 | 1.10 | 102 | 1 | 0.83 |
| 10 | 4 | 4.22 | 41 | 1 | 1.65 | 72 | 1 | 1.09 | 103 | 1 | 0.83 |
| 11 | 4 | 3.99 | 42 | 1 | 1.62 | 73 | 1 | 1.08 | 104 | 1 | 0.82 |
| 12 | 4 | 3.79 | 43 | 1 | 1.59 | 74 | 1 | 1.06 | 105 | 1 | 0.82 |
| 13 | 4 | 3.61 | 44 | 1 | 1.57 | 75 | 1 | 1.05 | 106 | 1 | 0.81 |
| 14 | 4 | 3.45 | 45 | 1 | 1.54 | 76 | 1 | 1.04 | 107 | 1 | 0.80 |
| 15 | 4 | 3.30 | 46 | 1 | 1.52 | 77 | 1 | 1.03 | 108 | 1 | 0.80 |
| 16 | 3 | 3.17 | 47 | 1 | 1.49 | 78 | 1 | 1.02 | 109 | 1 | 0.79 |
| 17 | 3 | 3.04 | 48 | 1 | 1.47 | 79 | 1 | 1.01 | 110 | 1 | 0.79 |
| 18 | 3 | 2.93 | 49 | 1 | 1.45 | 80 | 1 | 1.00 | 111 | 1 | 0.78 |
| 19 | 3 | 2.83 | 50 | 1 | 1.43 | 81 | 1 | 0.99 | 112 | 1 | 0.78 |
| 20 | 3 | 2.73 | 51 | 1 | 1.41 | 82 | 1 | 0.98 | 113 | 1 | 0.77 |
| 21 | 3 | 2.64 | 52 | 1 | 1.39 | 83 | 1 | 0.98 | 114 | 1 | 0.77 |
| 22 | 2 | 2.56 | 53 | 1 | 1.37 | 84 | 1 | 0.97 | 115 | 1 | 0.76 |
| 23 | 2 | 2.49 | 54 | 1 | 1.35 | 85 | 1 | 0.96 | 116 | 1 | 0.76 |
| 24 | 2 | 2.41 | 55 | 1 | 1.33 | 86 | 1 | 0.95 | 117 | 1 | 0.75 |
| 25 | 2 | 2.35 | 56 | 1 | 1.31 | 87 | 1 | 0.94 | 118 | 1 | 0.75 |
| 26 | 2 | 2.28 | 57 | 1 | 1.29 | 88 | 1 | 0.93 | 119 | 1 | 0.74 |
| 27 | 2 | 2.23 | 58 | 1 | 1.28 | 89 | 1 | 0.93 | 120 | 1 | 0.74 |
| 28 | 2 | 2.17 | 59 | 1 | 1.26 | 90 | 1 | 0.92 | 121 | 1 | 0.73 |
| 29 | 2 | 2.12 | 60 | 1 | 1.25 | 91 | 1 | 0.91 | 122 | 1 | 0.73 |
| 30 | 2 | 2.07 | 61 | 1 | 1.23 | 92 | 1 | 0.90 | 123 | 1 | 0.72 |
| 31 | 2 | 2.02 | 62 | 1 | 1.21 | 93 | 1 | 0.89 | 124 | 1 | 0.72 |

a = 0.7941, b = 3.6880, n = 124, $X^2$ = 6.66, FG = 98, P ≈ 1.00

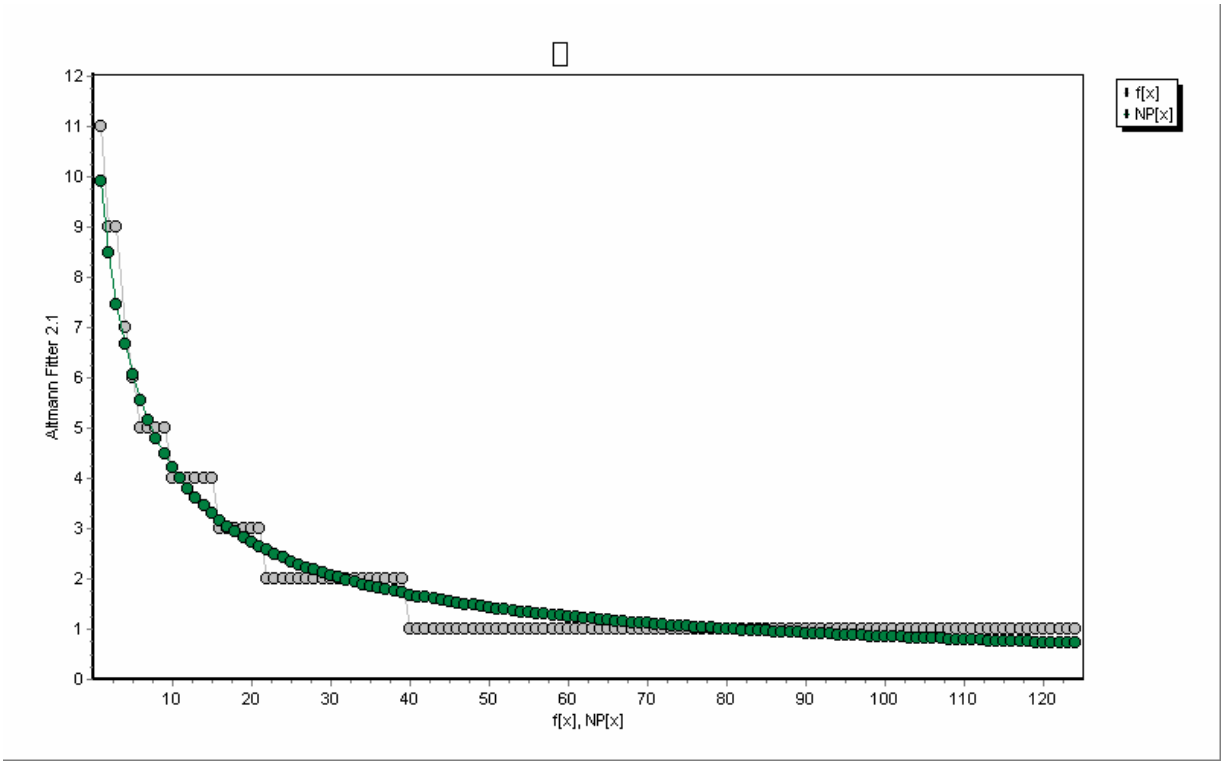


Abbildung 4.1. Ranghäufigkeitsverteilung der Wörter im „Erlkönig“ (Anpassung der Zipf-Mandelbrot-Verteilung)

### 4.1.4. Die Häufigkeitsverteilung (Häufigkeitsspektrum)

Die obige Ranghäufigkeitsverteilung kann man in dem Sinne „umdrehen“, dass man von unten anfangend zusammenzählt, wie viele Wörter $x$ = 1-mal vorkommen. Diese sind in Tabelle 4.1 oder im Schema (II) alle Wörter auf den Rängen 40 bis 124, d.h. $f_1$ = 85; dann zählt man die Wörter, die jeweils $x$ = 2-mal vorkommen. Man stellt leicht fest, dass es $f_2$ = 18 sind. Zählt man weiter, so bekommt man letztlich die Werte, die in den zwei ersten Spalten von Tabelle 4.2 dargestellt sind.

Wie sich mathematisch zeigen lässt (vgl. Boroda, Zörnig 1990), führt eine derartige Transformation wieder zu der Zipf-Mandelbrot-Verteilung mit neuen Parametern, die auch durch Transformation der Parameter der Ranghäufigkeitsverteilung entstehen. Da wir aber die Anpassungen iterativ durchführen, berechnen wir die neuen Parameter direkt und erhalten die Werte in der dritten Spalte der Tabelle 4.2, graphisch in Abbildung 4.2 dargestellt. Zu bemerken ist, dass nicht jedes Modell der Ranghäufigkeitsverteilung durch diese Transformation wieder sich selbst ergibt; und das Resultat der Transformation hängt auch davon ab, welche Transformation man vornimmt.

Tabelle 4.2
Häufigkeitsverteilung der Wörter im „Erlkönig“
(und die Anpassung der Zipf-Mandelbrot-Verteilung)

| $x$ | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 85 | 81.13 | 5 | 4 | 2.79 | 9 | 2 | 0.71 |
| 2 | 18 | 20.90 | 6 | 1 | 1.84 | 10 | 0 | 0.56 |
| 3 | 6 | 8.78 | 7 | 1 | 1.28 | 11 | 1 | 0.44 |
| 4 | 6 | 4.63 | 8 | 0 | 0.94 | | | |
| a = 2.4451, b = 0.2485, $X^2$ = 2.91, FG = 5, P = 0.71 | | | | | | | | |

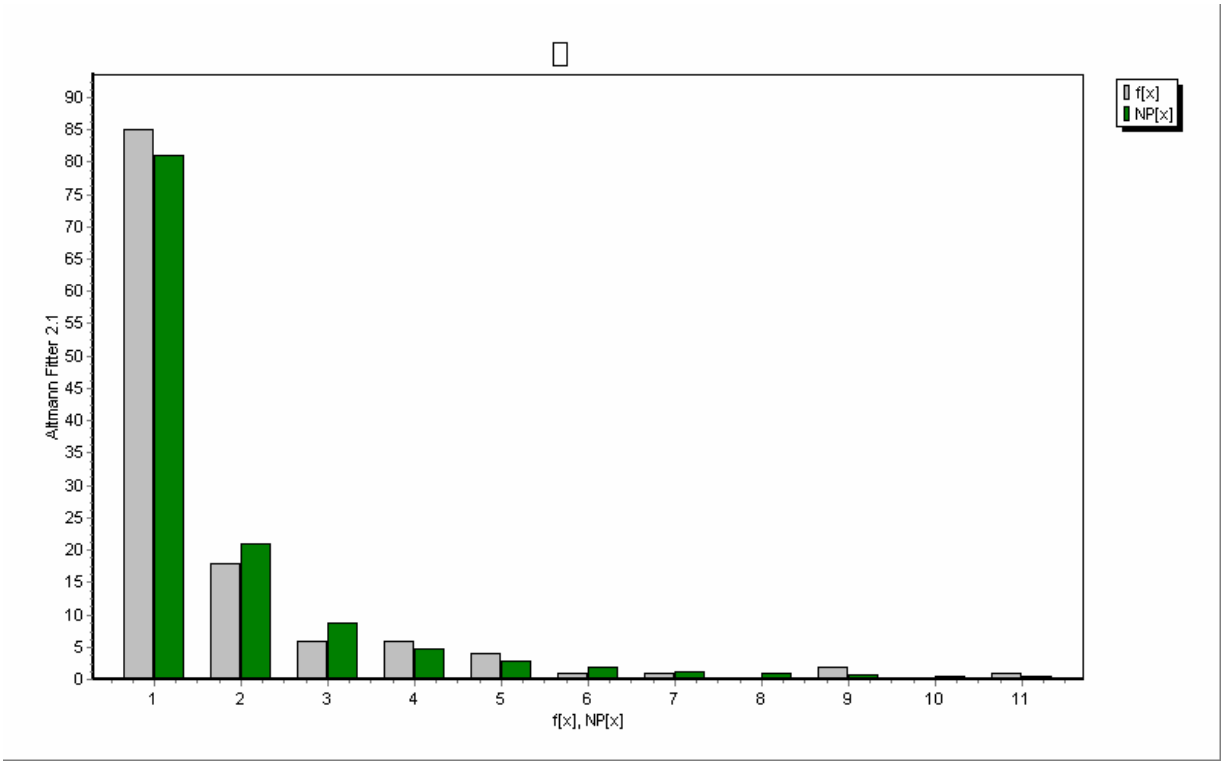


Abbildung 4.2. Anpassung der Zipf-Mandelbrot-Verteilung an die Worthäufigkeiten im „Erlkönig“

Auch bei den neuen Parametern ist deren Interpretation vorläufig unbekannt. Anpassungen dieser Art führt man daher eher zur Unterstützung der Theorie als zur Charakterisierung der Texte durch. Jeder Fall gibt uns aber mehr Vergleichsmöglichkeiten und öffnet Wege zur Erforschung von Zusammenhängen. Unser aktuelles Resultat zeigt nur, dass Wortformen gesetzesartig verteilt sind.

Bei eventueller Charakterisierung kann man die Parameter *a* und *b* als Koordinaten eines Punktes <*a*, *b*> benutzen und in einen Graphen eintragen, aber ohne Vergleichsinstanzen nutzt dieses Verfahren vorläufig nicht viel.

An dieser Stelle werden wir mit der Analyse der Wiederholungsrate und der Entropie einen weiteren Schritt tun, denn in diesem Fall sind diese Größen etwas angemessener als bei der Ranghäufigkeitsverteilung.

Sei die Wiederholungsrate definiert als

$$(4.7) \quad R = \sum_{i=1}^{K} \hat{p}_i^2$$

wobei wir $\hat{p}_i$ mit den relativen Häufigkeiten $f_i /N$ schätzen (vgl. 2.6). Unsere Aufgabe besteht darin, $R$ zu normalisieren, d.h. auf die Normalvariable zu transformieren, um

$$(4.8) \quad u = \frac{R - E(R)}{\sqrt{Var(R)}}$$

zu erhalten, wobei dann $u \sim N(0,1)$. Das resultierende $u$ ist ein Quantil der Normalverteilung, das uns zeigt, wie weit entfernt der gemessene R-Wert von der theoretischen Kurve liegt. Zu diesem Zweck müssen wir $E(R)$ und $Var(R)$ berechnen. Da die $p_i$-s in (4.7) multinomial verteilte Zufallsvariablen sind, die, empirisch gesehen, miteinander nach Umordnung im einfachsten Fall geometrisch angeordnet sind, erhalten wir, wie bereits bekannt, $E(R) = 2/K$ (vgl. Altmann, Lehfeldt 1980). Die Varianz berechnen wir durch Taylorentwicklung auf die übliche Weise (vgl. Kendall, Stuart 1967) aus

$$(4.9) \quad Var(R) = \sum_{i=1}^{K} \left( \frac{\partial R}{\partial \hat{p}_i} \right)_{p_i}^2 Var(\hat{p}_i) + \sum_{i=1}^{K} \sum_{j \neq i} \left( \frac{\partial R}{\partial \hat{p}_i} \right)_{p_i}^2 \left( \frac{\partial R}{\partial \hat{p}_j} \right)_{p_j}^2 Cov(\hat{p}_i, \hat{p}_j).$$

Da in der Multinomialverteilung

$$Var(\hat{p}_i) = \frac{p_i(1 - p_i)}{N}$$

und

$$Cov(\hat{p}_i, \hat{p}_j) = -\frac{p_i p_j}{N}$$

und die $p_i$ durch relative Häufigkeiten geschätzt werden, ergibt sich aus (4.9)

$$(4.10) \quad Var(R) = \frac{1}{V} \sum_{i=1}^{k} \left( 2\frac{f_i}{V} \right)^2 \frac{f_i}{V} \left( 1 - \frac{f_i}{V} \right) - \frac{1}{V} \sum_{i=1}^{k} \sum_{j \neq i} \left( 2\frac{f_i}{V} \right) \left( 2\frac{f_j}{V} \right) \frac{f_i f_j}{V}$$

$$= \frac{4}{V}\left[\sum_{i=1}^{k}\left(\frac{f_i}{V}\right)^3 - \left(\sum_{i=1}^{k}\left(\frac{f_i}{V}\right)^2\right)^2\right]$$

$$= \frac{4}{V}\left[\sum_{i=1}^{k}\hat{p}_i^3 - \hat{R}^2\right].$$

Hier ist *k* die Zahl der Häufigkeitsklassen, praktisch $x_{max}$ (hier 11), und *V* ist die Zahl der unterschiedlichen Formentypen, d.h. das Vokabular des Textes (hier 124).

Um (4.10) zu berechnen, brauchen wir

$$\sum_{i=1}^{k}\hat{p}_i^3 = \sum_{i=1}^{k}\left(\frac{f_i}{V}\right)^3 = \frac{85^3 + 18^3 + 2(6^3) + 4^3 + 2^3 + 3(1^3)}{124^3} = 0.325425$$

$$\sum_{i=1}^{k}\hat{p}_i^2 = \sum_{i=1}^{k}\left(\frac{f_i}{V}\right)^2 = \frac{85^2 + 18^2 + 2(6^2) + 4^2 + 2^2 + 3(1^2)}{124^2} = 0.497139$$

Setzen wir die entsprechenden Werte in (4.10) ein, so erhalten wir

$$Var(R) = \frac{4}{124}\left[0.3254 - 0.4971^2\right] = 0.0025255.$$

Jetzt setzen wir alle notwendigen Werte in (4.8) ein und bekommen

$$u = \frac{0.4971 - 2/124}{\sqrt{0.0025255}} = 9.57.$$

Dieser Wert stellt die normierte Abweichung der Wiederholungsrate von dem Erwartungswert dar. Die Abweichung ist hoch signifikant, und man kann davon ausgehen, dass der Text eine Idiosynkrasie enthält, die mit der geometrischen Verteilung – aus der $R = 2/K$ abgeleitet wurde – nicht harmoniert, und dies ist in der Tat der Fall. Aber auch andere Verteilungen liefern sehr niedrige Erwartungswerte, die mit dieser Wiederholungsrate nicht übereinstimmen. Dies ist ein Anlass, die Theorie erneut zu prüfen.

Formel (4.8) ermöglicht es uns, mit Hilfe eines Konfidenzintervalls zu bestimmen, wie etwa der erwartete Wert für *R* aussehen sollte. Mit einer Wahrscheinlichkeit etwa 0.95 kann man das Intervall wie folgt aufstellen

(4.11) $\hat{R} - 1.96\sqrt{Var(\hat{R})} \le R \le \hat{R} + 1.96\sqrt{Var(\hat{R})}$,

wobei $\hat{R}$ die beobachtete Wiederholungsrate bedeutet. In unserem Fall bekommen wir

$$0.4971 - 1.96\sqrt{0.0025255} \le R \le 0.4971 + 1.96\sqrt{0.0025255} = (0.3986, 0.5956)$$

Um zwei Texte aus dieser Sicht miteinander zu vergleichen, benutzt man das gleiche Kriterium und berechnet

(4.12) $$u = \frac{R_1 - 2/K_1 - (R_2 - 2/K_2)}{\sqrt{Var(R_1) + Var(R_2)}},$$

wo die Indizes 1 und 2 zwei unterschiedliche Texte bedeuten.

Im Allgemeinen eignet sich die Wiederholungsrate nicht immer gut für die Charakterisierung der Verteilungen, jedoch gut für den *Vergleich von zwei Verteilungen* (vgl. Ziegler, Altmann 2002: 52 ff.; Wimmer et al. 2003: 128f.), wofür es aber auch andere Methoden gibt.

Ähnlich kann man mit der Entropie umgehen. Mit der gleichen Methode der Taylorentwicklung finden wir, dass

(4.13) $$Var(\hat{H}) = \frac{1}{V}\left[\sum_i \frac{f_i}{V} ld^2 \frac{f_i}{V} - \hat{H}^2\right],$$

und in unserem Fall berechnen wir (bei der Zerlegung Vorsicht mit den Logarithmen!) wir mit $V = 124$

$$H = ld\, V - \frac{1}{V}\sum_i f_i\, ld\ f_i = \text{ld}(124) - [85 \text{ ld } 85 + 18 \text{ ld } 18 + 2(6) \text{ ld } 6 +$$
$$+ 4 \text{ ld } 4 + 2 \text{ ld } 2 + 3(1) \text{ ld } 1]/124 = 1.624548$$

$$\sum_i \frac{f_i}{V} ld^2 \frac{f_i}{V} = \frac{1}{V}\sum_i f_i\ ld^2\ f_i - ld^2 V + 2H\ ld\ V =$$

$$= [85 \text{ ld}^2 85 + 18 \text{ ld}^2 18 + 2(6) \text{ ld}^2 6 + 4 \text{ ld}^2 4 + 2 \text{ ld}^2 2 +$$
$$+ 3(1) \text{ ld}^2 1]/124 - 48.360846 + 2(1.624548)6.954196 =$$
$$= 5.70979222.$$

Damit ergibt sich

$Var(H) = (5.70979222 - 1.624548^2)/124 = 0.02476.$

Mit Hilfe dieses Wertes können wir wieder Konfidenzintervalle aufstellen, nämlich

(4.14) $\hat{H} - 1.96\sqrt{Var(\hat{H})} \leq H \leq \hat{H} + 1.96\sqrt{Var(\hat{H})}$,

was in unserem Fall ergeben würde

$1.624548 - 1.96(0.1574) \leq H \leq 1.624548 + 1.96(0.1574) = (1.32,\ 1.93).$

Bei der Charakterisierung eines Textes ist es ratsam solche Intervalle aufzustellen, da sich bei unterschiedlicher Interpretation dessen, was dasselbe Wort ist, Schwankungen ergeben können.

Analog zu (4.12) kann man auch den Unterschied zweier Texte testen. Hier kann man vorläufig von zwei theoretischen Erwartungswerten ausgehen, nämlich aufgrund der geometrischen Verteilung (cf. Altmann, Lehfeldt 1980)

(4.15) $E(H) = -ld\left[\left(\frac{4}{K+2}\right)\left(\frac{K-2}{K+2}\right)^{\frac{K-2}{4}}\right]$

und aufgrund der Zipf-Mandelbrot-Verteilung (cf. Zörnig, Altmann 1984)

(4.16) $E(H) = ld\, e \ln\left(\sqrt{(B+K)(B+1)}\ln\frac{B+K}{B+1}\right)$

wo $K$ die Zahl der Kategorien (Inventargröße, Types) ist. $B$ ist üblicherweise 0.61. Beide Kurven haben einen sehr ähnlichen Verlauf.

### 4.1.5. Textabdeckung

Viele Arbeiten widmen sich der sog. Textabdeckung, d.h. der Abschätzung dessen, welchen Teil (Proportion) eines Textes die Wörter bis zum Rang $x$ abdecken. Meistens bleibt man bei der empirischen Feststellung stehen, die man ganz einfach mit Hilfe der kumulativen Verteilung (Verteilungsfunktion) erfasst. Beispielsweise haben wir in Tabelle 4.1 (mit $P_x = \sum_{j=1}^{x} f_j / N$)

| $x$ | $f_x$ | $\sum_{j=1}^{x} f_j$ | $P_x$ |
|---|---|---|---|
| 1 | 11 | 11 | 0.0489 |
| 2 | 9 | 20 | 0.0889 |
| 3 | 9 | 29 | 0.1289 |
| 4 | 7 | 36 | 0.1600 |

usw., d.h., die ersten drei häufigsten Wörter decken 12.89 Prozent des Textes ab, die ersten 4 16.00% usw. Empirisch gesehen gilt: je flacher die *Ranghäufigkeitsverteilung*, desto größer der Vokabularreichtum, je steiler, desto ärmer der Text. Da man die Steilheit einer Verteilung mit dem Koeffizienten des Exzesses messen kann, d.h. als

$$(4.17)\quad \gamma = \frac{m_4}{m_2^2} - 3,$$

wo

$$(4.18)\quad m_r = \frac{1}{N}\sum_{i=1}^{V}(x_i - \bar{x})^r \; f_x,$$

gilt: je kleiner γ, desto größer der Vokabularreichtum. In (4.17) wird der empirische Exzess mit dem der Normalverteilung verglichen, der 3 beträgt. Offenbar ist der Exzess der Ranghäufigkeitsverteilung eines der wenigen adäquaten Maße des Vokabularreichtums. Da uns das Program FITTER bei der mechanischen Anpassung der Zipf-Mandelbrot-Verteilung (oder jeder anderen) automatisch die empirischen Momente liefert, erhalten wir für den „Erlkönig“

$$\gamma = \frac{4402011.0843}{1384.6025^2} - 3 = -0.70,$$

was einen erhöhten Reichtum signalisiert, da $\gamma < 0$. Die Kurve ist nämlich flacher als die Normalverteilung, es gibt im Text wenige häufig und viele einmalig vorkommende Wörter.

Vergleiche mit anderen Texten lassen sich durch Transformation auf die Normalverteilung durchführen, d.h. als

$$(4.19)\quad u = \frac{\gamma_1 - \gamma_2}{\sqrt{Var(\gamma_1) + Var(\gamma_2)}},$$

wobei man die Varianz $Var(\gamma)$ in Kendall, Stuart (1967) findet.

Bei der *Häufigkeitsverteilung* (Tab. 4.2) ist die Interpretation bezüglich des Vokabularreichtums genau umgekehrt. Je steiler die Verteilung, desto größer der Vokabularreichtum, denn je mehr einmalig vorkommende Wörter es gibt, desto reicher ist der Text. Hier bekommen wir für den „Erlkönig“

$$\gamma = \frac{113.5350}{2.9575^2} - 3 = 9.98$$

Dieses Resultat sollte mit denen für andere Texten verglichen werden.

Theoretisch gesehen, ist die Verteilungskurve $F_x$ eine monoton wachsende Kurve (hier eine Treppenkurve), die davon abhängt, welche theoretische Verteilung ihr zugrundeliegt. Wie bereits erwähnt, ist die Zipf-Mandelbrot-Verteilung nicht die einzig mögliche (vgl. Chitashvili, Baayen 1983). Benutzt man sie aber, so erhält man (vgl. Arapov 1981; Haitun 1983; Tuldava 1998)

$$(4.20)\quad F_x = \sum_{j=1}^{x} P_j = \sum_{j=1}^{x} \frac{K}{(j+b)^a}, \quad x = 1, 2, ..., n.$$

Die Summe in (4.20) wird approximiert mit einem Integral als

$$(4.21)\quad F_x \approx \int_1^x \frac{K}{(j+b)^a} dj = \frac{K}{1-a}\left[(x+b)^{1-a} - (1+b)^{1-a}\right].$$

Dies mag in vielen Fällen eine gute Approximation sein, führt aber trotzdem nicht immer zu guten Resultaten. Daher ist es ratsamer, zuerst in zahlreichen Texten γ zu berechnen, um überhaupt die empirische Variabilität dieser Größe zu eruieren und dadurch Texte zu charakterisieren. Möglicherweise ergibt sich dann auch für die Textabdeckung eine einfachere Kurve.

## 4.2. Wortarten

Eine Gruppierung der Wörter in eine kleinere Anzahl von nominalen (kategorischen) Klassen ist immer möglich. Oben (Abschnitt 4.1.1) haben wir nur die Wortart angegeben, es gibt aber zahlreiche andere Möglichkeiten, die aus einem Forschungsziel, aus einer Hypothese herrühren können. Die klassische Wortart ist deswegen interessant, weil man aus einer Zusammenstellung der Wortarten des Textes auf bestimmte Eigenschaften von ihm schließen kann. Die häufige Verwendung von Adjektiven und bestimmten Adverbien signalisieren einen ornamentalen Stil, wie er der Lyrik eigen ist; die häufige Verwendung von „aktiven“ Verben deutet eher Epik an; wenige Pronomina weisen auf Deskriptivität

hin, viele Personalpronomina auf Dramatik, alle Pronomina zusammen weisen auf eine bestimmte Diskurs-Gestaltung hin. Die Häufigkeit von Konjunktionen, die die Nominal- oder die Verbalphrasen oder ganze Sätze komplexer machen, lässt sich auf verschiedene Arten interpretieren: als Ausdruck einer Nichtlinearität des Geschehens, als Ornamentalität, als Spezifizität formaler Texte (amtlicher Stil, Wissenschaftstexte), z.B. als Deskriptivität; Zahlwörter sind Zeichen der Deskriptivität; Hilfswörter als ganze nutzt man oft in der forensischen Forschung, bei Autorschaftsproblemen usw. (vgl. Carroll 1969). Wortartenhäufigkeiten wurden am meisten für die Unterscheidung von Gattungen und die Charakterisierung von Autoren verwendet (vgl. Kločkova 1968; Jakubaitis 1981; Tiščenko 1987; Levickij, Hikow 2004).

Unter den genannten Aspekten sind die Wortarten bereits gut erforscht worden, das Problem verdient aber noch mehr Aufmerksamkeit und bietet gute Perspektiven. Bei der Aufstellung von Indizes muss zur Vorsicht geraten werden (vgl. Altmann 1978). Man sollte dabei auch bedenken, dass keine Klassifikation des sprachlichen Materials ganz scharf sein kann, dass immer Grenzfälle übrigbleiben, bei denen man sich nicht recht entscheiden kann, zu welcher Gruppe sie gehören sollen. So kann man z.B. „ächzende“ entweder den Verben oder den Adjektiven, „ächzend“ entweder den Verben oder den Adverbien zuordnen. Linguistische Kriterien helfen nicht immer, weil sie oft „schulbedingt“ sind. Die Zählung der Wortarten kann sich von derjenigen der Wörter unterscheiden, wie man in Schema (I) sieht. Das Wort „ein“ kann als Artikel, als Zahlwort oder in Form eines abgetrennten Präfixes als Adverb klassifiziert werden. So ist es im „Erlkönig“ der Fall in „…singen dich ein“ (Vers 20), oder „an“ in „…faßt er mich an“ (Vers 27), die analog zu „warm“ in „…hält ihn warm“ (Vers 4) behandelt werden. Ein anderes Problem ist die Vergleichbarkeit der Texte in dieser Hinsicht. In den Grammatiken der europäischen Sprachen hält man sich üblicherweise an die lateinischen Wortarten, syntaktische Kriterien liefern aber andere Resultate, und bei der Beschreibung anderer Sprachen gerät man ständig in Konflikte mit den Wortarten europäischer Sprachen.

Man kann aber ein externes Kriterium benutzen, das von nichtgrammatischen und nichtsemantischen Eigenschaften, d.h. grammatikunabhängig aufgestellt wird: Eine Klassifikation sprachlicher Entitäten ist dann effektiv, wenn die empirische Ranghäufigkeitsverteilung der Klassen der Zipf-Mandelbrot-Verteilung (oder einer anderen Rangverteilung) folgt. Es ist nämlich schwer sich vorzustellen, dass in einem „normalen“ Text chaotische Verhältnisse herrschen sollten. Die Hervorhebung einer Eigenschaft kann natürlich Verschiebungen der Häufigkeiten einzelner Klassen verursachen. Beispielsweise können Adjektive bei zu großer Ornamentalität des Textes auf die rangerste Stelle gesetzt werden, aber in diesem Fall werden die anderen Wortarten so proportioniert, dass die Verhältnisse konstant bleiben und weiterhin der Formel entsprechen. Man vergleicht dann nicht die Einteilung der Wortarten, sondern lediglich die Parameter der resultierenden Verteilung, d.h., unabhängig davon, wie die Einteilung in

Wortarten beschaffen gewesen sein mag, vergleicht man auf einer darüberliegenden höheren Abstraktionsebene.

Die Auszählung der Wortarten in Schema (I) ergibt die Häufigkeiten, die in Tabelle 4.3 dargestellt sind. Die Anpassung der Zipf-Mandelbrot-Verteilung ist in der dritten Spalte aufgeführt. Wie man sieht, ist sie akzeptabel, auch wenn es sicherlich noch bessere Anpassungen gibt. Zahleiche Autoren benutzen eher die sich immer mehr durchsetzende negative hypergeometrische Verteilung oder die Zipf-Alekseev-Verteilung, die Poisson- und die Hyperpoisson-Verteilung (vgl. Best 1994, 1997, 2000, 2001a, 2003; Hammerl 1990; Judt 1995; Schweers, Zhu 1991; Tuldava 1987; Ziegler 1998, 2001; Ziegler, Best, Altmann 2001).

Tabelle 4.3
Verteilung der Wortarten im „Erlkönig“

| Wortart | x | $f_x$ | $NP_x$ |
|---|---|---|---|
| Pronomen | 1 | 56 | 64.67 |
| Substantiv | 2 | 53 | 46.27 |
| Verb | 3 | 41 | 33.87 |
| Adverb | 4 | 23 | 25.30 |
| Artikel | 5 | 15 | 19.23 |
| Adjektiv | 6 | 15 | 14.84 |
| Präposition | 7 | 13 | 11.62 |
| Konjunktion | 8 | 9 | 9.21 |
| a = 4.5701, b = 12.1579, $X^2_4$ = 4.95, P = 0.29 | | | |

Man kann das Häufigkeitsspektrum der Wortarten auch in Form eines Vektors darstellen und für Vergleichszwecke benutzen. Ohne Rangordnung hätten wir es mit der Multinomialverteilung zu tun. Man definiert z.B.

(4.22) $V_T = [S, V, A, P, Adv, Pr, K, Art, Part, Num, I]/N,$

wobei $N$ die Gesamtsumme aller Wortarten ist. Die Kategorien kann man sicherlich auch anders darstellen, wenn man z.B. syntaktische Kriterien berücksichtigt. Man kann auch Matrizen benutzen, wenn man z.B. für *S* alle Deklinationskategorien, für *V* alle Konjugationsformen usw. in Betracht zieht, jedoch lohnt sich dies nur in längeren Texten. Sind die Kategorien in zwei Texten oder zwei Sprachen gleich definiert, dann kann man den Unterschied in Form eines Abstandsmaßes messen, z.B. als Euklidische Distanz zwischen Texten.

Bei den Wortarten kann man im Grunde alle Probleme bearbeiten, die man bei anderen Texteinheiten zu beschreiben pflegt, wie Übergangswahrscheinlichkeiten, Distanzen im Text, Entropie, Entwicklung des obigen Vektors im Verlauf der Strophen, mögliche Sprünge u.a. (vgl. Wimmer, Altmann 2001a;

Wimmer et al. 2003; Ziegler, Best, Altmann 2001; Levickij, Hikow 2004). Wir werden hier nur einige ausgewählte Probleme behandeln, die man ohne Vergleich mit anderen Texten bearbeiten kann.

### 4.2.1. Das Spektrum der Wortarten

Für den „Erlkönig“ stellen wir den Vektor (4.22) mit Hilfe der Zahlen aus Tabelle 4.3 folgendermaßen auf (alle Zahlen bezogen auf 225)

| | S | V | A | P | Adv | Pr | K | Art | Part | Num | Int |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $V_{Erl}$ = | 0.2356, | 0.1822, | 0.0667, | 0.2489, | 0.1022, | 0.0578, | 0.04, | 0.0667, | 0, | 0 , | 0]. |

Der Vektor ist für den Text charakteristisch, sagt aber ohne Vergleich wenig. Man kann nur intuitive Beurteilungen liefern, aber solche Urteile wie „erhöht“ oder „reduziert“ sind vage, man braucht Vergleichsmaterial oder eine „Norm“, wobei es unterschiedliche Normen geben kann: die des Gedichts, die des Autors, die der Lyrik, die der Sprache usw. Die Resultate können sich unterscheiden je nachdem, mit welcher Norm man z.B. eine Strophe vergleicht.

Man kann sofort sehen, dass das Quadrat der Euklidischen Distanz dieses Vektors von dem Ursprung (d.h. von 0) die Wiederholungsrate darstellt, die damit eine spezielle Interpretation bekommt. Wir hätten (mit $p_i$ als Elementen dieses Vektors $V_T$)

$$R = \sum_{i \in V} (p_i - 0)^2 = \sum_{i \in V} p_i^2 = p_S^2 + p_V^2 + p_A^2 + ... + p_{Int}^2 = 0.1759$$

Betrachten wir den Vektor (4.22) separat für jede Strophe und untersuchen die Ausprägung der Wortarten. Einfachkeitshalber lassen wir *Part*, *Num* und *Int* aus, da diese überall die Häufigkeit 0 haben. Für die erste Strophe, die folgendermaßen aussieht

P V Adv Adv Pr S K S
P V Art S Pr P S
P V Art S Adv Pr Art S
P V P Adv P V P Adv

erhalten wir

$V_1$ = [6, 5, 0, 8, 5, 3, l, 3]/31

= [0.194, 0.161, 0.00, 0.258, 0.161, 0.097, 0.032, 0.097].

Auf diese Weise zählt man das ganze Gedicht durch und bekommt so schließlich die Zahlen in Tabelle 4.4, wo sie in Form von Spaltenvektoren aufgeführt werden. Die Norm ist der Vektor für das ganze Gedicht (s.oben).

Tabelle 4.4
Wortartenvektoren

| | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ | $V_6$ | $V_7$ | $V_8$ | Norm |
|---|---|---|---|---|---|---|---|---|---|
| S | .194 | .333 | .222 | .240 | .154 | .276 | .188 | .296 | .236 |
| V | .161 | .111 | .185 | .200 | .269 | .103 | .219 | .185 | .182 |
| A | .000 | .000 | .148 | .120 | .077 | .069 | .063 | .074 | .067 |
| P | .258 | .259 | .259 | .240 | .231 | .276 | .344 | .148 | .249 |
| Adv | .161 | .111 | .037 | .080 | .077 | .172 | .125 | .037 | .102 |
| Pr | .097 | .037 | .111 | .040 | .038 | .034 | .000 | .111 | .058 |
| K | .032 | .037 | .000 | .040 | .115 | .034 | .031 | .037 | .040 |
| Art | .097 | .111 | .037 | .040 | .038 | .034 | .031 | .111 | .067 |

In dieser Tabelle kann man nun verschiedene Eigenschaften untersuchen.

**(a) Profil der Wortart**

Eine Wortart weist in den einzelnen Strophen unterschiedliche Proportionen auf. Der Verlauf dieser Proportionen (s. Zeilen in Tab. 4.4) bildet das Profil der gegebenen Wortart, wie es in Abb. 4.3 für die Substantive dargestellt wird.

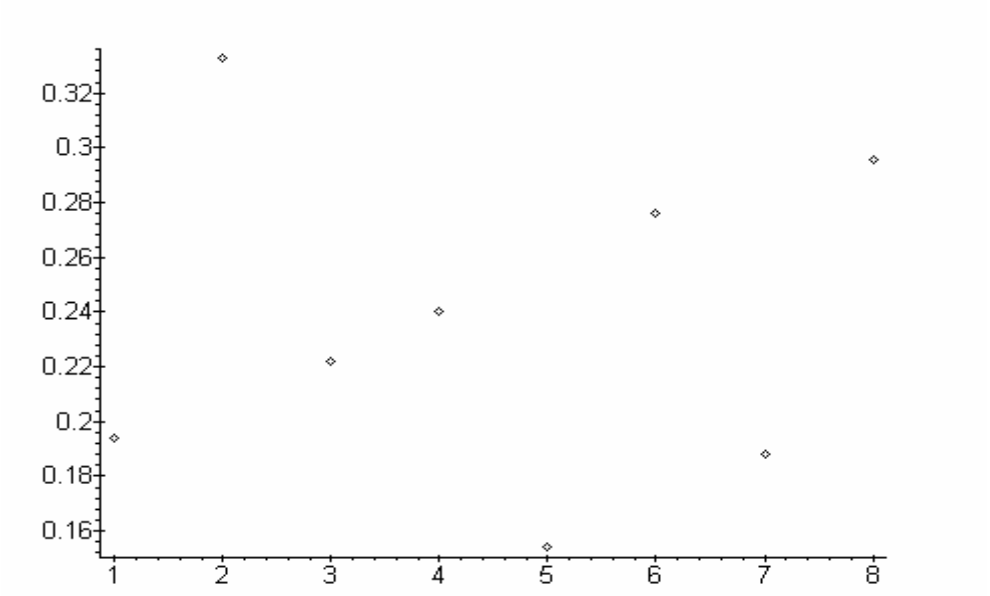


Abbildung 4.3. Profil der Substantive in den Strophen des „Erlkönigs“

In diesem scheinbaren Chaos erkennt man eine recht regelmäßige Wellenbewegung, die man eventuell mit Differenzengleichungen höherer Ordnung oder mit Fourieranalyse erfassen kann. Da wir aber erst nur einen Datensatz haben, und Erfahrungen aus anderen poetischen Texten nicht vorliegen, wäre die Anwendung dieser Methoden etwas verfrüht.

Die Profile einzelner Wortarten kann man miteinander vergleichen. Entweder benutzt man verschiedene Maße wie z.B. die Euklidische Distanz, den Korrelationskoeffizienten, verschiedenen Ähnlichkeitsmaße (vgl. Bock 1974), die man in jedem Lehrbuch der Klassifikation findet, oder man kann Konfidenzintervalle für Proportionen, Zusammenhänge zwischen den Wortarten und Tests zwischen Texten angeben, wozu man leider eine etwas kompliziertere Statistik braucht (cf. Wimmer, Altmann 2001a).

Eine andere Möglichkeit besteht darin, die Proportion einer Wortart von Strophe zu Strophe zu kumulieren. In unserem Fall würde das bedeuten, dass wir für eine Wortart eine Reihe aus der kumulativen Summe der entsprechenden Vektorelemente bilden, d.h.

$V_1$
$V_1 + V_2$
$V_1 + V_2 + V_3$

usw. bis zu $V_8$. So bekommen wir für die Substantive aus Tabelle 4.4 durch schrittweise durchgeführte Addition der Zahlen in der Zeile für *S* folgende Werte

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.194, | 0.527, | 0.749, | 0.989, | 1.143, | 1.419, | 1.607, | 1.903 |

Eingetragen in ein Koordinatensystem, ergeben diese Werte eine recht klare Gerade, wie man in Abb. 4.4 sehen kann. Die Gerade kann man als $S = 0.020286 + 0.232464P$ ($S$ = kumulativer Wert der Proportion der Substantive, $P$ = Position der Strophe) ausdrücken. Der Determinationskoeffizient ergibt D = 0.995.

Eine dritte Möglichkeit, die wir hier nur erwähnen werden, besteht darin, dass man jedem Wort seine Position im Vers zuschreibt, dann das Wort seiner Wortart zuordnet und schließlich den Zuwachs der Wortart per Position beobachtet. Dies ist nichts anderes als das übliche Type-Token-Verfahren, jedoch auf einzelne Wortarten beschränkt. Ziegler, Best und Altmann (2001a) haben eine Untersuchung dieser Art an einem deutschen Prosawerk durchgeführt und dabei festgestellt, dass sich alle Wortarten gemäß $y = ax^b$ entwickeln, manche sogar mit $b > 1$ (konvex). Die Kurven für einzelne Wortarten liefern dann das Wortartenspektrum des Textes in seiner Entfaltung.

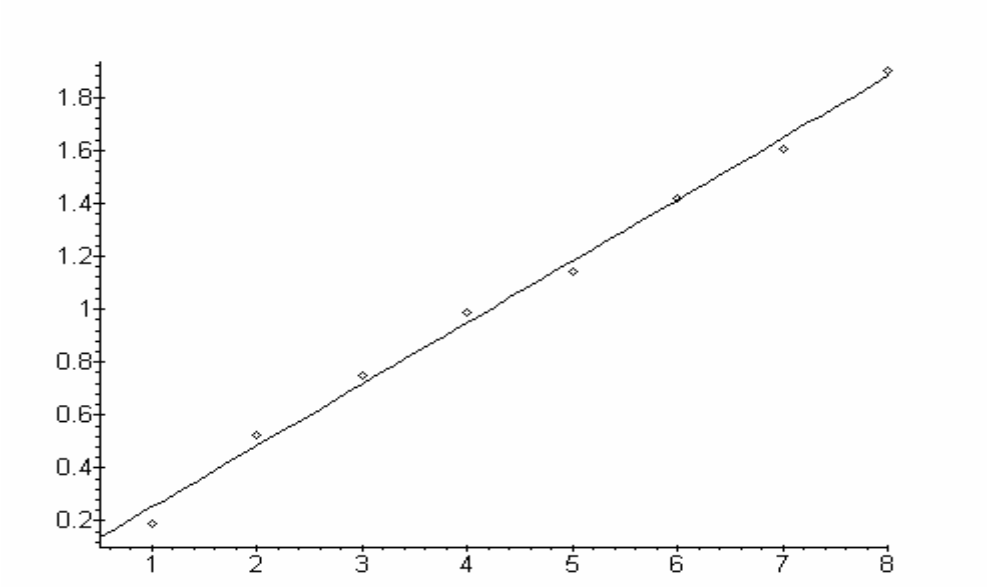


Abb. 4.4. Kumulative Proportionen der Substantive in einzelnen Strophen des „Erlkönigs“

**(b) Wortartstreuung**

Das Profil einer Wortart weist in den Strophen eine Streuung aus, die man numerisch ausdrücken kann. Man kann sie beispielsweise als die Summe der absoluten Abweichungen von der Norm, als Varianz oder als Standardabweichung u.ä. berechnen. Bezeichnen wir als $V_{ij}$ das *i*-te Element (*i*-te Wortart) des *j*-ten Vektors (der *j*-ten Strophe), die Norm (s. letzte Spalte von Tab. 4.4, die den Mittelwert darstellt) als $N_i$ und als *k* die Anzahl der Vektoren (Strophen) dann können wir ein Maß definieren als

$$(4.24)\quad W_i = \frac{100}{k}\sum_{j=1}^{k} |V_{ij} - N_i| .$$

So ergibt sich z.B. für die Substantive von Tabelle 4.4

$W_S$ = 100{0.194 – 0.236| + |0.333 – 0.236| + |0.222 – 0.236| + |0.240 – 0.236| +
+ |0.154– 0.236| + |0.276 – 0.236| + |0.198 – 0.236| + |0.296 – 0.236|}/8
= 4.71

Für die einzelnen Wortarten erhalten wir daher

$W_S$ = 4.71
$W_{Adv}$ = 4.23
$W_V$ = 3.98

$W_A$ = 3.64
$W_{Pr}$ = 3.58
$W_P$ = 3.46
$W_{Art}$ = 3.41
$W_K$ = 1.80

Wie man sieht, werden nur die Konjunktionen weniger strukturiert eingesetzt, alle anderen Wortarten weisen eine beträchtliche Variabilität in den Strophen auf.

Dieses Maß ist sozusagen lokal und nur eines von vielen möglichen. Vergleiche zwischen Texten müssen dann mit entsprechenden statistischen Methoden durchgeführt werden.

**(c) Strophenvektorprofil**

Wie wir bereits am Anfang von 4.2.1 gesehen haben, kann man für eine Strophe einen Wortartenvektor aufstellen (Spalten in Tab. 4.4). Da die Elemente des Vektors die Summe 1 ergeben, kann man für jeden Vektor dessen Entropie, Wiederholungsrate u.a. berechnen. Aus dem Verlauf einer derartigen Charakteristik kann man unter Umständen wiederum Schlüsse über inhaltliche Aspekte ziehen. Es geht nicht nur darum, irgendwelche Eigenschaften zu finden, die sich im Laufe des Gedichtes monoton entwickeln oder konstant bleiben, sondern um den Zusammenhang zwischen Inhalt und Form. Leider muss man sich bei dem inhaltlichen Aspekt auf die Urteile vieler Versuchspersonen verlassen, was unproportional viel mehr Arbeit bedeutet als die Auswertung von formalen Elementen. Die Aufstellung dieser möglichen Assoziationen bleibt dennoch ein wünschenswertes Forschungsziel. Hier berechnen wir nur die Wiederholungsrate für die einzelnen Strophenvektoren $V_i$:

$R_1$ = 0.1758
$R_2$ = 0.2181
$R_3$ = 0.1879
$R_4$ = 0.1808
$R_5$ = 0.1778
$R_6$ = 0.2002
$R_7$ = 0.2223
$R_8$ = 0.1770

Trägt man diese Werte in ein Koordinatensystem ein, so erhält man Abbildung 4.5. Eine mögliche Assoziation mit dem Inhalt wird hier nicht erörtert, dies gehört in die Kompetenz der Literaturwissenschaft. Die Entropie verhält sich fast spiegelbildlich: dort, wo die Wiederholungsrate ansteigt, sinkt die Entropie, und umgekehrt. Die Berechnung wird dem Leser überlassen. Jedoch haben auch solche recht „chaotischen“ Verläufe ihre Eigenschaften, die untersuchungswert sind.

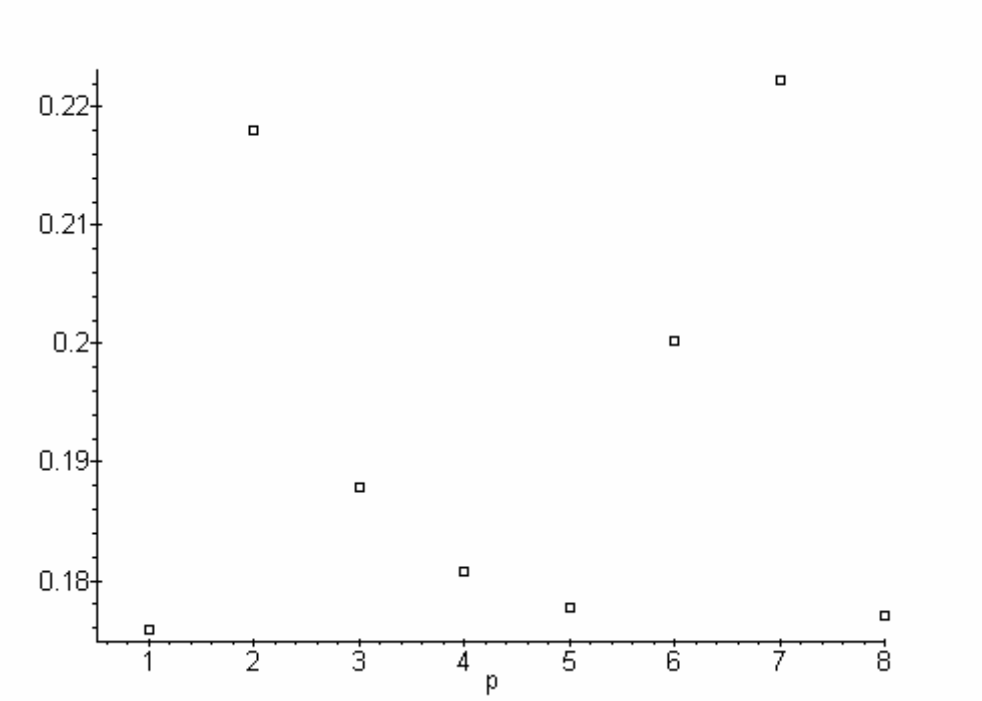


Abbildung 4.5. Verlauf der Wiederholungsrate der Wortartenvektoren im „Erlkönig“ (p = Position)

### 4.2.2. Der Aktionsquotient

Hat man eine Vorstellung davon, was eine Wortart ausdrückt, dann kann man durch deren Kombinationen zahlreiche Textattribute definieren. Dabei lässt sich eine Wortart weiter unterteilen und mit Unterklassen anderer Wortarten kombinieren. Beispielsweise kann man die Wörter „Überlegung“, „überlegen“ und „überlegt“ grammatisch drei Wortartenklassen zuordnen, aber unter einem anderen Gesichtspunkt kann man sie zu einer einzigen Klasse zusammenfassen. Von einer ähnlichen Idee ging Busemann (1925) aus, der den sogenannten Aktionsquotienten definiert hat, indem er die Zahl der Verben in einem Text zur Zahl der Adjektive ins Verhältnis setzte und den Wert des Koeffizienten als Ausprägung einer Eigenschaft auf der Achse „Deskriptivität-Aktivität“ betrachtete. Mit wachsendem Alter eines eines Autors werden dessen Texte angeblich immer „deskriptiver“. Untersuchungen zum Aktionsquotienten wurden von Boder (1940), Schlissmann (1948), Antosch (1953), Fischer (1969) und Altmann (1988) durchgeführt. Die formalen und die inhaltlichen Schwächen des Indexes wurden von Altmann (1978) analysiert, der die folgende Version vorgeschlagen hat.

$$Q = \frac{V}{A+V}. \tag{4.25}$$

Man kann natürlich zu den „Deskriptionsmitteln“ auch die Adverbien zählen, weil sie sowohl Verben als auch Adjektive spezifizieren, eventuell auch Zahlwörter oder andere Wortarten, je nach dem, wie man Deskriptivität-Aktivität auffassen möchte. In einem solchen Fall kommen alle betreffenden Wortarten in den Nenner.

In der Form (4.25) lässt sich der Index gut interpretieren. Er verläuft im Intervall <0, 1> und stellt einfach eine Proportion dar; bei Q = 0 haben wir minimale Aktivität, bei Q = 0.5 haben wir ein Deskriptivitäts-Aktivitäts Gleichgewicht, bei Q = 1 ist der Text maximal „aktiv“. Q = 0 ist im Grunde nur dann möglich, wenn man nicht alle Verben berücksichtigt, sondern nur diejenigen, die echte Aktivität ausdrücken. Verben wie „sein“, „haben“, „ruhen“, „glauben“, „bleiben“ usw. sind nicht aktiv, „reiten“, „tun“ sind aktiv, andere wieder haben einen Zwischenstatus wie „meinen“. Daher ist die Berechnung des Aktionsquotienten nicht nur eine formale Angelegenheit, man sollte bei Verben eventuell den Aktivitätsgrad messen, z.B. durch Urteile von Versuchspersonen. Hat man aber einmal beschlossen, welche Verben und Adjektive man in die Rechnung einbeziehen will, so kann man die *Hypothese der erhöhten Aktivität* mit Hilfe der Binomialverteilung testen, indem man

$$(4.26)\quad P(X \geq V) = \sum_{x=V}^{n} \binom{n}{x} 0.5^n$$

berechnet, bzw. die *Hypothese der erhöhten Deskriptivität* mit

$$(4.27)\quad P(X \leq V) = \sum_{x=0}^{V} \binom{n}{x} 0.5^n$$

testen, wobei $n = A + V$. Ist (4.26) oder (4.27) kleiner als ein vorgegebenes $\alpha$, z.B. $\alpha = 0.05$, dann wird die Hypothese angenommen. Die Binomialverteilung kann benutzt werden, weil ein in Betracht gezogenes Wort entweder deskriptiv (A) oder aktiv (V) ist, wie man sie auch immer bestimmen mag. Daher hat jeder Ausgang nur zwei Möglichkeiten, und die Wahrscheinlichkeit von genau $x$ gegebenen Ausgängen kann mit Hilfe der Binomialverteilung berechnet werden. Es reicht, wenn man die Zahl der Verben (= aktive Wörter) berücksichtigt; ist ihre Zahl größer als die der Adjektive, d.h. V > A, dann berechnet man (4.26); ist ihre Zahl kleiner als die der Adjektive, d.h. V < A, dann berechnet man (4.27).

Nehmen wir einfachkeitshalber an, dass alle Verben die Voraussetzung der Aktivität erfüllen, so haben wir im „Erlkönig“ nach Tabelle 4.3

V = 41, A = 15, A + V = n = 56.

Da es hier mehr Verben als Adjektive gibt, testen wir die Hypothese der erhöhten Aktivität (de facto ist dies die Alternativhypothese zu der Nullhypothese, dass keine erhöhte Aktivität existiert) und berechnen (4.26) als

$$P(X \geq 41) = \sum_{x=41}^{56} \binom{56}{x} 0.5^{56} = 0.0003,$$

d.h., der Text weist eine signifikante Aktivität auf, weil $P < 0.05$. Das Rechnen mit dieser Formel, die exakte Resultate liefert, ist nur mit einem Programm möglich. Bei großen Stichproben kann man sich die Arbeit beträchtlich erleichtern, indem man einfach einen Chiquadrat-Test durchführt, nämlich

$$(4.28)\quad X_1^2 = \frac{(V - A)^2}{V + A}.$$

In unserem Fall bekommen wir

$$X_1^2 = \frac{(41-15)^2}{41+15} = 12.07.$$

Ein Chiquadrat dieser Größe mit einem Freiheitsgrad ergibt die Überschreitungswahrscheinlichkeit von $P = 0.0005$ (zweiseitig).

Hat man für Charakterisierungszwecke den Index Q in (4.25) berechnet, so kann man bei großen Stichproben direkt das Kriterium

$$(4.29)\quad u = (2Q-1)\sqrt{n}$$

berechnen, wobei man $u^2 = X_1^2$ erhält. In unserem Fall ist $Q = 41/56 = 0.7321$, d.h.

$$u = [2(0.7321) - 1]\sqrt{56} = 3.4738$$

und $3.4738^2 = 12.07$, was mit der obigen Rechnung übereinstimmt. Die Wahrscheinlichkeit für $u$ findet man in den entsprechenden Tabellen der Normalverteilung (P = 0.0005 zweiseitig). Der Binomialtest ist einseitig und etwas exakter als die beiden anderen. Die Zusammenhänge findet man in Altmann (1978), Wimmer et al. (2003).

Vergleicht man zwei Texte auf ihre „Aktivität" hin, so kann man auch die Normalverteilung verwenden. Bezeichnen wir die Daten eines Textes mit dem Index 1, die des zweiten mit dem Index 2, so stellen wir das Testkriterium in der Form

$$(4.30)\quad u = \frac{(V_1 n_2 - V_2 n_1)\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2 (V_1 + V_2)(A_1 + A_2)}}$$

auf. Die Größe *u* ist normalverteilt *N*(0,1), wie oben bedeutet $n = A + V$ für die einzelnen Texte.

## 4.3. Das type-token Verhältnis (TTR)

Diese in der Überschrift benannte Problematik ergab sich ursprünglich aus dem Wunsch, den Vokabularreichtum eines Textes zu charakterisieren. Vokabularreichtum bedeutet eigentlich die Zahl unterschiedlicher Wörter (types) im Text. Man weiß, dass mit wachsender Textlänge auch die Zahl der types anwächst, jedoch ist schon sehr bald festgestellt worden, dass sich dieses Wachstum verlangsamt. Ist ein Text sehr lang, dann kommen neue Wörter immer seltener hinzu. Man hat versucht diesen Umstand immer dadurch zu berücksichtigen, dass die Textlänge (Zahl der tokens) und die Zahl unterschiedlicher Wörter (Zahl der Types) zueinander in Beziehung gesetzt wurden, man stellte Indizes auf und bemühte sich, diese stabil zu gestalten, oder man drückte den types-Zuwachs mit einer Kurve aus. Dabei wurde immer wieder festgestellt, dass ein Index oder eine Kurve, die für bestimmte Daten valide sind, für andere Daten verzerrte Resultate, z.B. unrichtige Voraussagen, lieferte. Es stellte sich weiter heraus, dass die Indizes eine stark anwachsende Streuung aufweisen und daher zur Unterscheidung von Texten ungeeignet sind. Manche Modelle gehen von linguistisch unhaltbaren Voraussetzung aus, tragen aber dennoch zur Klärung der Probleme bei. Man kann ruhig sagen, dass es genauso viele ungelöste Probleme wie Lösungen gibt. Schon deren Aufzählung würde ein ganzes Buch füllen, daher werden wir uns einschränken müssen.

Im allgemeinen betrachtet man die unterschiedlichen Indizes als Maße des Vokabularreichtums, die type-token-Kurven hingegen als Messungen einer Art des Informationsflusses im Text. Der Informationsfluss ist um so langsamer, je öfter sich bereits benutzte Wörter wiederholen, und um so schneller, je mehr neue Wörter hinzukommen. In didaktischen Texten müsste daher eine Kurve flacher verlaufen, in poetischen – zumindest in den meisten – und in journalistischen Texten steiler, weil die letzteren üblicherweise kurz sind und in "kurzer Zeit" viel mitteilen müssen, während man in poetischen Texten Wiederholungen eher aus stilistischen Gründen meidet. Dies muss aber nicht unbedingt immer so sein.

Ein rein philologisches Problem ist wieder die Frage, ob man Lemmata oder Wortformen zählen soll. Praktischer – auch bei der EDV – ist natürlich die Zählung der Wortformen, da die Lemmatisierung viele Probleme aufwirft, die in unterschiedlichen Sprachen jeweils anders zu lösen sind und vielfältige Schwierigkeiten bereiten. Bei der Wortformenzählung wird man etwa feststellen, dass

die Übersetzung eines lateinischen Textes ins Englische einen viel geringeren Vokabularreichtum aufweist als der Originaltext, dass also die Kurve des Informationsflusses viel flacher ist, obwohl es sich in gewissem Sinne um identische Texte handelt. Daher kann man bei der Zählung von Wortformen keine zwischensprachlichen Vergleiche durchführen, denn sonst stellt man fest, dass sich hochsynthetische Sprachen durch vokabularreiche Texte auszeichnen, was völlig unsinnig ist. Aber auch bei der Lemmatisierung gestaltet sich die Lage nicht viel einfacher, denn wie soll man z.B. Suppletivismus, abtrennbare Präfixe, komplexe Verbformen, Kontraktionen („am“, „grausets“), genusunterscheidende Affixe, Bindestrichkomposita usw. behandeln? Was ist ein Lemma in einigen indianischen Sprachen? Wie kann man bei TTR-Problemen die Grammatik ausschalten? Da es uns hier lediglich um ein Gedicht geht, werden wir bei den Wortformen bleiben, so dass man die im folgenden angegebenen Verfahren zumindest für die Untersuchung deutscher Texte verwenden kann.

Um sich den TTR-Verlauf vorzustellen, betrachten wir die erste Strophe des „Erlkönig“. Jede Wortform bekommt hier eine Zahl zugeordnet, die ihrer Stelle im Text entspricht:

1 2 3 4 5 6 7 8
Wer reitet so spät durch Nacht und Wind?
9 10 11 12 13 14 15
Es ist der Vater mit seinem Kind;
16 17 18 19 20 21 22 23
Er hat den Knaben wohl in dem Arm,
24 25 26 27 28 29 30 31
Er faßt ihn sicher, er hält ihn warm.

Die Zahl der types ($V_i$) (= unterschiedliche Wortformen) wächst gleichmäßig mit der Zahl der tokens ($L_i$ = Textlänge bis zum *i*-ten Wort), die erste Wiederholung erfolgt in $L_{24}$, wo sich „er“, das bereits in Position $L_{16}$ vorkommt, wiederholt. Nochmals begegnet man „er“ in $L_{26}$. So erhält man bis 23 immer $L_i = V_i$ , ab 23 gestaltet sich die Auszählung folgendermaßen:

| $L_i$ | $V_i$ | *Wortform* |
|---|---|---|
| 23 | 23 | Arm |
| 24 | 23 | **er** |
| 25 | 24 | faßt |
| 26 | 25 | ihn |
| 27 | 26 | sicher |
| 28 | 26 | **er** |
| 29 | 27 | hält |
| 30 | 27 | **ihn** |
| 31 | 28 | warm |

…….

das heißt, dort, wo eine bereits benutzte Wortform auftritt, wird $V_i$ nicht erhöht. Auf diese Weise erhält man eine monoton wachsende Treppenkurve, die sich auf verschiedene Weisen modellieren lässt. Mathematiker arbeiten lieber mit stochastischen Prozessen, während stärker linguistisch orientierte Forscher eher Kurven benutzen, oftmals auch nur aus Kurven abgeleitete Indizes. Die Literatur zu diesem Problem ist umfangreich (vgl. z.B. Brainerd 1972, 1982; Gani 1975; Simon 1955; Haight, Jones 1974; Lánský, Radil-Weiss 1980; Herdan 1964, 1966, Müller 1971; Maas 1972; Nešitoj 1975; Ratkowsky, Halstead, Hantrais 1980; Tuldava 1995, 1998; Orlov, Boroda, Nadarejšvili 1982; Ejiri, Smith 1993; Baayen 2001 und die gesamte französische Schule; vgl. die Bibliographie von Köhler 1995).

Das Resultat der Auszählung des „Erlkönig" findet sich in Tabelle 4.5. Hier werden wir nur einige Kurven anpassen und deren Aussagen vergleichen. Eine bessere empirische Aussage für den „Erlkönig" bedeutet nicht, dass diese auch für andere Daten die „beste" sein müsse. Wir sind zwar fest davon überzeugt, dass der Informationsfluss in Texten gesetzesartig verläuft, aber Gesetze wirken nur, wenn bestimmte Anfang- und Randbedingungen gegeben und erfüllt sind. Diese aber können in Texten sehr unterschiedlich ausfallen, sogar in Texten desselben Autors. Wir können davon ausgehen, dass es für Texte unterschiedliche Gleichgewichtszustände geben kann – d.h. solche Zustände, bei denen sie verstanden werden können – und solche Zustände kann man als Attraktoren bezeichnen. Die Attraktoren können aber so weit voneinander entfernt liegen, dass man sie mit einer bloßen Variation der Parameter einer Kurve nicht erfassen kann, sondern durch Änderung sämtlicher Voraussetzungen zu völlig anderen Kurven kommen muss.

Ein anderer Umstand, der zu der gleichen Überlegung führt, ist die Tatsache, dass nur kurze Texte „in einem Atemzug" entstehen können. Bei der Erzeugung längerer Texte macht der Autor notgedrungen Pausen. Während einer Pause kann das bereits benutzte Vokabular mental deaktiviert werden, und beim Wiedereinsetzen des Schreibvorgangs kann dies die bis dahin glatte empirische Kurve mit einem „Sprung" verzerren. Solche „Sprünge" sind nicht nur Signale für neue Abschnitte, etwa für neue Kapitel, sondern kommen auch innerhalb solcher Abschnitte vor, wenn diese nur lang genug sind. Ein derartiger Neuanfang nach einem Sprung kann wiederum einen ganz anderen Attraktor signalisieren. Textanalytiker, die den gesamten Verlauf von $V_i = f(L_i)$ mit einer einzigen Kurve modellieren wollen, geraten unvermeidlich in Schwierigkeiten, greifen zu Polynomen, Überlagerungen, Mischungen usw. Doch auch, wenn sie schließlich gute Resultate erreichen, kann man diese weder interpretieren, noch in eine Theorie einbetten. Bei langen Texten ist es sogar empfehlenswert, sie in kohärente Teile aufzuspalten und alle Teile einzeln durchzurechnen.

Tabelle 4.5
Der type-token-Verlauf im „Erlkönig“ (Wortformen)

| $L_i$ | $V_i$ | $L_i$ | $V_i$ | $L_i$ | $V_i$ | $L_i$ | $V_i$ | $L_i$ | $V_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 46 | 39 | 91 | 65 | 136 | 88 | 181 | 108 |
| 2 | 2 | 47 | 39 | 92 | 65 | 137 | 88 | 182 | 108 |
| 3 | 3 | 48 | 40 | 93 | 65 | 138 | 88 | 183 | 109 |
| 4 | 4 | 49 | 40 | 94 | 65 | 139 | 88 | 184 | 109 |
| 5 | 5 | 50 | 41 | 95 | 65 | 140 | 88 | 185 | 109 |
| 6 | 6 | 51 | 41 | 96 | 65 | 141 | 88 | 186 | 109 |
| 7 | 7 | 52 | 42 | 97 | 66 | 142 | 88 | 187 | 109 |
| 8 | 8 | 53 | 42 | 98 | 67 | 143 | 88 | 188 | 110 |
| 9 | 9 | 54 | 42 | 99 | 68 | 144 | 88 | 189 | 110 |
| 10 | 10 | 55 | 42 | 100 | 69 | 145 | 88 | 190 | 110 |
| 11 | 11 | 56 | 42 | 101 | 70 | 146 | 89 | 191 | 110 |
| 12 | 12 | 57 | 43 | 102 | 70 | 147 | 90 | 192 | 110 |
| 13 | 13 | 58 | 44 | 103 | 70 | 148 | 90 | 193 | 110 |
| 14 | 14 | 59 | 44 | 104 | 70 | 149 | 91 | 194 | 110 |
| 15 | 15 | 60 | 45 | 105 | 70 | 150 | 92 | 195 | 110 |
| 16 | 16 | 61 | 45 | 106 | 71 | 151 | 93 | 196 | 110 |
| 17 | 17 | 62 | 46 | 107 | 72 | 152 | 93 | 197 | 111 |
| 18 | 18 | 63 | 47 | 108 | 73 | 153 | 93 | 198 | 112 |
| 19 | 19 | 64 | 47 | 109 | 73 | 154 | 93 | 199 | 112 |
| 20 | 20 | 65 | 48 | 110 | 73 | 155 | 93 | 200 | 112 |
| 21 | 21 | 66 | 49 | 111 | 74 | 156 | 93 | 201 | 113 |
| 22 | 22 | 67 | 50 | 112 | 75 | 157 | 94 | 202 | 113 |
| 23 | 23 | 68 | 51 | 113 | 76 | 158 | 94 | 203 | 113 |
| 24 | 23 | 69 | 52 | 114 | 76 | 159 | 95 | 204 | 114 |
| 25 | 24 | 70 | 53 | 115 | 76 | 160 | 95 | 205 | 114 |
| 26 | 25 | 71 | 53 | 116 | 76 | 161 | 96 | 206 | 114 |
| 27 | 26 | 72 | 54 | 117 | 77 | 162 | 97 | 207 | 114 |
| 28 | 26 | 73 | 55 | 118 | 77 | 163 | 98 | 208 | 115 |
| 29 | 27 | 74 | 56 | 119 | 78 | 164 | 99 | 209 | 116 |
| 30 | 27 | 75 | 57 | 120 | 79 | 165 | 99 | 210 | 117 |
| 31 | 28 | 76 | 58 | 121 | 80 | 166 | 100 | 211 | 117 |
| 32 | 29 | 77 | 59 | 122 | 81 | 167 | 100 | 212 | 118 |
| 33 | 30 | 78 | 59 | 123 | 82 | 168 | 101 | 213 | 118 |
| 34 | 31 | 79 | 60 | 124 | 82 | 169 | 101 | 214 | 119 |
| 35 | 32 | 80 | 61 | 125 | 82 | 170 | 102 | 215 | 119 |
| 36 | 33 | 81 | 62 | 126 | 83 | 171 | 103 | 216 | 120 |
| 37 | 33 | 82 | 62 | 127 | 83 | 172 | 104 | 217 | 120 |
| 38 | 34 | 83 | 62 | 128 | 84 | 173 | 104 | 218 | 121 |
| 39 | 35 | 84 | 63 | 129 | 85 | 174 | 105 | 219 | 121 |
| 40 | 36 | 85 | 64 | 130 | 85 | 175 | 105 | 220 | 122 |
| 41 | 37 | 86 | 64 | 131 | 86 | 176 | 106 | 221 | 122 |
| 42 | 37 | 87 | 64 | 132 | 86 | 177 | 106 | 222 | 122 |
| 43 | 37 | 88 | 64 | 133 | 87 | 178 | 106 | 223 | 122 |
| 44 | 37 | 89 | 64 | 134 | 87 | 179 | 107 | 224 | 123 |
| 45 | 38 | 90 | 64 | 135 | 88 | 180 | 107 | 225 | 124 |

Da wir es nur mit einem kurzen Text zu tun haben, werden wir im folgenden neun bekannte Kurven und den Umgang mit ihnen darstellen. Dabei bedeuten $V_i$ immer das Vokabular (Zahl der types bis zur Stelle *i*) und $L_i$ die Länge des Textes (Zahl der tokens bis zur Stelle *i*).

**(1) Herdan**

Vielleicht die einfachste Kurve ist diejenige, die G. Herdan (1964) aufgrund einfacher Proportionalität abgeleitet hat:

(4.31) $V_i = L_i^b$.

Hat man keine Software zur Verfügung, die iterativ die Anpassungverbessert, z.B. NLREG von P.H. Sherrod, SPSS oder Matlab, dann muss man den Parameter *b* schätzen. In diesem Fall kann man (4.31) logarithmieren, um

(4.32) $\log V_i = b \log L_i$

zu erhalten, woraus man leicht den Schätzer

(4.33) $$\hat{b} = \frac{1}{n-1}\sum_{i=2}^{n}\frac{\log V_i}{\log L_i}$$

aufstellen kann. Hier ist *n* die Textlänge, d.h. 225, und die Summierung läuft von $i = 2$, weil $\log 1 = 0$. Dieser Schätzer ergibt sich als $b = 0.92583$, und der Determinationskoeffizient beträgt $D = 0.87$.

Mit Hilfe der Methode der kleinsten Quadrate bekommt man aufgrund von (4.32) den Schätzer

(4.34) $$\hat{b} = \frac{\sum_{i=1}^{n}\log L_i(\log V_i)}{\sum_{i=1}^{n}(\log L_i)^2},$$

und durch einfache Rechnung erhält man den Zähler als

$$\log 1(\log 1) + \log 2(\log 2) + \ldots + \log 225(\log 124) = 4230.4065.$$

Der Nenner ist einfach die Summe der Quadrate von Logarithmen natürlicher Zahlen von 1 bis 225, d.h. 4625.6019. Daher ergibt sich dieser Schätzer als

$$\hat{b} = \frac{4230.4065}{4625.6019} = 0.9146\,.$$

Setzt man diesen Wert in Formel (4.31) ein, so erhält man den Determinationskoeffizienten $D = 0.95$, der sicherlich besser ist als der erste Schätzer. Die empirischen type-Werte sind in Abbildung 4.6 als Punkte, die Kurve mit dem zweiten Schätzer als glatte Linie eingezeichnet.

Wie man sieht, erfasst diese Kurve den Verlauf der types etwa bis zur Hälfte ganz gut, dann wird der Abstand immer größer. Auch durch Optimierung erreicht man eine nur unwesentliche Verbesserung.

Praktisch identisch mit Herdans Kurve ist diejenige, die man aus Ejiris und Smiths (1993) Index erstellen kann. Sie hat eine multiplikative Konstante, die aber gleich 1 sein muss, weil bei $L_1 = 1$ sich $V_1 = 1$ ergeben muss.

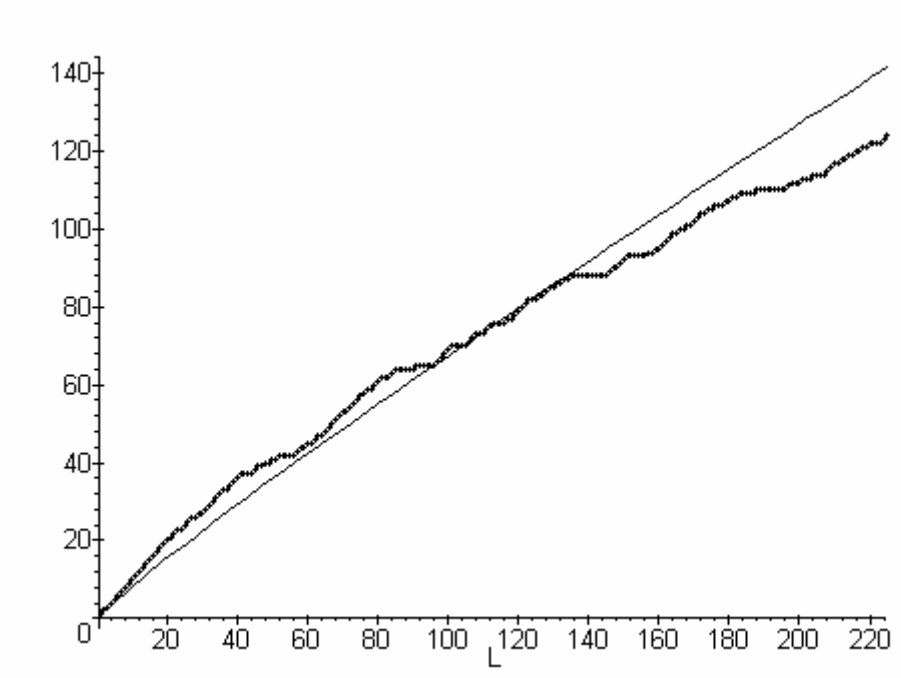


Abbildung 4.6. Die types im „Erlkönig“ und Herdans Kurve

Verläufe dieser Art, wo z.B. die erste Hälfte der Daten über der Kurve, die zweite unter der Kurve liegen, zeigen, dass das Vokabular zwar schnell anwächst, dann aber, wegen der Konzentration auf das engere Thema, nur wenig Neues hinzukommt. Wäre das Gedicht länger, so käme es eventuell zu einer noch größeren Diskrepanz. Gerade solche Überlegungen zwingen uns nach anderen Kurven zu suchen und eventuell die Interpretation des Informationsflusses zu modifizieren. Jedoch nehmen wir an, dass ein Parameter mit der thematischen Konzentration zusammenhängen kann. Bei dieser Kurve drückt *b* nur das Gesetz im allgemeinen aus, nimmt aber keine Rücksicht auf die Randbedingungen.

Aus dem Verlauf der empirischen Werte ($V_i$) kann man im Text Brüche an den Stellen vermuten, wo sich nach einer längeren fast oder ganz horizontalen Bewegung der Punkte plötzlich ein steiler Anstieg bemerkbar macht. Jedoch müssen solche Stellen strikt getestet werden, intuitive Aussagen sind wenig wert.

**(2) Tuldava**

Die graduelle Abweichung des Vokabulars von der gerade besprochenen theoretischen Kurve, die in anderen Wissenschaften als Potenzgesetz oder auch als allometrisches Gesetz bekannt ist, veranlasste Tuldava (1974, 1998: 86), nach anderen Möglichkeiten zu suchen. Als erste benutzte er die Tornquist-Kurve

$$(4.35)\quad V_i = \frac{aL_i}{L_i + b},$$

die für kleinere Stichproben geeignet sein sollte. Die Parameter *a* und *b* kann man nach der Linearisierung[2] in der Form

$$(4.36)\quad V_iL_i + bV_i = aL_i$$

mit der Methode der kleinsten Quadrate aus

$$(4.37)\quad \sum_i (V_iL_i + bV_i - aL_i)^2 = \min!$$

erhalten. Man bekommt die Schätzer

$$(4.38)\quad \hat{a} = \frac{\sum_i V_iL_i^2 \sum_i V_i^2 - \sum_i V_iL_i \sum_i V_i^2L_i}{\sum_i V_i^2 \sum_i L_i^2 - \left(\sum_i V_iL_i\right)^2}$$

$$(4.39)\quad \hat{b} = \frac{\sum_i V_iL_i^2 \sum_i V_i\ L_i - \sum_i L_i^2 \sum_i V_i^2L_i}{\sum_i V_i^2 \sum_i L_i^2 - \left(\sum_i V_iL_i\right)^2}.$$

[2] Unter Linearisierung versteht man die Überführung einer Formel, die Potenzen von x oder x im Exponenten usw. enthält, in eine Form, in der es nur lineare Funktionen von x gibt, z.B. ergibt $y = 3x^2$ durch Logarithmieren ln y = ln 3 + 2 ln x. Man schreibt ln y = Y, ln x = X und bekommt Y = ln 3 + 2X, allgemein Y = A + BX, eine Funktion, die eine Gerade darstellt.

Berechnet man diese Werte aus den Daten, so erhält man

$$\hat{a} = 287.3216 \qquad \hat{b} = 309.9098\,.$$

Setzt man diese Werte in (4.35) ein, so erhält man die Kurve, die in Abbildung 4.7 zu sehen ist. Optisch scheint sich diese Kurve sehr gut den Daten anzupassen. Der Determinationskoeffizient beträgt jetzt D = 0.997. Für unsere Daten ist daher Tuldavas Kurve ausgezeichnet geeignet, es gibt aber keine Garantie, dass dies auch bei anderen Daten so sein wird.

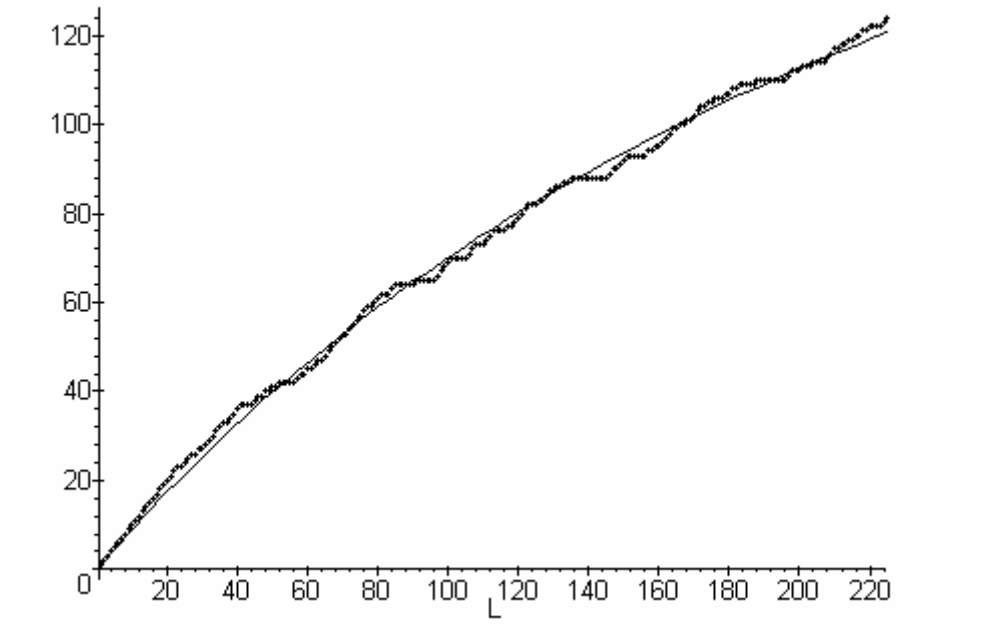


Abbildung 4.7. Anpassung der Kurve von Tuldava an Types in Erlkönig

Wie man leicht sieht, wenn man in (4.35) $L_1 = 1$ einsetzt, erhält man $V_1 = 0.92$, einen Zustand also, wie er in der Empirie nicht möglich ist. Besteht man darauf, dass bei $L_1 = 1$ auch $V_1 = 1$ sein muss, dann wird die Formel etwas einfacher, denn diese Forderung ist nur dann erfüllt, wenn $b = a$-1. Aus (4.35) wird dann

$$(4.40)\quad V_i = \frac{aL_i}{a-1+L_i},$$

wobei man nur einen Parameter schätzen muss. Löst man (4.40) nach $a$ auf, dann erhält man

$$a = \frac{V_i(L_i-1)}{L_i - V_i}.$$

Den Schätzer von *a* bekommt man dann aus dem Durchschnitt dieser Zahlen, d.h. als

$$(4.41)\quad \hat{a} = \frac{1}{n-k}\sum_{i=k}^{n}\frac{V_i(L_i-1)}{L_i-V_i},$$

wobei *k* die erste Position ist, bei der $L > V$, sonst bekämen wir eine 0 im Nenner. In Tabelle 4.5 sieht man, dass $k = 24$, weil hier $L = 24$ und $V = 23$. Konkret haben wir beim „Erlkönig“

$$\hat{a} = \frac{1}{225-24}\sum_{i=24}^{225}\frac{V_i(L_i-1)}{L_i-V_i} = 251.5539.$$

Setzt man diese Zahl in (4.40) und passt die Kurve an, so erhält man einen Determinationskoeffizienten $D = 0.996$, d.h., diese Kurve ist keineswegs schlechter als die Kurve mit zwei Parametern.

Bei der Methode der kleinsten Quadrate linearisiert man (4.40) zuerst als

$$(4.42)\quad (L_i + a - 1)V_i = aL_i$$

und berechnet auf dem üblichen Wege

$$(4.43)\quad \hat{a} = \frac{\sum_{i=1}^{n}V_i(L_i-V_i)(L_i-1)}{\sum_{i=1}^{n}(L_i-V_i)^2}.$$

In unserem Falle ergibt sich (4.43) als $\hat{a} = 252.5603$ und liefert einen Determinationskoeffizienten $D = 0.996$, d.h. eine gleich gute Anpassung

**(3) Köhler-Martináková**

R. Köhler und Z. Martináková benutzen für TTR in der Musik (1998) eine etwas verallgemeinerte Version der Kurve von Tuldava, nämlich

$$(4.44)\quad V_i = \frac{aL_i}{1-b+bL_i}.$$

Nach der Linearisierung in der Form

(4.45) $(1-b)V_i + bV_iL_i = aL_i$

erhält man die Kleinstquadratschätzungen der Parameter *a* und *b* als

(4.46) $$\hat{a} = \frac{\sum V_i L_i^2 (\sum V_i^2 - \sum V_i^2 L_i) + \sum V_i L_i (\sum V_i^2 L_i^2 - \sum V_i^2 L_i)}{D}$$

(4.47) $$\hat{b} = \frac{\sum L_i^2 (\sum V_i^2 - \sum V_i^2 L_i) - \sum V_i L_i (\sum V_i L_i - \sum V_i L_i^2)}{D}$$

wobei

$$D = \sum L^2 (\sum V^2 - 2\sum V^2 L + \sum V^2 L^2) - (\sum V L - \sum V L^2)^2 .$$

Die etwas langwierige Berechnung ergibt

$$\hat{a} = 0.8994453, \qquad \hat{b} = 0.0029675$$

und liefert einen Determinationskoeffizienten $D$ = 0.997, d.h., genauso gut wie die Kurve von Tuldava.

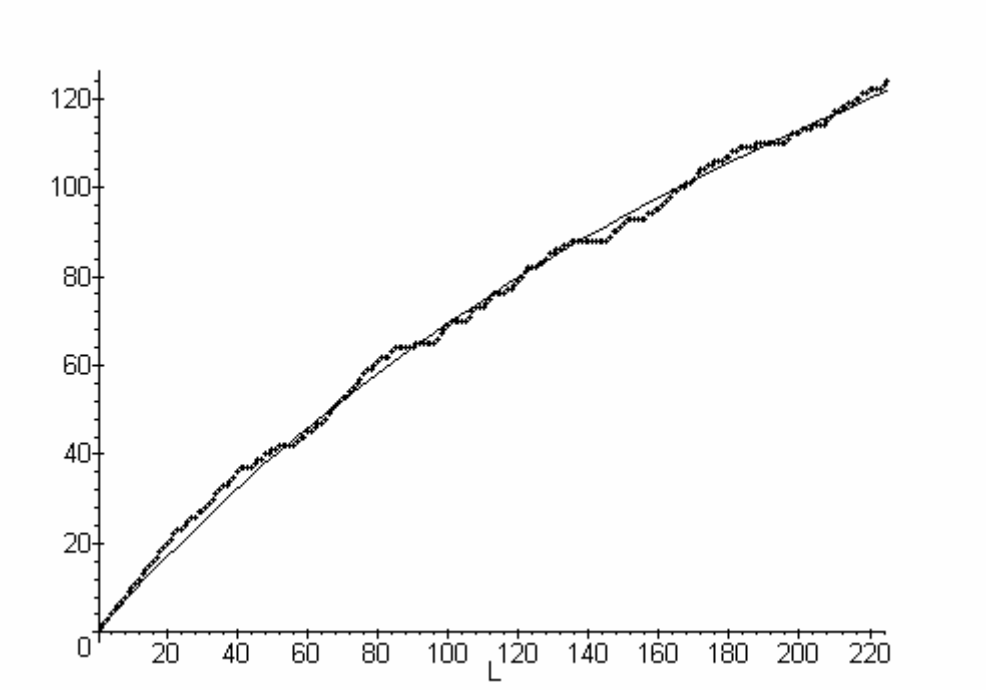


Abbildung 4.8. Anpassung der Kurve von Köhler-Martináková an die types im „Erlkönig“

**(4) Brunet**

Aus Brunets Überlegungen (1978) lässt sich die Kurve

(4.48) $V_i = b(\ln\ L_i)^c$

ableiten, die gleichfalls zwei Parameter hat. Die linearisierte Form ergibt

(4.49) $\ln V_i = \ln b + c\ln(\ln L_i)$,

wobei wir $\ln b = A$ setzen, d.h. $b = e^A$. Die übliche Kleinstquadratmethode ergibt dann die Schätzer

$$(4.50)\quad \hat{c} = \frac{n\sum_i \ln V_i \ln(\ln L_i) - \sum_i \ln V_i \sum_i \ln(\ln L_i)}{n\sum_i [\ln(\ln L_i)]^2 [\sum_i \ln(\ln L_i)]^2},$$

$$(4.51)\quad A = \frac{\sum_i \ln\ V_i - \hat{c}\sum_i \ln(\ln L_i)}{n}.$$

Bei der Formel (4.48) sehen wir aber, dass bei $L_1$ sich $V_1$ als 0 ergibt, was unmöglich ist. Ebenso ergibt sich in (4.49) bei $L_1$ eine Singularität. Dies zwingt uns zuerst, Formel (4.48) zu

(4.52) $V_i = 1 + b(\ln\ L_i)^c$

zu modifizieren und die Summierungen in (4.50) und (4.51) erst von $i$ = 2 ab durchzuführen. Diese Formeln bleiben aber gültig, wenn wir statt $n$ an allen drei Stellen $n$-1 setzen und statt ln $V_i$ in beiden Formeln ln($V_i$ – 1) schreiben.

Auf diese Art ergeben sich die Schätzer für (4.52) als

$\hat{c}$ = 2.7821, $\hat{b}$ = 0.9948

und ein $D$ = 0.973. Die so gewonnene Anpassung ist in Abbildung 4.9 dargestellt. Optisch ergibt sie kein so gutes Bild wie die Kurven von Tuldava und Köhler-Martináková, sie ist aber ohne weiteres akzeptierbar.

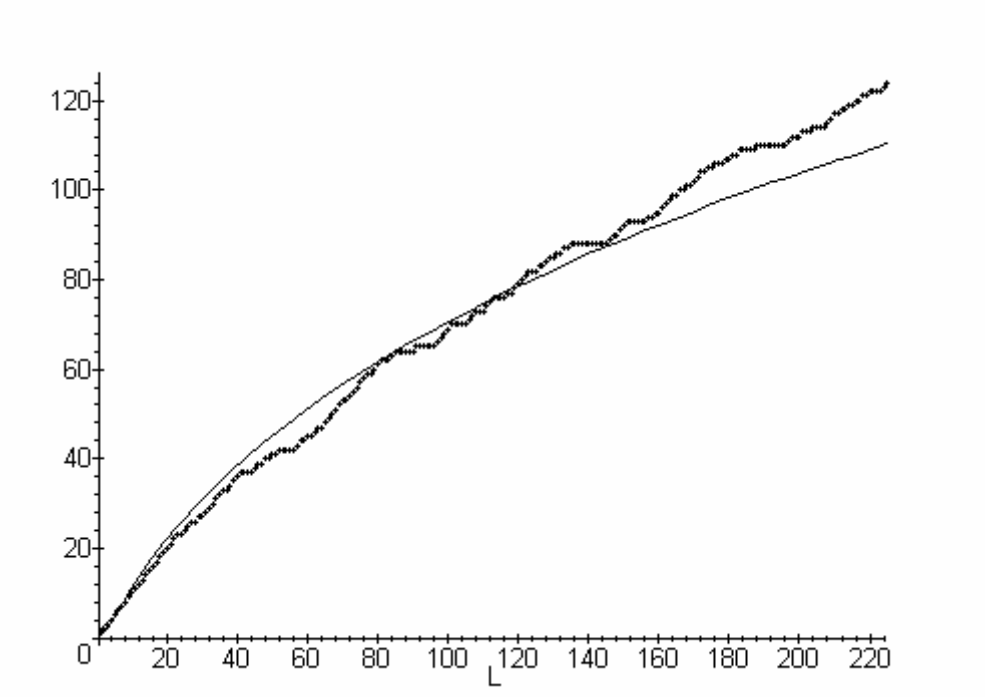


Abbildung 4.9. Anpassung der Kurve von Brunet an die types im „Erlkönig“.

**(5) Maas**

Aus dem Index von Maas (1972) erhält man die Kurve

(4.53) $V_i = L_i(\ln L_i)^a$,

für die man durch Linearisierung

(4.54) $\ln V_i = \ln L_i + a \ln(\ln L_i)$

mit Hilfe der kleinsten Quadrate den Parameter *a* als

(4.55) $$\hat{a} = \frac{\sum_{i=2}^{n} \ln(\ln\ L_i) \ln V_i - \sum_{i=2}^{n} \ln L_i \ln(\ln\ L_i)}{\sum_{i=2}^{n} [\ln(\ln\ L_i)]^2}$$

schätzen kann. Daraus ergeben sich die Parameter *a* = -0.257375 und der Determinationskoeffizient *D* = 0.92. Graphisch ist die Anpassung in Abb. 4.10 dargestellt.

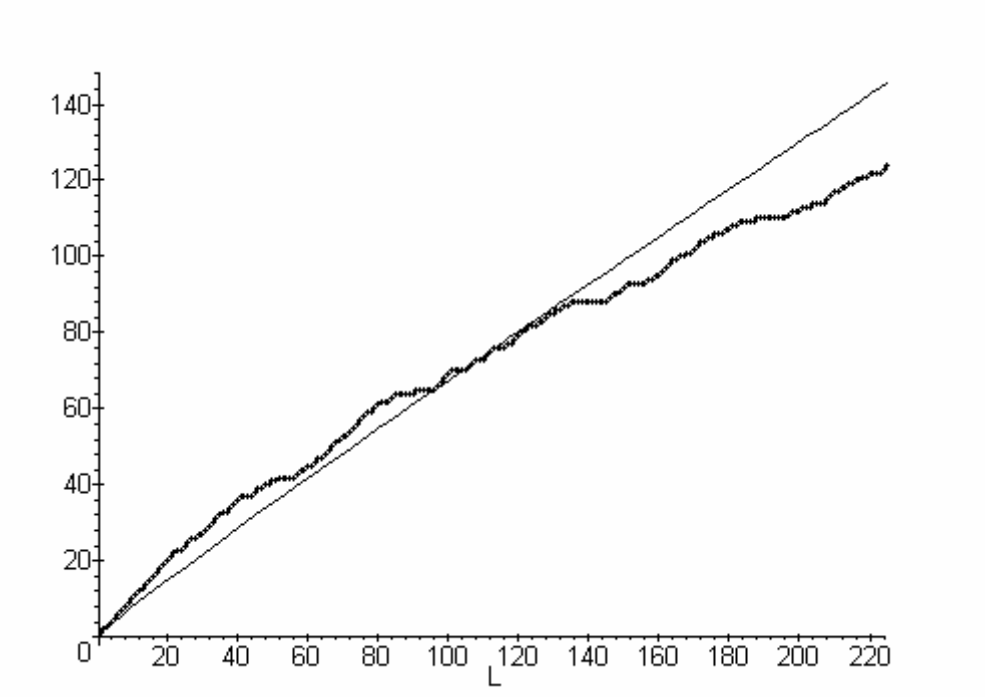


Abbildung 4.10. Anpassung der Kurve von Maas an die types im „Erlkönig“

Die Anpassung ist zwar gut akzeptabel, aber optisch etwas schlechter als z.B. die von Köhler-Martináková.

**(6) Dugast**

Aus dem Index von Dugast (1978,1979,1979a,1980) lässt sich die Kurve

(4.56) $V_i = L_i^{(1+a \ln L_i)}$

aufstellen. Hier bekommt man durch Logarithmieren die Form

(4.57) $\ln V_i = (1 + a \ln L_i) \ln L_i$

und aus dieser auf die bereits bekannte Weise

(4.58) $$\hat{a} = \frac{\sum_{i=1}^{n} (\ln L_i)^2 \ln V_i - \sum_{i=1}^{n} (\ln L_i)^3}{\sum_{i=1}^{n} (\ln L_i)^4}$$

was in unserem Fall $a$ = -0.01835 ergibt und eine Anpassungsgüte von $D$ = 0.987 signalisiert. Die Anpassung ist in Abbildung 4.11 dargestellt.

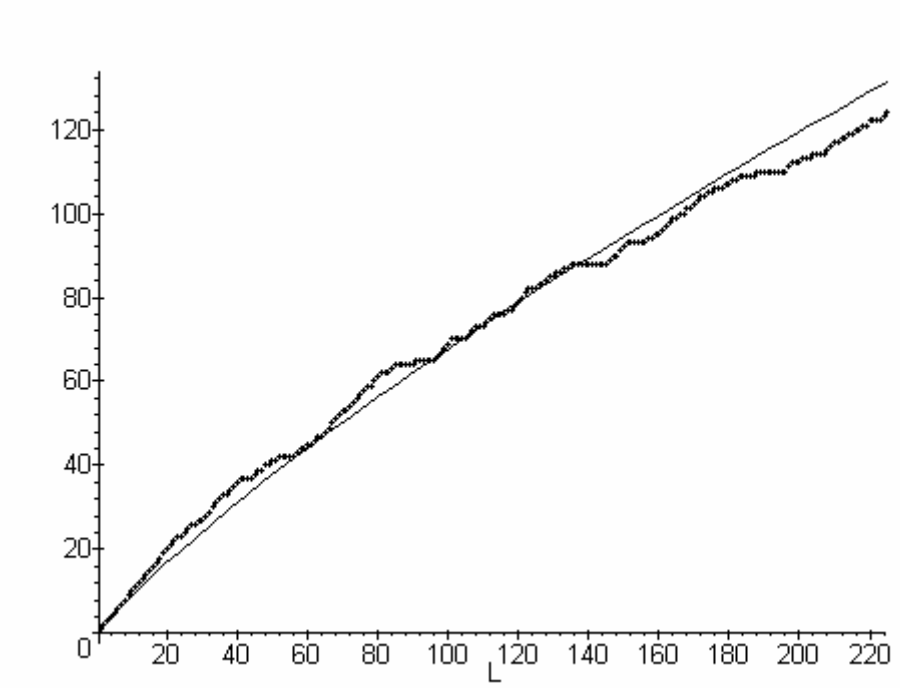


Abbildung 4.11. Anpassung der Kurve von Dugast an Types im Erlkönig

**(7) Orlov**

J. K. Orlov (s. Orlov, Boroda, Nadarejšvili 1982) ist es gelungen zu zeigen, dass es für einen Text den sogenannten „Zipfschen Umfang“ *Z* gibt (heutzutage als Zipf-Orlovscher Umfang bezeichnet), der sich für einen Text aus der im voraus „geplanten“, beabsichtigten Textlänge ergibt. Will jemand einen Roman verfassen, dann gestaltet sich der Informationsfluß im Text ganz anders, als wenn man beabsichtigt, nur einen Zeitungsartikel zu schreiben. Mit Hilfe des Zipfschen Umfangs kann man auch den Verlauf der types im Text rekonstruieren. Orlov hat hierfür die Formel

$$V_i = \frac{Z \ln(\frac{Z}{L_i})}{\ln(Zp)\left(\frac{Z}{L_i} - 1\right)}. \quad (4.59)$$

eingeführt. Hier ist *Z* einfach ein Parameter, *p* ist die relative Häufigkeit des häufigsten Wortes im Text. In unserem Text kennen wir diese Häufigkeit zwar, sie beträgt 11/225 = 0.0489 (für das Wort „mein“), aber wenn man TTR untersucht, kennt man sie normalerweise nicht, d.h., auch *p* ist nur ein Parameter, der geschätzt werden muss. (4.59) kann man alternativ schreiben als

$$(4.60)\quad V_i = \frac{Z(\ln Z - \ln L_i)L_i}{(\ln Z + a)(Z - L_i)}.$$

Hier ist $a = \ln p$. Da sich die Schätzungen als sehr komplizierte Formeln ergeben, wurde (4.60) in dieser Form iterativ den Daten angepasst. Es ergab sich $a$ = -2.935269 und $Z$ = 2755.94655 mit einem $D$ = 0.9985. Diese Anpassung ist in Abbildung 4.12 dargestellt. Sie ist von allen Anpassungen bisher die beste (für diese Daten). Orlov hat sie aber auch für viele andere Daten erfolgreich getestet. Es ist schwer, sich eine bessere Anpassung vorzustellen. Gleichzeitig kann festgestellt werden, dass $a$ nichts mit der größten empirischen relativen Häufigkeit zu tun hat und $Z$ als „geplanter Umfang" an der Realität völlig vorbeigeht. Es ist einfach nur ein Parameter. Nehmen wir aber diese Orlovs Interpretation als gegeben an, dann können wir (4.60) als eine dreiparametrige Kurve

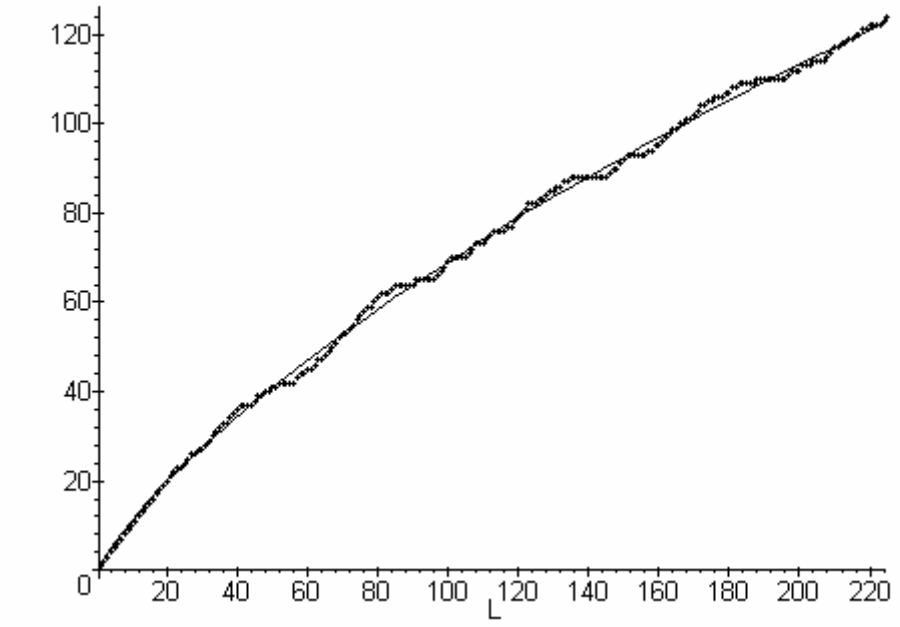


Abbildung 4.12. Anpassung der Orlov-Kurve an Types in Erlkönig

$$(4.61)\quad V_i = \frac{L_i(A - B\ln L_i)}{C(B - L_i)}$$

schreiben, wo $A = Z \ln Z$, $B = Z$ und $C = a + \ln Z$. Die Kleinstquadratschätzer ergeben aber hier äußerst lange Ausdrücke mit jeweils zwei Lösungen für $B$ und $C$, so dass auch dieser Weg sehr beschwerlich ist. Die iterative Anpassung von (4.54) ergibt jedoch genau dieselben Ergebnisse wie (4.60).

Tweedie und Baayen (1998) haben die Problematik von $Z$ erkannt und diese Größe durch eine Funktion $aL_i^b$ ersetzt, wodurch die Kurve wieder drei Para-

meter hat und bei der Kleinstquadratschätzung die gleichen Probleme auftauchen. Zumindest in unserem Fall bekommt man auf diese Weise keine besseren Resultate.

**(8) Somers**

Auf dem Wege zur Stabilisierung des „Wortschatzindexes“ haben viele Forscher mit Logarithmen gearbeitet. Somers (1959, 1966) hat das angestrebte Ziel durch eine doppelte Logarithmierung erreicht und den Index c = ln(ln V)/ln(ln L) aufgestellt. Löst man diese Formel nach V, so erhält man die Kurve

(4.62) $V_i = \exp(\ln^c L_i)$.

Die Schätzungen für den Parameter c erhält man recht einfach als

$$(4.63)\quad \hat{c} = \frac{\sum_{i=2}^{n} \ln(\ln V_i)}{\sum_{i=2}^{n} \ln(\ln N_i)}$$

mit dem man c = 0.94686 und $D$ = 0.9447 bekommt. Alternativ kann man $c$ abschätzen als

$$(4.64)\quad \hat{c} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{\ln(\ln V_i)}{\ln(\ln N_i)},$$

mit dem man $c$ = 0.95045 und $D$ = 0.9107 bekommt. Schließlich kann man auch mit dem Kleinstquadratschätzer

$$(4.65)\quad \hat{c} = \frac{\sum_{i=2}^{n} \ln(\ln V_i) \ln(\ln L_i)}{\sum_{i=2}^{n} [\ln(\ln N_i)]^2}$$

operieren, mit dem man $c$ = 0.94475 und $D$ = 0.9597 bekommt. Diese letzte Variante ist in Abbildung 4.13 dargestellt.

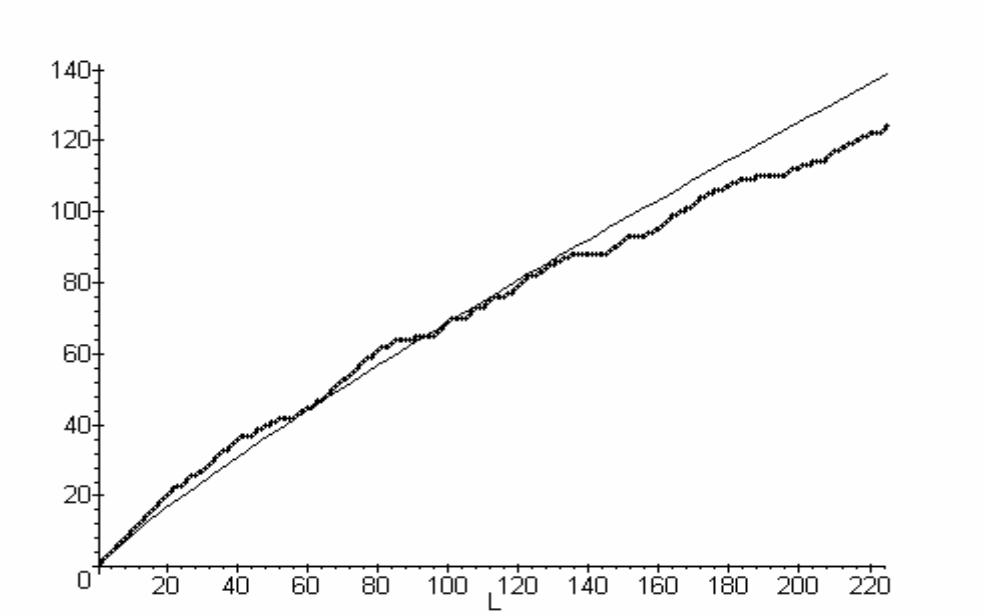


Abbildung 4.13. Anpassung der Somers-Kurve an die types im „Erlkönig“

**(9) Sichel**

Eine etwas kompliziertere Formel stammt von Sichel (1986), die für Daten der von uns untersuchten Art sehr geeignet ist:

(4.66) $V_i = \frac{2}{bc}\left[1 - e^{b(1-\sqrt{1+cL_i})}\right].$

Auch für diese Formel erhält man die Kleinstquadratschätzer nur mühsam. Es empfiehlt sich die iterative Anapassung, die $b$ = 0.130287, $c$ = 0.020518 und $D$ = 0.9982 liefert. Die graphische Darstellung ist in Abbildung 4.14 zu sehen

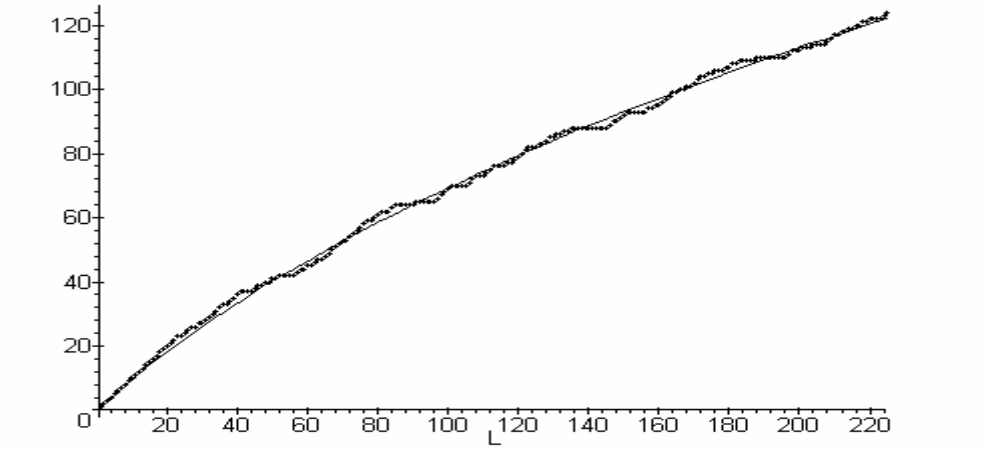


Abbildung 4.14. Anpassung der Sichel-Kurve an Types in Erlkönig

Es gibt zahlreiche andere Kurven dieser Art, die hier nicht alle angeführt werden können (zu Übersichten und einigen Eigenschaften dieser Kurven s. z.B. Pawlowski 1994; Wimmer, Altmann 1999).

Bei einer derartigen Fülle theoretischer Ansätze muss man unweigerlich einige Entscheidungen treffen und Schlüsse ziehen:

(1) Diese Kurven stellen nicht den Vokabularreichtum eines Textes oder eines Autors dar, sondern den Informationsfluss in einem Text. Auch die Annahme, die Asymptote einer Kurve (falls es sie gibt) repräsentiere die Menge der Wörter, die der Autor für das Schreiben des gegebenen Textes „zur Verfügung gestellt hatte", ist abwegig. Vokabularreichtum ist ein bisher nur ungenügend explizierter Begriff.

(2) Der Informationsfluss ist das Resultat eines sowohl selbstorganisatorischen als auch eines selbstregulierenden Prozesses, in dem so viele Faktoren eine Rolle spielen, dass es nie möglich sein wird, sie alle detailliert zu separieren und zu identifizieren. Auch eine Befragung der Autoren – falls sie überhaupt möglich ist – wäre müßig. Wir werden uns in Zukunft damit abfinden müssen, dass es hier zahlreiche Attraktoren gibt (die unterschiedliche Kurven repräsentieren) und dass sich ´mal eine, ´mal eine andere Kurve bewähren wird. Es ist viel empirische Arbeit nötig, um herauszufinden, welche Kurven für welche Texte „am besten" sind, ob bestimmte Kurven sogar für bestimmte Textsorten geeignet sind.

(3) Bei der Anpassung einer Kurve an TTR-Daten wähle man zuerst die einfachste, d.h. diejenige mit den wenigsten Parametern und mit einfachen Funktionen, und kompliziere das Verfahren nur dann, wenn es nötig ist, z.B., wenn man dadurch eine starke Zunahme des Wertes des Determinationskoeffizienten erzielt

(4) Zwar haben wir an den meisten Stellen Schätzer angegeben, aber es ist empfehlenswert, bei der Kurvenanpassung möglichst ein iteratives Verfahren zu benutzen. Diese selbst fangen entweder mit festen Werten oder mit der Methode der kleinsten Quadrate an und verbessern die Anpassung bei jeder Iteration schrittweise. Fertige Software wie NLREG oder MATLAB sind hier sehr hilfreich.

## 4.4. Wortlänge

Die Wortlänge verhält sich in Texten bekanntlich sehr regulär, und die Erforschung dieser Regularität gehört zu den ältesten Forschungsproblemen der quantitativen Linguistik. In Deutschland war es der Physiker W. Fucks, der als erster zu einer Gesetzeshypothese gelangte (Fucks 1955a,b, 1956). Diese wurde später von Grotjahn (1982) erweitert. Heutzutage gibt es bereits eine ausgebaute Theorie der Wortlänge (vgl. Grotjahn, Altmann 1993; Wimmer et al. 1994; Wimmer, Altmann 1996; Wimmer, Witkovský, Altmann 1999; Uhlířová, Wimmer 2003 u.a.) und die empirische Untersuchung der Wortlänge ist wohl das umfang-

reichste Forschungsprojekt der quantitativen Linguistik (vgl. Schmidt 1996; Best 1997, 2001; Grzybek 2005; http://gwdu05.gwdg.de/~kbest/).

Die Wortlänge wird als Anzahl der Silben gemessen, weil die Silbe phonetisch die unmittelbare Komponente des Wortes ist. Andere Möglichkeiten sind aber nicht ausgeschlossen, z.B. die Zählung der Phoneme, der Buchstaben, der Morpheme, der Moren, und einige Eigenschaften der Länge bleiben bei jeglicher Art von Messung erhalten (vgl. Hřebíček 1997). Schwierigkeiten, wie sie bei der Längenzählung entstehen können, werden in der angegebenen Literatur ausführlich besprochen, der Leser wird auf diese Literatur verwiesen. Die Wortlänge im Text folgt einem bestimmten Rhythmus, der sich im Laufe der Textentfaltung ändern, durch Eingriffe in den Text gestört werden und besonders nach Pausen Brüche aufweisen kann. Es gibt drei Arten der Einheiten, deren Länge gemessen wird: Wörter aus dem Wörterbuch, wo sich nur Lemmata befinden; Wörter aus einem homogenen Text, wo sich wiederholte Wortformen befinden und Wörter aus dem Frequenzwörterbuch einer Sprache, das eine Mischung von Texten enthält. Es empfiehlt sich, die Wortlänge in Einzeltexten zu messen (nicht in Textmischungen etwa wie in Korpora) und bei verallgemeinernden Schlüssen sehr vorsichtig zu sein, da man dabei Aussagen über Grundgesamtheiten trifft, die es gar nicht gibt. So gibt es keine Wortlängenverteilung für das Deutsche oder für die Sprache Goethes oder für die Poesie usw. Schon die Mischung von zwei Texten kann zu Verzerrungen führen, die sich bei der Anpassung eines Modell stark auswirken (vgl. Altmann 1992). Fucks´ Versuch, für alle Sprachen eine einzige Verteilung zu finden, endete mit einer gemischten Verteilung, bei der man die Parameterzahl so lange vergrößern muss, bis die Verteilung endlich passt – ein Trick, den man meiden sollte. Auch wenn man in bestimmten Sprachen oft eine bevorzugte Verteilung findet, sollte man nicht annehmen, dass sich diese in allen Texten durchsetzt. Die Wortlängentheorie resultiert nicht in einer einzigen Verteilung, sondern in ganzen Familien von Verteilungen.

Sowohl phylogenetisch als auch ontogenetisch gehen wir davon aus, dass es „am Anfang“ einsilbige Wörter gab und gibt (oder Wiederholungen der gleichen Silbe), erst später, wenn durch zu viele Monosyllaba die Redundanz verringert wird und komplexe Sachverhalte lexikalisch erfasst werden müssen, entwickelt man zweisilbige Wörter. Diese entwickeln sich in einem proportionalen Verhältnis zu den Einsilbigen, etwa in der Form

$$P_2 \sim P_1,$$

wo $P_i$ die Wahrscheinlichkeit (oder die relative Häufigkeit) der Klassen der Länge $i$ bedeutet. Da diese Proportionalität nicht konstant zu sein braucht, sondern sich bei unterschiedlichen Längenklassen ändern kann, entwickelt sich eine Funktion $g(x)$, die die Häufigkeiten steuert, so dass wir schreiben können

$$P_x = g(x)P_{x-1}. \quad (4.67)$$

Je nachdem, wie man die Funktion $g(x)$ einsetzt, erhält man unterschiedliche Verteilungen. Man interpretiert diese Funktion als $g(x) = s(x)/h(x)$, wo $s(x)$ ein dem Sprecher, $h(x)$ ein dem Hörer zugeschriebener Faktor ist. Diese Funktion sichert auch die Konvergenz der Verteilung.

Im „Erlkönig“ ist die Zählung der Wortlänge relativ einfach. Man erhält die Längenverteilung, wie in der ersten und der zweiten Spalte von Tabelle 4.6 dargestellt.

Tabelle 4.6
Silbische Länge der Wortformen im Erlkönig

| Silbenzahl im Wort $x$ | Beobachtete Häufigkeit $f_x$ | Theoretische Häufigkeit $NP_x$ |
|---|---|---|
| 1 | 152 | 153.85 |
| 2 | 65 | 58.48 |
| 3 | 6 | 11.12 |
| 4 | 2 | 1.55 |
| | $a = 0.3801$, $X^2 = 3.23$, FG = 2, $P = 0.20$ | |

Obwohl hier die Art des Textes den Verfasser möglicherweise zur Wahl kürzerer Wörter gezwungen hat, nehmen wir auch in einem Fall wie diesem an, dass sich die Verteilung „regulär“ verhält und einem „Attraktor“ folgt. Um diesen Attraktor zu erfassen, halten wir uns an die vorhandene Theorie und wählen zunächst eine Funtion $g(x)$ mit kleiner Anzahl von Parametern, die sich besonders bei Daten mit wenigen Klassen bewähren. Es besteht nämlich die Gefahr, dass die Freiheitsgrade (FG) für das Testen des Modells sonst nicht ausreichen. In unserem Fall wählen wir

$$g(x) = \frac{a}{x-1}$$

und erhalten die Differenzgleichung

$$(4.68)\quad P_x = \frac{a}{x-1} P_{x-1}, \quad x = 2,3,4,...$$

Die Lösung ergibt die von W. Fucks ursprünglich eingeführte 1-verschobene Poisson-Verteilung

$$(4.69)\quad P_x = \frac{e^{-a} a^{x-1}}{(x-1)!}, \quad x = 1,2,3,...$$

Zwar ist die Schätzung des Parameters *a* in diesem Fall sehr einfach, wir benutzen für die Anpassung trotzdem die vorhandene Software (Altmann-Fitter 1997), die uns die Werte in der letzten Spalte von Tabelle 4.6 liefert. Die berechnete Wahrscheinlichkeit $P = 0.20$ signalisiert eine zufriedenstellende Anpassung, die graphisch in Abbildung 4.15 dargestellt ist. Es gibt sicherlich auch besser passende Verteilungen, aber da die Poisson-Verteilung sehr einfach ist und nur einen Parameter hat, besteht kein Bedarf, nach einer anderen zu suchen.

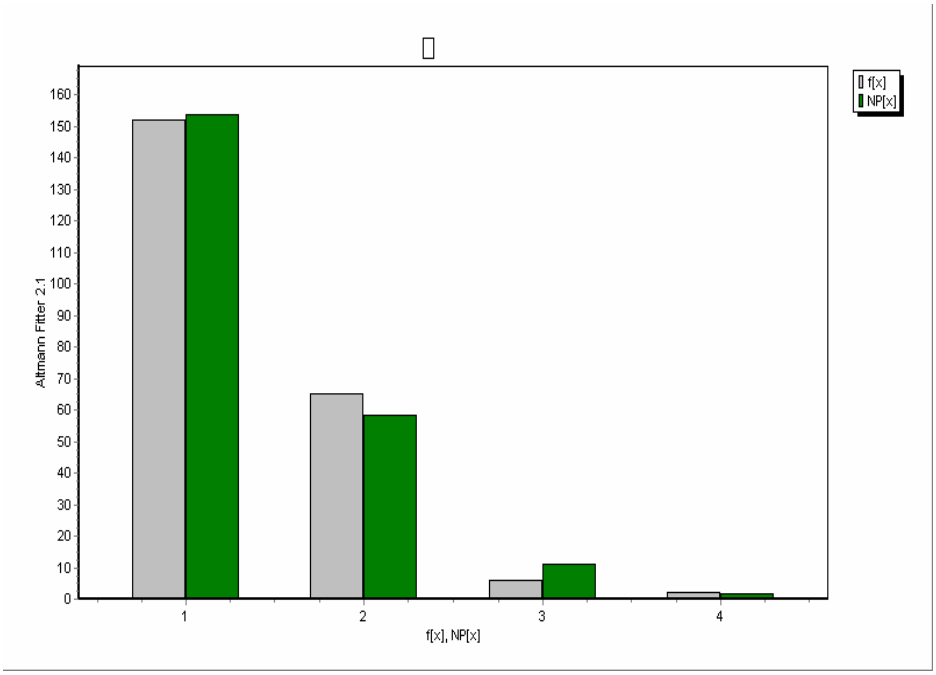


Abbildung 4.15. Anpassung der 1-verschobenen Poisson-Verteilung an die Wortlängen im „Erlkönig“

Die Wortlänge ist mit vielen anderen Eigenschaften des Wortes selbst und mit Eigenschaften anderer Entitäten verbunden (vgl. Köhler 1986, 2002). Diese Zusammenhänge sind aber nicht Eigenschaften des gegebenen Textes selbst, sondern eher allgemeine Sprachmechanismen. In einem individuellen Text können sie aber eine spezielle Form annehmen. Bekanntlich ist die Wortlänge von der Worthäufigkeit abhängig, hängt zusammen mit der Polysemie, der Synonymiezahl, der Polytextie, der Phonemdistribution, der Existenz von Tönen, dem Synthetismus der Sprache, mit der Silben- und der Klausenlänge (Menzerathsches Gesetz) usw. Jedoch ist unser Text zu kurz, um zumindest einige von diesen Zusammenhängen markant aufzuweisen, manche lassen sich unter Zuhilfenahme des Wörterbuchs oder eines Korpus ermitteln.

# 5. Denotative Analyse

Die klassischen Spracheinheiten wie Phonem, Silbe, Wort, Satz u.a. sind nicht die einzig möglichen und nicht einmal die „richtigen“ oder „wahren“ oder „natürlichen“ Einheiten. Sie sind keine Bestandteile der Realität, sie korrespondieren lediglich in bestimmter Weise mit realen Entitäten, sind also unsere konzeptuellen Konstrukte, die wir zunächst intuitiv, später mit Hilfe bestimmter Kriterien – die wiederum keineswegs dateninhärent sind – etablieren. Spracheinheiten sind also nicht gegeben sondern konstruiert. Diese Tatsache wird uns klar, wenn wir ein Gespräch in einer Sprache hören, von der wir kein einziges Wort verstehen. Jeder Linguist weiß, wie schwer es ist, bereits etablierte Einheiten in einem Text zu identifizieren und zu segmentieren. Die Literatur darüber füllt eine umfangreiche Bibliothek. Die Diskussion über Segmentierung und Identifizierung rührt daher, dass man meint, mit einem Wort, das mit einem Begriff assoziiert ist, eine reale Entität exakt bestimmen, erfassen zu können. Hinter diesem Glauben steckt ein Überbleibsel der „Magie des Wortes“, wie Carnap es treffend bezeichnet hat. Dieser Irrtum führt zu dem Glauben, dass es in der Natur z.B. Zentimeter gäbe, und beruht auf der Vermischung der Ontologie mit der Methodologie. Die Dinge, die in der Natur existieren, sind nicht identisch mit den Begriffen, die wir über sie ersinnen. Wenn wir aber die etablierten Begriffe anwenden möchten, dann brauchen diese weder „richtig“ noch „falsch“ zu sein, sondern müssen lediglich fruchtbar sein, d.h., sie müssen bestimmten Zwecken dienen, ob es nun unsere Orientierung in der Natur, deren Beschreibung, deren Klassifikation oder, im besten Falle, ihre Anwendung bei der Formulierung von Hypothesen ist. Von der theoretischen Seite her ist die letztgenannte Anwendung maßgebend. Ist die Forderung nach Anwendbarkeit in Hypothesen nicht erfüllt, dann sind die Begriffe theoretisch unfruchtbar. In einer Theorie sind sie nur Ziegelsteine, die man mit dem Mörtel der Hypothesen verbindet.

Von dieser Warte aus hat der Textanalytiker also zunächst volle Freiheit. Er kann nach Belieben neue Begriffe prägen und sie Wörtern zuordnen – und dies ist auch die übliche Praxis in den Geisteswissenschaften. Alles andere bereitet nur Schwierigkeiten und wird gerne vernachlässigt. Ein Begriff muss nämlich zuerst operationalisiert werden, damit man sein Korrelat in Texten oder in der Sprache eindeutig identifizieren kann. Der Begriff „sozialer Status“ etwa kann auf zahlreiche unterschiedliche Weisen operationalisiert werden, und keine davon ist wahr oder falsch, sondern nur fruchtbar oder unfruchtbar. Weiter muss ein Begriff quantifiziert werden, d.h., es müssen Regeln aufgestellt werden, wie man ihm Werte zuschreiben kann, damit sich mit seiner Hilfe testbare Hypothesen aufstellen lassen. Schließlich müssen Messungen durchgeführt und die aufgestellten Hypothesen getestet werden. Die Fruchtbarmachung eines Begriffes ist daher kein einfacher Vorgang, sondern ein langwieriger Prozess der Reifung einer Disziplin. Ohne solche Verfahren ist ein Begriff nur eine Schublade, in die bestimmte Erscheinungen hineingelegt werden. Heute wissen wir aber, dass diese

Schubladen nicht immer geschlossen sind, sondern vage (fuzzy) Mengen bilden, denen Sprach- oder Textentitäten nur bis zu einem bestimmten Grade angehören.

Analysiert man einen Text auf der Wortebene, dann kann man entweder Wortformen berücksichtigen – wie wir es gemacht haben –, oder man kann diese auf Lemmata reduzieren. Eine weitere Reduktion kann sich dadurch ergeben, dass alle Entitäten, die die gleiche reale Entität denotieren, zusammengefasst werden. Es sind irgendwelche denotativen Einheiten, die wir unter dem Namen „Denotationshreb" analysieren werden. Die Denotationshrebs sind von Ziegler und Altmann (2002) eingeführt worden und bilden eine Analogie zu den Satzhrebs auf der Satzebene. Satzhrebs, benannt nach ihrem Entdecker L. Hřebíček (s. besonders 1993a, 1997), sind Gruppen von nicht notwendigerweise benachbarten Sätzen, die entweder das gleiche Zeichen (Wort, Wortform) enthalten oder aufeinander referieren. Denotationshrebs sind Gruppen von nicht unbedingt gleichlautenden formalen Einheiten (Wortteilen, Wörtern, Phrasen, …), die sich auf dieselbe reale Entität beziehen, dasselbe denotieren. Der Begriff selbst wurde unter dem Namen „nominative Kette" bereits von Viehweger (1978) eingeführt, die Operationalisierung, Eigenschaften, Hypothesen und Testprozeduren wurden von Ziegler und Altmann (2002, 2003), Ziegler, Jüngling, Altmann (2005) besprochen.

An dieser Stelle werden wir den Denotationshreb durch eine Anzahl von Kriterien etwas anders erfassen als in Ziegler, Altmann (2002), um die Tragweite dieses Begriffes zu testen. Kriterien sind zum Glück lediglich Operationsanleitungen, die sich gemäß unserem Wissensstand ändern können. Sie sind Konventionen, sie können nur effektiv oder ineffektiv sein.

## 5.1. Etablierung von Denotationshrebs

Bei der Etablierung von Hrebs geht man einen Text Wortform für Wortform bzw. Morph für Morph, Phrase für Phrase usw. durch und bildet Mengen, wobei eine Menge alle Wortformen, Wortteile oder Phrasen enthält, die dieselbe Denotation haben, sich auf dasselbe in der Realität oder im Text beziehen. Hier lassen wir Phrasen beiseite. Wie weiter gezeigt wird, ist es notwendig, die Positionen, in denen sich Hrebelemente befinden, zu nummerieren, denn berücksichtigt man auch Wortteile als denotative Elemente, dann können dort, wo nur ein Wort steht, mehrere Positionen gleichzeitig vorhanden sein. Wir befolgen folgendes Sparsamkeitsprinzip

> *Soweit wie möglich soll man bei der Etablierung der Hrebs sparsam umgehen, d.h., der Text muss zwar vollständig zerlegt werden, aber die Zahl der Hrebs soll mit Hilfe von Kriterien niedrig gehalten werden.*

Die Kriterien sind eigentlich Entscheidungshilfen und können bei Bedarf nach semantischen, grammatischen oder textologischen Doktrinen usw. variiert wer-

den. Hier wird nur eine Variante möglicher Kriterien benutzt, der Anwender soll sich frei fühlen, andere Kriterien zu wählen, da diese Art der Textanalyse erst im Entstehen ist.

### *Kriterien*

(1) **Komposita** soll man nicht in ihre Bestandteile zerlegen. So bildet z.B. „Nebelstreif " nur einen Hreb, wir zerlegen das Wort nicht in „Nebel“ und „Streif“, es sei denn, beide Bestandteile kommen im Text auch unabhängig vor und werden als Hrebs etabliert. In anderen Sprachen kann man die Wörter als eine Einheit betrachten, auch wenn die einzelnen Bestandteile unabhängig vorkommen. So bildet das französische "chemin de fer" einen Hreb und nicht drei, das indonesische "kereta api" (Wagen + Feuer = ) Zug auch nur einen Hreb. Ein Konstrukt aus mehreren Wörtern denotiert eine Entität, die als Ganzheit betrachtet wird. Man kann diese Entscheidung auch dann treffen, wenn man nicht sicher ist, ob man es mit einem Kompositum, einer Phrase oder einem Idiom zu tun hat. Zerlegt man ein Kompositum, so betrachten wir die resultierenden Hrebs als deterministisch koinzident (s.u.).

(2) Ein Wort kann gleichzeitig zu **mehreren Hrebs** gehören. Dies betrifft vor allem Verben, die Personalendungen aufweisen, und Nomina mit Possessivaffixen. So empfiehlt es sich, das Verb „birgst“ im 5. Vers zwei Hrebs zuzuordnen, nämlich „bergen“ und „du“ = Sohn. Die Endung der dritten Person bei Verben bezieht sich immer auf das entsprechende Nomen oder Pronomen. Auch wenn im Deutschen in der ersten Person manchmal die Personalendung entfällt („spiel ich“ in Vers 10, „brauch ich“ in Vers 26), signalisieren diese Formen dennoch die 1. Person Sg. (oder eine andere Person im Imperativ). In einigen Sprachen gibt es Possessivsuffixe, z.B. signalisiert in dem ungarischen Wort "házam" mein Haus „-am“ die erste Person. Die resultierenden Hrebs betrachtet man als deterministisch koinzident.

(3) **Zusammengesetzte Verben** bzw. Verben mit abtrennbaren Präfixen („er hält ihn warm“, „faßt er mich an“) bilden auch bei Trennung nur einen einzigen Hreb: „warm halten“, „anfassen“.

(4) **Artikel** und **Demonstrativpronomina** kann man sozusagen als Ausdruck unserer Wissenslage (bekannt, unbekannt) oder Objektnähe (dieser, jener) auffassen und dem Nomen zuordnen. Im Grunde müsste man in Sätzen wie „Gestern sah ich eine Schlange. Schlangen sind giftig. Diese Schlange nicht“ drei Hrebs für „Schlange“ etablieren, weil sie unterschiedliche Dinge denotieren. Aufgrund des Sparsamkeitsprinzips (s.o.) etablieren wir aber nur einen Hreb − je nach Umständen. In einigen Sprache ist der Artikel ein nominales Affix (Bulgarisch).

(5) **Polysemantische Wörter** sind am schwierigsten zu behandeln. Man kann eventuell ihre Bedeutung durch Paraphrasierung exakter bestimmen. So gibt es zahlreiche Bedeutungen vom Verb „sein“, z.B. „Es ist der Vater ...“

(Identifikation) „... Blumen sind an dem Strand“ (= sich an einem Ort befinden), „sei ruhig“ (= sich verhalten) oder „X ist gut“ (= Eigenschaft haben), „X ist Lehrer“ (= zu Klasse Y gehören) usw.; das Verb „haben“ hat im Erlkönig drei Bedeutungen (halten, besitzen, Perfekt); „so“ wird als Adverb und als Konjunktion gebraucht („so spät“, „so brauch ich Gewalt“) usw. Die einzelnen Bedeutungen findet man in jedem größeren Wörterbuch. Diese Unterscheidung der Bedeutungen, die korrekt ist, kann man eventuell ignorieren, weil die gegebene Sprache dies eben auf diese Weise gestaltet, d.h. die Bedeutung diversifiziert hat. Zu detaillierte Unterscheidung der Bedeutungen führt zur Vermehrung der Hrebs. Auf jeden Fall sollte man am Anfang differenzieren, denn zusammenfassen kann man auch nachträglich.

(6) Die **Negation** bei Verben wird nicht als separater Hreb, sondern als Teil des Verbs betrachtet. In einigen Sprachen wird sie mit einem Suffix (Japanisch), in anderen mit einem Präfix (slawische Sprachen) gebildet.

(7) **Präpositionen** und **Konjunktionen** muss man je nach Sprache behandeln. Man kann im Deutschen z.B. „und“, „mit“, „zusammen“ in einen oder in mehrere Hrebs platzieren, je nach den Umständen. Ortbedeutungen drückt man im Deutschen mit Präpositionen aus, im Japanischen mit Postpositionen, im Ungarischen mit Suffixen. Falls die Ortbedeutung relevant ist, so kann man sie als Hreb etablieren (so z.B. bei Verben im Nimboranischen, Neu Guinea), im Ungarischen kann man sie auch ignorieren. Das Deutsche liegt zwischen diesen Extremen, so dass man eine der beiden Alternativen wählen kann. Das Sparsamkeitsprinzip spricht eher für die Ignorierung der Präpositionen mit Ortbedeutung als separate Hrebs.

(8) Hat man Bedenken über die Aufteilung eines Wortes oder seine Zuordnung zu mehreren Hrebs, dann sollte man sparsam vorgehen.

Zur Illustration analysieren wir ausführlich die erste Strophe unseres Gedichtes:

> Wer reitet so spät durch Nacht und Wind?
> Es ist der Vater mit seinem Kind;
> Er hat den Knaben wohl in dem Arm,
> Er faßt ihn sicher, er hält ihn warm.

Hier ergeben sich folgende Hrebs (Elemente in der Wortform angegeben):

[Wer, reite_**t**, der Vater, seinem, er, ha_**t**, er, faß_**t**, er, warmhäl_**t**]
[Kind, den Knaben, ihn, ihn]
[es, is_$\mathbf{t}_1$]
[reiten], [$so_1$], [spät], [durch], [Nacht], [$und_1$], [Wind], [$ist_1$], [mit], [$hat_1$], [wohl], [in], [dem Arm], [fasst], [sicher], [warmhält]

Vier Wörter haben wir mit dem Index 1 gekennzeichnet, weil es von ihnen noch andere Vorkommen mit anderen Bedeutungen gibt. Man könnte überlegen, ob man [und] und [mit] zusammenlegen sollte, weil sie hier dieselbe Beziehung denotieren.

In sparsamerer Fassung würden wir den Hreb [in] nicht etablieren, weil die Präposition zu [in dem Arm] gehören würde; [durch] würde zu [durch Nacht] gehören.

Bei einer denotativen Analyse, die eine Zerlegung des Textes mit Hilfe von Phrasen anstrebt, werden die meisten oben genannten Kriterien und Entscheidungen irrelevant (R.Köhler, pers. Mitteilung).

Man kann nun die Strophe als Folge von Positionen schreiben, an denen die Hrebs stehen. Mit einem Schrägstrich zwischen zwei Positionen bezeichnen wir Wörter, die zu zwei Hrebs gehören, unterstrichen sind Hrebs die aus zwei Wörtern bestehen. Diese werden nebeneinander in die Position des ersten Elements gestellt.

Für den „Erlkönig“ erhalten wir

Wer reitet so spät durch Nacht und Wind?
1 2/3 4 5 6 7 8 9
Es ist der Vater mit seinem Kind;
10 11/12 13 14 15 16
Er hat den Knaben wohl in dem Arm.
17 18/19 20 21 22 23
Er fasst ihn sicher, er warmhält ihn.
24 25/26 27 28 29 30/31 32

Mein Sohn, was birgst du so bang dein Gesicht?
33 34 35 36/37 38 39 40 41 42
Siehst nicht, Vater, du den Erlkönig?
43/44 45 46 47
Den Erlkönig mit Kron und Schweif?
48 49 50 51 52
Mein Sohn, es ist ein Nebelstreif.
53 54 55 56/57 58

Du, liebes Kind, komm, geh mit mir!
59 60 61 62/63 64/65 66 67
Gar schöne Spiele spiel ich mit dir;
68 69 70 71/72 73 74 75
Manch Blumen bunte sind an dem Strand,
76 77 78/79 80 81
Meine Mutter hat manch Gewand gülden.
82 83 84/85 86 87
Mein Vater, mein Vater, und hörest nicht du,

88 89 90 91 92 93/94 95
Was Erlenkönig mir leise verspricht?
96 97 98 99 100/101?
Sei ruhig, bleibe ruhig, mein Kind:
102/103 104 105/106 107 108 109
In dürren Blättern säuselt der Wind
110 111 112 113/114 115

Willst, feiner Knabe, du mit mir gehn?
116/117 118 119 120 121 122 123
Meine Töchter sollen dich warten schön;
124 125 126/127 128 129 130
Meine Töchter führen den Reihn nächtlichen
131 132 133/134 135 136
Und wiegen und tanzen und einsingen dich.
137 138/139 140 141/142 143 144/145 146

Mein Vater, mein Vater, und siehst nicht du dort
147 148 149 150 151 152/153 154 155
Erlkönigs Töchter am düstern Ort?
156 157 158 159 160

Mein Sohn, mein Sohn, ich sehe es genau:
161 162 163 164 165 166/167 168 169
Es scheinen die Weiden alten so grau.
170 171/172 173 174 175 176

Ich liebe dich, mich reizt deine schöne Gestalt;
177 178/179 180 181 182/183 184 185 186
Und bist nicht du willig, so brauch ich Gewalt.
187 188/189 190 191 192 193/194 195 196
Mein Vater, mein Vater, jetzt faßt an er mich!
197 198 199 200 201 202/203 204 205
Erlkönig hat getan mir ein Leids!
206 207/208 209 210

Dem Vater grausets, er reitet geschwind,
211 212/213 214 215/216 217
Er hält in Armen das Kind ächzende,
218 219/220 221 222 223 224
Erreicht den Hof mit Mühe und Not:
225/226 227 228 229 230 231
In seinen Armen das Kind war tot.
232 233 234 225 236/237 238

Es gibt verschiedene Auflistungsmöglichkeiten für Hrebs:
(1) *Datenhrebs* mit Angabe des Wortes (der Wörter, der Wortteile) und der Position. Man schreibt solche Hrebs in runden Klammern, Wortteile markiert man mit Fettdruck, falls möglich, oder besser mit tiefgesetztem Bindestrich, weil dieser auch dann sichtbar ist, wenn das Morphem fehlt, z.B. „komm_“. Aus den Datenhrebs kann man den Text vollständig rekonstruieren. Beispiel:

Vater = (wer 1, reite_**t** 3, der Vater 13, seinem 15, er 17, ha_**t** 18, er 24, faß_**t** 26, er 29, warmhäl_**t** 31, mein 33, sieh_**st** 44, Vater 45, du 46, mein 53, Vater 89, Vater 91, hör_**est** 94, du 95, Vater 148, Vater 150, sieh_**st** 153, du 154, mein 161, mein 163, ich 165, seh_**e** 167, Vater 198, Vater 200, dem Vater 211, er 214, reite_**t** 261, er 218, häl_**t** 220, erreich_**t** 226, seinen 233)

(2) *Listenhrebs* enthalten nur die Wörter ohne Position. Man schreibt sie in eckige Klammern. In Tabelle 5.1 werden sie nicht aufgeführt, weil sie mit den Datenhrebs − bis auf die Positionen − identisch sind. Beispiel:

Vater = [wer, reite**t**, der Vater, seinem, er, ha**t**, er, faß**t**, er, warmhäl**t**, mein, sieh**st**, Vater, du, mein, Vater, Vater, hör**est**, du, Vater, Vater, sieh**st**, du, mein, mein, ich, seh**e**, Vater, Vater, dem Vater, er, reit**et**, er, häl**t**, erreich**t**, seinen]

(3) *Mengenhrebs*, in denen nur die Lemmata, Morpheme oder denotierenden Kategorien aufgeführt werden. Sie werden in geschweifte Klammern gesetzt. Beispiel:

Vater = {wer, Vater, sein, er, mein, du, ich, Personalendung .P.Sg, Personalendung 2.P.Sg}

(4) *Positionshrebs*, die nur die Positionen enthalten und mit spitzen Klammern symbolisiert werden. Beispiel:

Kind = <16,20,27,32,34,37,38,41,54,59,61,63,65,75,88,90,98,103,106, 109,116,119,120,128,146,147,149,162,164,180,184,189,190, 197,199,205,209, 223,235,237> .

Mengenhrebs werden nur dann angegeben, wenn die Zahl der Elemente größer als 1 ist. Listenhrebs und Positionshrebs kann man mühelos aus den Datenhrebs gewinnen.

Analysiert man das ganze Gedicht und ordnet dann die Hrebs nach ihrem Umfang, so bekommt man die in Tabelle 5.1 angegeben Resultate. In der Spalte ganz rechts ist die Zahl aller Elemente in einem Listenhreb und darunter die Zahl der Grundbegriffe in einem Mengenhreb angegeben. In der Tabelle haben

wir nur die „großen“ Hrebs detailliert angegeben, in den restlichen kann man die anderen Hrebtypen leicht rekonstruieren.

Tabelle 5.1
Hrebs im „Erlkönig“

| | | |
|---|---|---|
| 1 | **Kind**<br>(Kind 16, den Knaben 20, ihn 27, ihn 32, Sohn 34, birg_**st** 37, du 38, dein 41, Sohn 54, du 59, Kind 61, komm_ 63, geh_ 65, dir 75, mein 88, mein 90, mir 98, sei_ 103, bleib_**e** 106, Kind 109, will_ st 116, Knabe 119, du 120, dich 128, dich 146, mein 147, mein 149, Sohn 162, Sohn 164, dich 180, deine 184, bi_**st** 189, du 190, mein 197, mein 199, mich 205, mir 209, das Kind 223, das Kind 235, war_ 237)<br>{Kind, Knabe, er, Sohn, du, dein, mein, ich, Personalendung 2.P.Sg., Personalendung 2.P.Sg.Imperativ, Personalendung 3.P.Sg Imperfekt}<br><16,20,27,32,34,37,38,41,54,59,61,63,65,75,88,90,98,103,106,109, 116,119,120, 128,146,147,149,162,164,180,184,189,190,197,199,205, 209,223, 235,237> | 40<br><br><br><br><br><br><br><br><br><br><br>11 |
| 2 | **Vater**<br>(wer 1, reite_**t** 3, der Vater 13, seinem 15, er 17, ha _**t** 18, er 24, faß_**t** 26, er 29, warmhäl_**t** 31, mein 33, sieh_**st** 44, Vater 45, du 46, mein 53, Vater 89, Vater 91, höre_**st** 94, du 95, mein 108, Vater 148, Vater 150, sieh_**st** 153, du 154, mein 161, mein 163, ich 165, seh_**e** 167, Vater 198, Vater 200, dem Vater 211, er 214, reite_**t** 261, er 218, häl_**t** 220, erreich_**t** 226, seinen 233)<br>{wer, Vater, sein, er, mein, du, ich, Personalendung 3.P.Sg, Personalendung 2.P.Sg} | 37<br><br><br><br><br><br><br><br>9 |
| 3 | **Erlkönig**<br>(den Erlkönig 47, den Erlkönig 48, mir 67, spiel_ 72, ich 73, meine 82, Erlenkönig 97, versprich_**t** 101, mir 122, meine 124, meine 131, Erlkönigs 156, ich 177, lieb_**e** 179, mich 181, brauch_ 194, ich 195, anfaß_**t** 203, er 204, Erlkönig 206, hat getan 208)<br>{Erlkönig, ich, mein, er, Personalendung 2.P.Sg, Personalendung 3.P.Sg} | 21<br><br><br><br><br><br>6 |
| 4 | **und**<br>(und, 8, und, 51, und$_2$ 92, und, 137, 140, 143, und, 151, 187, und, | 9 |
| 5 | **Töchter**<br>(Töchter 125, soll_**en** 127, Töchter 132, führ_**en** 134, wieg_**en** 139, tanz_**en** 142, einsing_**en** 145, Töchter 157)<br>{Tochter, Personalendung 3.P.Pl.)} | 8<br><br><br>2 |
| 6 | **es**<br>(es 10, is_**t** 12, es 55, is_**t** 57, es 168, es 170, grause_**ts** 213)<br>{es, Personalendung 3.P.Sg.) | 7<br>2 |

| | | |
|---|---|---|
| 7 | sein [[1 Identität, 2 sich befinden, 3 sich verhalten, 4 Zustand]]<br>($ist_1$ 11, $ist_1$ 56, $sind_2$ 78, $sei_3$ 102, bist nich$h_3$ 188, $war_4$ 236)<br>{sein} | 6<br><br>1 |
| 8 | (mit 14, 49, 66, 74, 121, 228) | 6 |
| 9 | so [[$so_1$ Adverb, $so_2$ Konjunktion]]<br>($so_1$ 4,39,175, $so_2$ 192) | 4 |
| 10 | (in 22, 110, 221, 232) | 4 |
| 11 | **Wind**<br>(Wind 9, säusel_**t** 114, der Wind 115)<br>{Wind, Personalendung 3.P.Sg.} | 3<br><br>2 |
| 12 | (dem Arm 23, Armen 222, Armen 234) | 3 |
| 13 | (siehst nicht 43, siehst nicht 152, sehe 166) | 3 |
| 14 | (schöne 69, schön Adv. 130, schöne 185) | 3 |
| 15 | (reitet 3, reitet 215). | 2 |
| 16 | **haben** [[1 halten, 2 besitzen]]<br>($hat_1$ 18, $hat_2$ 84) | 2<br>1 |
| 17 | ($was_1$ 35, $was_2$ 96) | 2 |
| 18 | (geh 64, gehn 123) | 2 |
| 19 | **Blume**<br>(manch Blumen 76, sind_ 79)<br>{Blume, Personalendung 3.P.Pl} | 2<br><br>2 |
| 20 | (an 80, am 158) | 2 |
| 21 | **Mutter**<br>(Mutter 83, ha_**t** 85)<br>{Mutter, Personalendung 3.P.Sg} | 2<br><br>2 |
| 22 | (ruhig 104, 107) | 2 |
| 23 | **Weide**<br>(die Weiden 173, schein_**en** 172)<br>(Weide, Personalendung 3.P.P1} | 2<br>2 |
| 24 | **Gestalt**<br>(reiz_**t** 83, Gestalt 186)<br>{Gestalt, Personalendung 3.P.Sg.} | 2<br>2 |
| 25 | (spät 5) | 1 |
| 26 | (durch 6) | 1 |
| 27 | (Nacht 7) | 1 |
| 28 | (wohl 21) | 1 |
| 29 | (faßt 25) | 1 |

| | | |
|---|---|---|
| 30 | (sicher 28) | 1 |
| 31 | (warmhält 30) | 1 |
| 32 | (birgst 36) | 1 |
| 33 | (bang 40) | 1 |
| 34 | (Gesicht 42) | 1 |
| 35 | (Kron 50) | 1 |
| 36 | (Schweif 52) | 1 |
| 37 | (ein Nebelstreif 58) | 1 |
| 38 | (liebes 60) | 1 |
| 39 | (komm 62) | 1 |
| 40 | (gar 68) | 1 |
| 41 | (Spiele 70) | 1 |
| 42 | (spiel 71) | 1 |
| 43 | (bunte 77) | 1 |
| 44 | (dem Strand 81) | 1 |
| 45 | (manch Gewand 86) | 1 |
| 46 | (gülden 87) | 1 |
| 47 | (hörest nicht 93) | 1 |
| 48 | (leise 99) | 1 |
| 49 | (verspricht 100) | 1 |
| 50 | (bleibe 105) | 1 |
| 51 | (dürren 111) | 1 |
| 52 | (Blättern 112) | 1 |
| 53 | (säuselt 113) | 1 |
| 54 | (willst 116) | 1 |
| 55 | (feiner 118) | 1 |
| 56 | (sollen 126) | 1 |
| 57 | (warten 129) | 1 |
| 58 | (führen 133) | 1 |
| 59 | (den Reihn 135) | 1 |
| 60 | (nächtlichen 136) | 1 |

| | | |
|---|---|---|
| 61 | (wiegen 138) | 1 |
| 62 | (tanzen 141) | 1 |
| 63 | (einsingen 144) | 1 |
| 64 | (dort 155) | 1 |
| 65 | (düstern 159) | 1 |
| 66 | (Ort 160) | 1 |
| 67 | (genau 169) | 1 |
| 68 | (scheinen 171) | 1 |
| 69 | (alten 174), | 1 |
| 70 | (grau 176) | 1 |
| 71 | (liebe 178) | 1 |
| 72 | (reizt 182) | 1 |
| 73 | (willig 191) | 1 |
| 74 | (brauch 193) | 1 |
| 75 | (Gewalt 196) | 1 |
| 76 | (jetzt 201) | 1 |
| 77 | (anfaßt 202) | 1 |
| 78 | (hat getan 207) | 1 |
| 79 | (ein Leids 210) | 1 |
| 80 | (grausets 212) | 1 |
| 81 | (geschwind 217) | 1 |
| 82 | (hält 219) | 1 |
| 83 | (ächzende 224) | 1 |
| 84 | (erreicht 225) | 1 |
| 85 | (den Hof 227) | 1 |
| 86 | (Mühe 229) | 1 |
| 87 | (Not 231) | 1 |
| 88 | (tot 238) | 1 |

In der Tabelle haben wir die Polysemie nur angedeutet, das Wort nicht aber in separate Hrebs aufgeteilt. Es fehlt zunächst an Erfahrung, wie hier die Analyse durchgeführt werden sollte. Jedoch ermöglichen die auf die obige Weise dargestellten Daten alle weiteren Analysen.

Die hier illustrierte Art der Zerlegung des Texts in Hrebs bedeutet eine recht starke Reduktion der Daten. Von der Textlänge, die 225 Entitäten beträgt, zu den Formentypes, deren Anzahl 124 ist, sind wir nicht auf grammatischem, sondern auf semantischem Wege zu 88 denotativen Elementen gekommen. Schon diese Tatsache zeigt, dass ein Text nicht nur eine Sammlung von Wörtern, sondern eine Konstruktion ist, die man je nach den Kriterien auf unterschiedliche Konstrukte reduzieren kann. Wie immer bei solchen Gelegenheiten, verliert man etwas Information, gewinnt aber dafür eine andere. Offensichtlich ist diese Bevorzugung bestimmter Informationen nicht die einzig mögliche Reduktion, sie hängt immer vom Aspekt ab, von dem aus wir den Text untersuchen.

Hat man die Hrebs etabliert, so fragt man sich sofort, ob sie vielleicht irgendwelche Eigenschaften haben, die man operationalisieren und messen könnte, eventuell auch, ob man in diesen Eigenschaften korrespondierende „texteigene“ oder „intuitiv etablierte“ Eigenschaften sehen könnte. Denn das Ziel quantitativer textologischer Analysen ist es, intuitive Interpretationen entweder zu ersetzen oder sie zu objektivieren. Man kann zweifellos behaupten, dass man das, was man in einem Text qualitativ finden kann, auch quantitativ tun kann, aber umgekehrt geht es nicht immer. Schon die einfachste Eigenschaft, die Häufigkeitsverteilung, die bei jedem Massenphänomen vorhanden ist, lässt sich qualitativ auf keine Weise erfassen.

Im folgenden werden wir versuchen spezielle Klassen von Hrebs zu bestimmen, ihre Eigenschaften zu quantifizieren und ihre gegenseitigen Zusammenhänge zu berechnen. Die Kürze des untersuchten Textes erlaubt es uns, ausführliche Illustrationen zu bringen.

## 5.2. Verteilungen

Ebenso wie bei Wörtern, so kann man auch hier fragen, ob die Denotationshrebs den üblichen Verteilungen folgen. Wenn das nicht der Fall sein sollte, könnte man sie nicht als theoretisch fruchtbar betrachten. Die Unterordnung unter ein Verteilungsgesetzes ist ein Indiz für die Akzeptabilität einer Entität als Spracheinheit. Es lässt sich leicht zeigen, dass unsere Hrebs diesen Test hervorragend bestehen. In Tabelle 5.2 findet man die Anpassung der Zipf-Mandelbrot-Verteilung an die rangierten Häufigkeiten der Hrebs (aus Tabelle 5.1, letzte Spalte).

Tabelle 5.2
Anpassung der Zipf-Mandelbrot-Verteilung an die Ranghäufigkeiten der Hrebs im „Erlkönig“

| x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 40 | 38.09 | 31 | 1 | 1.56 | 61 | 1 | 0.74 |
| 2 | 37 | 23.40 | 32 | 1 | 1.51 | 62 | 1 | 0.73 |
| 3 | 21 | 16.68 | 33 | 1 | 1.46 | 63 | 1 | 0.72 |
| 4 | 9 | 12.87 | 34 | 1 | 1.41 | 64 | 1 | 0.70 |
| 5 | 8 | 10.43 | 35 | 1 | 1.37 | 65 | 1 | 0.69 |
| 6 | 7 | 8.73 | 36 | 1 | 1.32 | 66 | 1 | 0.68 |
| 7 | 6 | 7.50 | 37 | 1 | 1.29 | 67 | 1 | 0.67 |
| 8 | 6 | 6.55 | 38 | 1 | 1.25 | 68 | 1 | 0.66 |
| 9 | 4 | 5.81 | 39 | 1 | 1.21 | 69 | 1 | 0.65 |
| 10 | 4 | 5.21 | 40 | 1 | 1.18 | 70 | 1 | 0.64 |
| 11 | 3 | 4.72 | 41 | 1 | 1.15 | 71 | 1 | 0.63 |
| 12 | 3 | 4.31 | 42 | 1 | 1.12 | 72 | 1 | 0.62 |
| 13 | 3 | 3.96 | 43 | 1 | 1.09 | 73 | 1 | 0.61 |
| 14 | 3 | 3.67 | 44 | 1 | 1.06 | 74 | 1 | 0.58 |
| 15 | 2 | 3.41 | 45 | 1 | 1.04 | 75 | 1 | 0.59 |
| 16 | 2 | 3.18 | 46 | 1 | 1.01 | 76 | 1 | 0.60 |
| 17 | 2 | 2.98 | 47 | 1 | 0.99 | 77 | 1 | 0.57 |
| 18 | 2 | 2.81 | 48 | 1 | 0.97 | 78 | 1 | 0.56 |
| 19 | 2 | 2.65 | 49 | 1 | 0.94 | 79 | 1 | 0.55 |
| 20 | 2 | 2.51 | 50 | 1 | 0.92 | 80 | 1 | 0.55 |
| 21 | 2 | 2.38 | 51 | 1 | 0.9 | 81 | 1 | 0.54 |
| 22 | 2 | 2.26 | 52 | 1 | 0.88 | 82 | 1 | 0.53 |
| 23 | 2 | 2.16 | 53 | 1 | 0.87 | 83 | 1 | 0.53 |
| 24 | 2 | 2.06 | 54 | 1 | 0.85 | 84 | 1 | 0.52 |
| 25 | 1 | 1.97 | 55 | 1 | 0.83 | 85 | 1 | 0.51 |
| 26 | 1 | 1.89 | 56 | 1 | 0.81 | 86 | 1 | 0.51 |
| 27 | 1 | 1.81 | 57 | 1 | 0.8 | 87 | 1 | 0.50 |
| 28 | 1 | 1.74 | 58 | 1 | 0.78 | 88 | 1 | 0.49 |
| 29 | 1 | 1.68 | 59 | 1 | 0.77 | | | |
| 30 | 1 | 1.62 | 60 | 1 | 0.75 | | | |
| | a = 1.1199, b = 0.8346, n = 88, FG = 62, $X^Z$ = 27.26, P ~ 1.00 | | | | | | | |

Die Anpassung erfolgte mit zahlreichen Zusammenfassungen von Häufigkeitsklassen. Die Software beginnt am Ende der Folge, stellt fest, ob eine theoretische Klasse ein $NP_x > 1$ hat, und, wenn nicht, dann addiert sie sie zu der vorherigen

Klasse bzw. setzt mehrere Klassen zusammen, damit $NP_x > 1$ wird. Dadurch verliert man zwar mehrere Freiheitsgrade, aber in unserem Fall reichen die übriggebliebenen völlig aus, um eine hervorragende Anpassung zu liefern. Das Resultat sagt, dass die durch unser Verfahren gewonnenen Denotationshrebs berechtigte Texteinheiten sind.

Das Resultat ist in Abbildung 5.1 graphisch dargestellt.

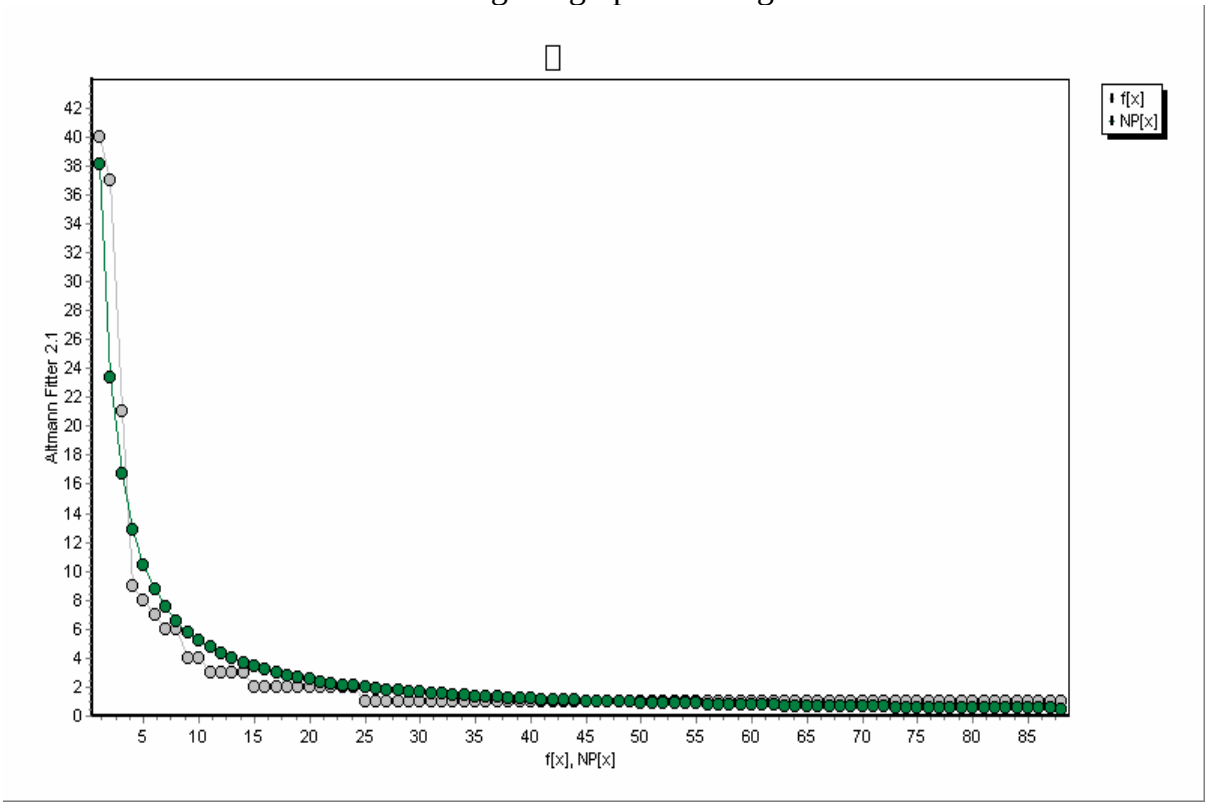


Abbildung 5.1. Rang-Häufigkeitsverteilung der Denotationshrebs im Erlkönig

Es ist bemerkenswert, dass in diesem Fall für die Anpassung sogar die einfache Zipf-Verteilung (= Zeta-Verteilung) ausgereicht hätte ($P$ = 0.52). Die starke Regularität gewährleistet auch eine zufriedenstellende Anpassung vieler anderer aus der Theorie herrührender Verteilungen.

Transformiert man diese Rangverteilung in eine Häufigkeitsverteilung, so müsste nach Zörnig und Boroda (1992) wieder die Zipf-Mandelbrot-Verteilung herauskommen. In unserem Fall lässt sich zwar zeigen, dass dies zutrifft, jedoch hat die Anpassung dann „nur“ $P$ = 0.13, unter sehr vielen anderen, die viel besser passen, zumal von G. Herdan für das Häufigkeitsspektrum eher die Waring-Verteilung propagiert wurde. Altmann (1993) zeigt mit einem analytischen „Trick“, der eine direkte Konsequenz der allgemeinen Theorie (vgl. Wimmer, Altmann 2005) ist, dass man aus der Zipf-Mandelbrot-Verteilung die Waring-Verteilung bekommt. Es handelt sich lediglich um den Übergang von einem stetigen zu einem diskreten Modell.

Betrachten wir die stetige Funktion

(5.1) $$y = \frac{K}{(b+x)^a}$$

die ein stetiges Analogon der Zipf-Mandelbrot-Verteilung ist. Wir stellen ihre Differentialgleichung auf, nämlich

(5.2) $$\frac{dy}{dx} = -aK(b+x)^{-a-1} = \frac{-ay}{b+x}.$$

Separiert man die Variablen, so erhält man

(5.3) $$\frac{dy}{y} = \frac{-a\,dx}{b+x}.$$

Diese Gleichung kann man diskretisieren, indem man für *dy* wie üblich *Δy* und für *dx Δx* schreibt. So erhält man

(5.4) $$\frac{\Delta y_x}{y_x} = \frac{-a}{b+x}\Delta x.$$

Da $\Delta y_x = y_{x+1} - y_x$ und $\Delta x = x+1-x = 1$, erhalten wir

(5.5) $$\frac{y_{x+1} - y_x}{y_x} = \frac{-a}{b+x}.$$

Setzen wir nun $y_x = P_x$ – da es um Wahrscheinlichkeiten geht – , so erhalten wir

(5.4) $$P_{x+1} = \left(1 - \frac{a}{b+x}\right)P_x = \left(\frac{b+x-a}{b+x}\right)P_x$$

oder

(5.5) $$P_x = \left(\frac{b-a+x-1}{b+x-1}\right)P_{x-1},$$

was bereits die Rekursionsformel der Waring-Verteilung bildet. Um auf die gleichen Parameter zu kommen, die in der Sofware benutzt werden, setzen wir a = *b´*+1 und *b* = *b´*+ *n* + 1 und schreiben

$$(5.6)\quad P_x = \left(\frac{n+x-1}{b'+n+x}\right)P_{x-1}$$

und erhalten so die Waring-Verteilung in der Form

$$(5.7)\quad P_x = \frac{b'}{b'+n}\frac{n^{(x)}}{(b'+n+1)^{(x)}},\ x=0,1,2,\ldots,$$

wo $b'$, $n > 0$ und $k^{(x)} = k(k+1)\ldots(k+x-1)$, $k^{(0)} = 1$ ist. Bei der Anpassung verschieben wir diese Verteilung um einen Schritt nach rechts und bekommen ohne Veränderung

$$(5.8)\quad P_x = \frac{b'}{b'+n}\frac{n^{(x-1)}}{(b'+n+1)^{(x-1)}},\ x=1,2,\ldots$$

Aus Tabelle 5.2 erhalten wir durch einfache Addition in der zweiten Spalte Tabelle 5.3. Die graphische Darstellung der Anpassung ist in Abbildung 5.2 zu sehen.

Tabelle 5.3
Anpassung der Waring-Verteilung an die Hreb-Häufigkeiten im „Erlkönig“

| x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ | x | $f_x$ | $NP_x$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 64 | 64.00 | 15 | 0 | 0.13 | 29 | 0 | 0.03 |
| 2 | 10 | 10.63 | 16 | 0 | 0.12 | 30 | 0 | 0.03 |
| 3 | 4 | 4.26 | 17 | 0 | 0.10 | 31 | 0 | 0.03 |
| 4 | 2 | 2.27 | 18 | 0 | 0.09 | 32 | 0 | 0.03 |
| 5 | 0 | 1.40 | 19 | 0 | 0.08 | 33 | 0 | 0.02 |
| 6 | 2 | 0.94 | 20 | 0 | 0.07 | 34 | 0 | 0.02 |
| 7 | 1 | 0.68 | 21 | 1 | 0.06 | 35 | 0 | 0.02 |
| 8 | 1 | 0.51 | 22 | 0 | 0.06 | 36 | 0 | 0.02 |
| 9 | 1 | 0.39 | 23 | 0 | 0.05 | 37 | 1 | 0.02 |
| 10 | 0 | 0.31 | 24 | 0 | 0.05 | 38 | 0 | 0.02 |
| 11 | 0 | 0.26 | 25 | 0 | 0.04 | 39 | 0 | 0.02 |
| 12 | 0 | 0.21 | 26 | 0 | 0.04 | 40 | 1 | 0.59 |
| 13 | 0 | 0.18 | 27 | 0 | 0.04 | | | |
| 14 | 0 | 0.15 | 28 | 0 | 0.03 | | | |
| | b´ = 1.1337, n = 0.4251, FG = 6, $X^2$ = 1.60, P = 0.95 | | | | | | | |

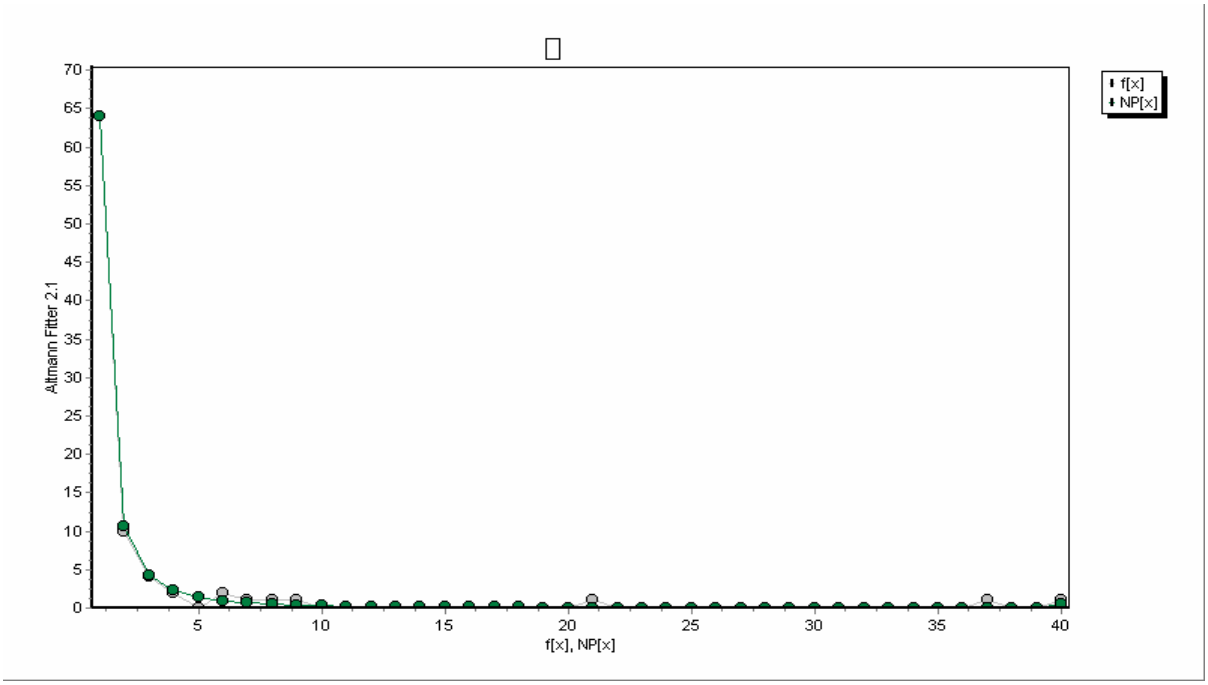


Abbildung 5.2. Anpassung der Waring-Verteilung an die Häufigkeiten der Denotationshrebs im Erlkönig

## 5.3. Die Suche nach dem Textkern

Den Kern eines Textes kann man auch intuitiv bestimmen, aber bekanntlich, eignet sich die Intuition nur zur Entdeckung, nicht aber zur Begründung oder Argumentation. Daher sollte sie durch eine objektivere Methode ersetzt werden.

Zum Kern des Textes gehören Entitäten, die selbst etwas tun oder die im Zentrum der Aufmerksamkeit stehen, um die sich sozusagen alles dreht. Durch eine solche Fokusierung wird die fragliche Entität öfters erwähnt, und zwar entweder direkt oder durch verschiedene Referenzen, Synonyma, Appellativa usw. Da diese aber zu unterschiedlichen Lexemen gehören, vergrößern sie dadurch den Mengen-Hreb der Entität, nicht aber unbedingt den Listen-Hreb.

Als Kern eines Textes bezeichnen wir daher alle Hrebs, deren Mengen-Hrebs mindestens zwei Elemente enthalten. Wenn es genau zwei Elemente gibt, dann muss das zweite ein selbständiges Lexem sein und kein Affix, z.B. keine Personalendung beim Verb oder Nomen (z.B. im Ungarischen). Möglicherweise wird sich diese Bedingung bei längeren Texten ändern, mit Sicherheit bei der Analysen von Texten in anderen Sprachen.

Betrachten wir Tabelle 1, so werden durch die obige Definition die Hrebs

Kind = {Kind, Knabe, Sohn, ich, mein, du, dein, er, 2PSg, 2PSgImp, 3PSg}
Vater = {Vater, wer, ich, mein, du, er sein, 2PSg, 3PSg}
Erlkönig = {Erlkönig, ich, mein, er, 2PSg, 3PSg}

erfasst, während die obige beschränkende Bedingung die Hrebs

Tochter = {Tochter, 3PPl}
es = {es, 3PSg}
Wind = {Wind, 3PSg}
Blume = {Blume, 3PSg}
Mutter = {Mutter, 3PSg}
Weide = {Weide, 3PSg}
Gestalt = {Gestalt, 3PSg}

als Kernhrebs disqualifiziert. Den Kern des Textes bildet also die Menge

Kern = {{Kind), {Vater}, {Erlkönig}}.

Dies hätte man sicherlich auch intuitiv erraten, weil es in diesem Text auf der Hand liegt. Bei langen Texten wird es aber sicherlich Probleme geben. Der Vorteil dieses Verfahrens liegt besonders in seiner Objektivität.

Die **Mächtigkeit des Kerns**, |Kern|, ergibt sich als die Summe der Mächtigkeiten aller Kern-Hrebs. Wir haben

|Kind| = |{Kind}| = 11
|Vater| = |{Vater}| = 9
|Erlkönig| = |{Erlkönig}| = 6,

daher ist |Kern| = 11 + 9 + 6 = 26. Die **relative Mächtigkeit des Kerns** ist das Verhältnis der Mächtigkeit des Kern zur Summe aller Mächtigkeiten. Zählt man diese in Tabelle 5.1 zusammen, so erhält man 128. Daher ist die relative Mächtigkeit des Kerns gleich

$|\text{Kern}|_{\text{rel}} = 26/128 = 0.20.$

Weiter definieren wir die **Topikalität eines Kern-Hrebs** als dessen Proportion im Kern, d.h.

(5.9) $$T(Hreb) = \frac{|\{Hreb\}|}{|Kern|},$$

was für die obigen Kern-Hrebs

$$T(Kind) = \frac{|\{Kind\}|}{|Kern|} = \frac{11}{26} = 0.42$$

T(Vater) = 0.35
T(Erlkönig) = 0.23

bedeutet. Interessant ist die Tatsache, dass, obwohl hier eine andere Etablierungstechnik für Hrebs angewandt wurde als in Ziegler, Altmann (2002), der Kern derselbe ist und sich die Topikalitäten nur minimal unterscheiden. Daher lässt sich schließen, dass Mengen-Hrebs gute Indikatoren für den Kern und für die Topikalität dessen Elemente sind und die Methode ihrer Bestimmung recht zuverlässig ist.

Sowohl die relative Mächtigkeit des Textkerns als auch die Topikalität des Hrebs sind ganz einfache Proportionen und können mit allen dazu geeigneten statistischen Mitteln als solche behandelt werden. So kann die Entropie im Kern ermittelt werden, Erwartung, Varianz usw. und Textkerne können miteinander statistisch verglichen werden. Das eröffnet wiederum weitere Forschungsmöglichkeiten wie z.B. die Entwicklung eines Autors, den Stand einer Textsorte, die Gestaltung der Kerne in unterschiedlichen Sprachen u.a.

Eine andere Schicht des Textes wird durch den deterministischen Kern dargestellt. Dieser wird durch alle Hrebs gebildet, deren Elemente deterministisch, z.B. grammatisch, verbunden sind (vgl. Ziegler, Altmann (2003); Wimmer et al. 2003). Als deterministisch können wir eine solche Verbindung erachten, die bei der Verbindung eines Regens-Wortes bei dem Rectum-Wort nicht fehlen darf. So ist z.B. bei „Vater singt“ „Vater“ Regens, und „singen“ ist Rectum. Statt „singen“ kann man andere Verben einsetzen, aber die Endung der dritten Person Singular muss immer vorhanden sein.

Kern-Hrebs können, müssen aber nicht unbedingt zum deterministischen Kern gehören, wie an einem englischen Sonett gezeigt wurde (Ziegler, Altmann 2003). So finden wir z.B. im „Erlkönig“ (Tabelle 5.1), dass „Kind“ mit folgenden Verben auf diese deterministische Weise verbunden ist (d.h. dass es im Hreb „Kind“ folgende grammatisch verbundenen Wörter gibt): {bergen, kommen, gehen, sein, bleiben, wollen}. Diese Verben gehören nicht zum Hreb „Kind“, aber ihre Endung wird vom „Kind“ regiert.

Stellt man alle solche Verbindungen über alle Hrebs des Textes zusammen, dann bekommt man den deterministischen Kern (DK) als

$DK_{Erlkönig}$ = {Kind, bergen, kommen, gehen, sein, bleiben, wollen, Vater, reiten, haben, fassen, warmhalten, halten, sehen, hören, erreichen, Erlkönig, spielen, versprechen, lieben, brauchen, anfassen, tun, Tochter, sollen, führen, wiegen, tanzen, einsingen, es, grausen, Wind, säuseln, Blume, Mutter, Weide, scheinen, Gestalt, reizen}

oder nach Nomina und Verben geordnet

{Kind, Vater, Erlkönig, Tochter, Wind, Blume, Mutter, Weide, Gestalt, es; bergen, kommen, gehen, sein, bleiben, wollen, reiten, haben, fassen, warmhalten,

sehen, hören, halten, erreichen, spielen, versprechen, lieben, brauchen, anfassen, tun, sollen, führen, wiegen, tanzen, einsingen, grausen, säuseln, scheinen, reizen}.

Man kann aus dem deterministischen Kern „es" auslassen, z.B. in „es regnet", sonst kommt es sowieso nicht als Hreb vor, weil es nur ein Nomen vertritt. In unserem Fall schließt der deterministische Kern den Textkern ein. Auch hier kann man Maße einführen, z.B. die relative Größe des deterministischen Kerns als Proportion seiner Mächtigkeit zu allen Wörtern des Textes, d.h.

$$(5.10)\quad DK_{rel} = \frac{|DK|}{L},$$

wo $L$ die Textlänge in Anzahl der Wörter bedeutet. Für unseren Fall mit $L = 225$ erhalten wir

$$DK_{Erlkönig} = 38/225 = 0.17.$$

## 5.4. Kompaktheit, Zentralisiertheit, Diffusität

Der Text ist auf der denotativen Seite umso kompakter, je weniger Hrebs es in ihm gibt. Ornamentale Texte, z.B. poetische, enthalten viele Hrebs, wissenschaftliche relativ weniger, weil sich in diesen der Inhalt stark konzentriert. Die Zahl der Hrebs (n) hängt natürlich auch von der Textlänge (L) ab, wobei L in diesem Fall die Zahl der Positionen ist, und die Textlänge hängt wiederum mit der Morphologie der Sprache zusammen. Bei stark synthetischen Sprachen wird ein nach Positionen analysierter Text viel länger sein als die Zahl der Wörter in ihm; in stark analytischen Sprachen werden sich die beiden Längen einander annähern. Da bisher nur kurze Texte verarbeitet worden sind, gibt es keine Vergleichsmöglichkeiten. Daher benutzen wir hier das Maß der Kompaktheit so, wie es von Ziegler und Altmann (2002) vorgeschlagen worden ist:

$$(5.10)\quad K = \frac{1-\frac{n}{L}}{1-\frac{1}{L}}.$$

Hier symbolisiert n die Zahl der Hrebs und $L$ die Zahl der Positionen im Text (nicht die der Wörter!).

Da n minimal 1 und maximal $L$ sein kann, bewegt sich der Index im Intervall <0,1>. Je kompakter der Text, desto größer der Wert von $K$. Man erkennt im

Zähler sofort eine der vielen (weniger adäquaten) Varianten des Maßes für Vokabularreichtum. Der Index wird aus diesem Grunde sicherlich einmal modifiziert werden, im Augenblick bietet er aber eine gute Vergleichsmöglichkeit. Für den „Erlkönig“ mit $n$ = 88 Hrebs in $L$ = 238 Positionen bekommen wir

$$K = \frac{1-88/238}{1-1/238} = 0.63.$$

Man kann die Kompaktheit zweier Texte mit Hilfe der Formel (5.10) nur dann vergleichen, wenn man diese aufgrund der gleichen Zerlegungskriterien in Hrebs analysiert hat. Die strengeren Kriterien, die wir hier benutzt haben, haben die Zahl der Positionen vermehrt und die der Hrebs verringert, so dass wir hier eine etwas größere Kompaktheit bekommen haben als in Ziegler, Altmann (2002), wo sich R = 0.58 ergab.

Während es sich bei der Kompaktheit um eine bloße Reduktion der Zahl der Hrebs (Denotationsverengung) handelt, wobei alle Hrebs die gleiche Häufigkeit haben können, misst die Wiederholungsrate $R$, die bereits in Kapitel 2.3 vorgestellt worden ist, die **Zentralisiertheit des Textes**, d.h. die Fokussierung der Aufmerksamkeit auf einige der vorhandenen Hrebs. Würden alle Hrebs eines Textes mit gleicher Häufigkeit vorkommen, so wäre die Konzentration minimal; würde aber ein Hreb sehr häufig vorkommen und alle anderen nur jeweils einmal, so wäre die Konzentration auf diesen Hreb sehr stark.

Die Berechnung der üblichen Wiederholungsrate lässt sich aufgrund von Tabelle 5.1 oder von Tabelle 5.2 leicht durchführen:

$$R = (40^2 + 37^2 + 21^2 + ... + 1^2 + 1^2)/238^2 = 0.0679.$$

Als relatives Maß empfiehlt sich wieder McIntosh`s (1967) Index in der Form

$$(5.11)\quad R_{rel} = \frac{1-\sqrt{R}}{1-\frac{1}{\sqrt{n}}},$$

das hier, nach Einsetzung von $\sqrt{0.0674} = 0.2596$ und $n$ = 88, $R_{rel}$ = 0.83 ergibt. Auf ähnliche Weise kann man auch die Entropie oder andere Maße als Charakteristika der Zentralisiertheit benutzen. Andere Kriterien der Hrebbildung würden natürlich andere Werte ergeben. Man kann sich jedoch leicht davon überzeugen, dass der berechnete Wert von $R$ recht weit von dem theoretischen Wert entfernt liegt, der aus der Zipf-Mandelbrot-Verteilung folgt. Entweder gilt hier nicht die Beziehung zwischen Wiederholungsrate und Inventargröße, wie wir sie aus dem phonischen Bereich kennen, oder es gibt ein anderes Problem, das untersucht werden muss. Falls die Beziehung zwischen Inventargröße und Wiederholungsrate korrekt ist, dann

kann das Problem auch in der Aufstellung der Kriterien liegen. Die Vergleichbarkeit der Werte dieses Indexes ist aber trotzdem möglich, s. Kap. 4.1.

Die **Diffusität** eines Hrebs misst dessen Streuung im Text. Es bieten sich sehr viele Charakterisierungsmöglichkeiten an, die man aus der Statistik übernehmen kann, hier beschränken wir uns auf drei Maße. Beim ersten ziehen wir nur die Position des ersten und des letzten Vorkommens eines Hrebs im Text in Betracht und definieren

$$(5.12)\quad D(Hreb) = \frac{\sup < Hreb > - \inf < Hreb >}{|Hreb|},$$

d.h., wir subtrahieren die erste Position von der letzten und dividieren das Resultat durch die Gesamtzahl der Positionen (Umfang des Hrebs). In Tabelle 5.1 sieht man beim ersten Hreb (Kind) den Positionshreb in der letzten Zeile, daher sup<Kind> = 237 (im Hilfsverb war), inf<Kind> = 16 und |Kind| = 40, woraus sich

$$D(Kind) = \frac{237 - 16}{40} = 5.53$$

ergibt. Der Hreb (Vater) ergibt *D*(Vater) = (233 - 1)/36 = 6.44. Die Diffusität ist ein Maß der positionalen Konzentration eines Hrebs, das Maß seiner Dichte. Dieser Index lässt sich auf verschiedene Weisen normieren. Er ist nur dann sinnvoll anwendbar, wenn der fragliche Hreb mindestens zweimal vorkommt, d.h. $|Hreb| \geq 2$.

Für die einzelnen Hrebs im „Erlkönig“, die wir als Diffusitäthrebs (DH) bezeichnen, ergeben sich die Diffusitäten als

| | | | |
|---|---|---|---|
| Kind | 5.53 | sehen | 41.00 |
| Vater | 6.27 | schön | 38.67 |
| Erlkönig | 7.67 | reiten | 106.00 |
| und | 24.67 | haben | 33.00 |
| Tochter | 4.00 | was | 30.50 |
| es | 27.00 | gehen | 29.50 |
| sein | 37.50 | Blume | 1.50 |
| mit | 35.67 | an | 39.00 |
| so | 47.00 | Mutter | 1.00 |
| in | 52.50 | ruhig | 1.50 |
| Wind | 35.33 | Gestalt | 1.50 |
| Arm | 70.33 | Weide | 0.50 |

Wie man sieht, haben die Hrebs des Kerns kleine Diffusitäten. Diejenigen Hrebs, deren Diffusität kleiner ist als die größte Diffusität im Kern, betrachten wir als zum **erweiterten Kern** (*EK*) gehörend. Hier gehören dazu 9 Hrebs:

$\text{Kern}_{\text{erweitert}}$ = {Kind, Vater, Erlkönig, Tochter, Mutter, Blume, Gestalt, Weide, ruhig}.

Aus den Einzeldiffusitäten berechnen wir zwei weitere Charakteristika: Die **durchschnittliche Diffusität des erweiterten Kerns**

$$\bar{D}_{EK} = \frac{1}{|EK|} \sum_{i \in EK} D_i \,, \tag{5.13}$$

die sich hier als

$$\bar{D}_{EK} = \frac{1}{9}(5.53 + 6.27 + 7.67 + 4.0 + 1.5 + 1.0 + 1.5 + 1.5 + 1.5 + 0.5) = 3.44$$

ergibt, und die **durchschnittliche Diffusität des Texts**, die sich aus allen Einzeldiffusitäten derjenigen Hrebs, die mindestens zweimalmal vorkommen, zusammensetzt, d.h.

$$\bar{D}_{Text} = \frac{1}{|DH|} \sum_{i \in DH} D_i \,. \tag{5.14}$$

Für den „Erlkönig“ erhalten wir durch Summierung aller Diffusitäten der 24 oben angeführten Diffusitätshrebs

$$\bar{D}_{Erlkönig} = \frac{1}{24}(5.53 + 6.64 + ... + 1.5 + 0.5) = 28.22 \,.$$

Dieser Index wird wahrscheinlich noch weiter relativiert werden müssen, weil er möglicherweise von der Textlänge abhängt.

Die bisher definierten Arten von Kernen stehen in einer bestimmten Beziehung zueinander: Der Textkern ist offensichtlich eine Teilmenge des deterministischen Kerns, der neben dem Textkern noch einige weitere Nomina enthält, der Rest sind die Tätigkeiten dramatis personarum. Der deterministische Kern ist also eine Art Erweiterung des Textkerns, ist aber nicht identisch mit dem hier definierten erweiterten Kern, mit dem er nur einen Mengendurchschnitt hat. Wenn später im Laufe der Forschung andere Erweiterungen des Textkerns definiert werden, so wird ihre gegenseitige Beziehung eine spezielle Eigenschaft des Textes darstellen.

## 5.5. Rhematische Schichtung des Textes

Nach der automatischen Etablierung des Textkerns, des erweiterten Textkerns und des deterministischen Textkerns ist natürlich zu fragen, ob der gesamte Text irgendwie geschichtet ist. Textanalytiker sprechen auch von der thematischen Progression des Textes (Daneš 1970). Eine thematische Schicht (das, was schon bekannt ist) wird rhematisch erweitert, diese neue Schicht wird wieder um eine neue Schicht erweitert usw. Es besteht immer die Möglichkeit, eine solche Schichtung interpretativ zu ermitteln, uns geht es aber darum, sie automatisch ermitteln zu können. Falls es überhaupt eine Schichtung gibt, dann muss diese einer Regularität folgen, die theoretisch ableitbar und empirisch überprüfbar sein müsste. In der Tat lässt sich eine derartige Schichtung mit Hilfe des Menzerathschen Gesetzes erkennen. Nach diesem Gesetz (vgl. Altmann 1980; Altmann, Schwibbe 1989) gilt allgemein, dass die Größe der Konstituenten einer sprachlichen Entität um so kleiner wird, je größer das ganze Konstrukt ist. An unzähligen Fällen wurde gezeigt, dass das Gesetz für alle formalen Ebenen der Sprache und des Textes gilt (vgl. Hřebíček 1995, 1997, 2000). Jedoch handelt es sich in unserem Fall nicht um eine formale, sondern um eine semantisch-denotative Einheit, die aus formal disparaten Wortformen oder Wortteilen bestehen kann. Gerade diese Randbedingung führt dazu, dass sich das gesamte denotative Feld in Schichten aufteilt, in denen es unterschiedliche formale Homogenitäten gibt. Das Menzerathsche Gesetz trennt diese Schichten mechanisch voneinander.

Das Menzerathsche Gesetz hat die Form

(5.15) $y = Ax^b$,

wobei *X* die Konstruktgröße, hier den Listen-Hreb-Umfang, und *Y* die durchschnittliche silbische Länge der Konstituenten (Elementen) des Hrebs bedeuten. A und *b* sind Parameter. Bei der Messung der Längen gelten folgende Regeln:

(i) Es wird immer die Silbenzahl in einem ganzen Hreb-Element ermittelt, z.B. „Kind“ enthält 1 Silbe, „das Kind“ 2 Silben.

(ii) Ein Affix, das ein Hrebelement darstellt, hat die Länge 0, falls es keinen Vokal enthält, sonst die Länge 1, 2, ... (falls es 1, 2, ... Vokale enthält). So z.B. ist „t“ in „erreicht_t“ im (Vater)-Hreb 0 Silben lang, im (erreichen)-Hreb hingegen 2 Silben lang.

(iii) Es werden jeweils alle Hrebs des gleichen Umfangs gemittelt.

So ergeben sich z.B. für $x = 3$ (d.h. alle Hrebs des Umfangs 3 in Tabelle 5.1) folgende Elementlängen

1,0,2,2,2,2,2,2,2,2,1,2

(in den Hrebs Nr. 11, 12, 13, 14). Die Anzahl der Elemente ist $n = 12$, die Summe der Längen $\Sigma\ l_i = 1 + 0 + 2 +... + 1 + 2 = 20$, woraus sich $y = 20/12 = 1.67$ ergibt.

Berechnet man auf diese Weise alle Durchschnitte, so bekommt man die Resultate in Tabelle 5.4. Es ist zu betonen, dass wir die Hrebs aus Wortteilen zusammengesetzt haben, die eine bestimmte silbische Länge haben.

Tabelle 5.4
Durschnittliche silbische Länge
des Hrebs mit dem Umfang *x*

| x | y |
|---|---|
| 1 | 1.78 |
| 2 | 1.35 |
| 3 | 1.67 |
| 4 | 1.00 |
| 6 | 1.08 |
| 7 | 0.57 |
| 8 | 1.38 |
| 9 | 1.00 |
| 21 | 1.52 |
| 37 | 1.11 |
| 40 | 0.98 |

Die rhematischen Schichten entstehen aus einer (theoretisch) monoton fallenden Folge von Werten. In Tabelle 5.4 sieht man, dass es hier vier Schichten gibt (1-2, 3-7, 8-9, 22-40), wobei die letzte genau dem Textkern entspricht, d.h., es befinden sich in dieser Schicht genau die Kern-Hrebs, während die restlichen Elemente des erweiterten Kerns in 8-9 (Tochter) und 1-2 (Mutter, Blume, Gestalt, Weide, ruhig) verteilt sind. Da bisher nur wenige Texte untersucht worden sind, ist nicht bekannt, wie sich die Elemente des erweiterten Kerns im Lichte des Menzerathschen Gesetzes verhalten.

Im allgemeinen reicht eine optische Trennung der Schichten hin, wie sie in Tabelle 5.4 angedeutet wird, aber in längeren Texten wird es wahrscheinlich zu Entscheidungsproblemen kommen, wenn die empirischen Daten innerhalb einer Schicht nicht ganz monoton abnehmen. Daher berechnen wir die Kurve (5.15) für diejenigen Schichten, die mindestens drei unterschiedliche *x*-Werte haben. Die Parameter *A* und *b* ergeben sich aus

$$A = e^a$$

wobei

$$(5.16)\quad a = \frac{\sum_i (\ln x_i)^2 \ln y_i - \sum_i \ln x_i \sum_i \ln x_i (\ln y_i)}{n\sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2}$$

und

$$(5.17)\quad b = \frac{n\sum_i \ln x_i (\ln y_i) - \sum_i \ln x_i \sum_i \ln y_i}{n\sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2}.$$

So ergibt sich für die zweite Schicht

$y = 4.61x^{-0.9775}$ mit dem Determinationskoeffizienten $D = 0.78$ (S. Anhang I)

wo $D$ etwas klein ist, weil $x = 6$ eine Anomalität aufweist. Für die vierte Schicht mit den Kern-Hrebs erhalten wir

$y = 10.53x^{-0.6341}$ mit dem Determinationskoeffizienten $D = 0.98$,

was ein sehr gutes Resultat darstellt. In Tabelle 5.5 sind alle berechneten Werte in der dritten Spalte angegeben

Tabelle 5.5
Beobachtete und theoretische Werte
der rhematischen Schichtung im „Erlkönig“

| x | y | $y_{teor}$ |
|---|---|---|
| 1 | 1.78 | |
| 2 | 1.35 | |
| 3 | 1.67 | 1.57 |
| 4 | 1.00 | 1.19 |
| 6 | 1.08 | 0.80 |
| 7 | 0.57 | 0.69 |
| 8 | 1.38 | |
| 9 | 1.00 | |
| 21 | 1.52 | 1.53 |
| 37 | 1.11 | 1.07 |
| 40 | 0.98 | 1.02 |

Erst nach der Analyse vieler Texte wird es möglich sein zu sagen, ob sich die rhematischen Schichten durch bestimmte Größen der Parameter voneinander signifikant unterscheiden lassen oder nicht.

## 5.6. Informationsfluss

Ebenso wie mit Wörtern, so lässt sich der Informationsfluss in Form des Type-TokenVerhältnisses auch mit Hrebs messen. Zu diesem Zweck schreibt man den Text als Folge von Hrebs, die man mit Zahlen symbolisiert. Es ist am besten, wenn man den Text satz- oder versweise in unterschiedlichen Zeilen schreibt, weil man eine solche Datei noch für andere Zwecke benutzen kann. Bei der Transkription benutzen wir den positionsgemäß dargestellten Text von Abschnitt 5.1 und die Numerierung der Hrebs in Tabelle 5.1, so dass wir für den „Erlkönig“ folgendes Schema bekommen:

Schema (I)

2, 15, 2, 9, 25, 26, 27, 4, 11
6, 7, 6, 2, 8, 2, 1
2, 16, 2, 1, 28, 10, 12
2, 29, 2, 1, 30, 2, 31, 2, 1

2, 1, 17, 32, 1, 1, 9, 33, 1, 34
13, 2, 2, 2, 3
3, 8, 35, 4,36
2, 1, 6, 7, 6, 37

1, 38, 1, 39, 1, 18, 1, 8, 3
40, 14, 41, 42, 3, 3, 8, 1
19, 43, 7, 19, 20, 44
3, 21, 16, 21, 45, 46

1, 2, 1, 2, 4, 47, 2, 2
17, 3, 1, 48, 49, 3
7, 1, 22, 50, 1, 22, 2, 1
10, 51, 52, 53, 11, 11

54, 1, 55, 1, 1, 8, 3, 18
3, 5, 56, 5, 1, 57, 14
3, 5, 58, 5, 59, 60
4, 61, 5, 4, 62, 5, 4, 63, 5, 1

1, 2, 1,2, 4, 13, 2, 2, 64
3, 5, 20, 65, 66
2, 1, 2, 1, 2, 13, 2, 6, 67
6, 68, 23, 23, 69, 9, 70

3, 71, 3, 1, 3, 72, 24, l, 14, 24
4, 7, 1, 1, 73, 9, 74, 3, 3, 75
1, 2, 1, 2, 76, 77, 3, 3, 1
3, 78, 3, 1, 79

2, 80, 6, 2, 15, 2, 81
2, 82, 2, 10, 12, 1, 83
84, 2, 85, 8, 86, 4, 87
10, 2, 12, 1, 7, 1, 88.

Die TTR-Folge ergibt sich daraus so, wie in Tabelle 5.6 angegeben.

Tabelle 5.6
Die TTR-Folge für die Hrebs im „Erlkönig“

| $L_i$ | $V_i$ | $L_i$ | $V_i$ | $L_i$ | $V_i$ | $L_i$ | $V_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 61 | 29 | 121 | 52 | 181 | 70 |
| 2 | 2 | 62 | 30 | 122 | 52 | 182 | 71 |
| 3 | 2 | 63 | 30 | 123 | 52 | 183 | 72 |
| 4 | 3 | 64 | 31 | 124 | 52 | 184 | 72 |
| 5 | 4 | 65 | 31 | 125 | 53 | 185 | 72 |
| 6 | 5 | 66 | 31 | 126 | 54 | 186 | 72 |
| 7 | 6 | 67 | 31 | 127 | 54 | 187 | 72 |
| 8 | 7 | 68 | 32 | 128 | 54 | 188 | 72 |
| 9 | 8 | 69 | 33 | 129 | 55 | 189 | 72 |
| 10 | 9 | 70 | 34 | 130 | 55 | 190 | 72 |
| 11 | 10 | 71 | 35 | 131 | 55 | 191 | 73 |
| 12 | 10 | 72 | 35 | 132 | 55 | 192 | 73 |
| 13 | 10 | 73 | 35 | 133 | 56 | 193 | 74 |
| 14 | 11 | 74 | 35 | 134 | 56 | 194 | 74 |
| 15 | 11 | 75 | 35 | 135 | 57 | 195 | 74 |
| 16 | 12 | 76 | 36 | 136 | 58 | 196 | 75 |
| 17 | 12 | 77 | 37 | 137 | 58 | 197 | 75 |
| 18 | 13 | 78 | 37 | 138 | 59 | 198 | 75 |
| 19 | 13 | 79 | 37 | 139 | 59 | 199 | 75 |
| 20 | 13 | 80 | 38 | 140 | 59 | 200 | 75 |
| 21 | 14 | 81 | 39 | 141 | 60 | 201 | 76 |
| 22 | 15 | 82 | 39 | 142 | 60 | 202 | 77 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 23 | 16 | 83 | 40 | 143 | 60 | 203 | 77 |
| 24 | 16 | 84 | 40 | 144 | 61 | 204 | 77 |
| 25 | 17 | 85 | 40 | 145 | 61 | 205 | 77 |
| 26 | 17 | 86 | 41 | 146 | 61 | 206 | 77 |
| 27 | 17 | 87 | 42 | 147 | 61 | 207 | 78 |
| 28 | 18 | 88 | 42 | 148 | 61 | 208 | 78 |
| 29 | 18 | 89 | 42 | 149 | 61 | 209 | 78 |
| 30 | 19 | 90 | 42 | 150 | 61 | 210 | 79 |
| 31 | 19 | 91 | 42 | 151 | 61 | 211 | 79 |
| 32 | 19 | 92 | 42 | 152 | 61 | 212 | 80 |
| 33 | 19 | 93 | 43 | 153 | 61 | 213 | 80 |
| 34 | 19 | 94 | 43 | 154 | 61 | 214 | 80 |
| 35 | 20 | 95 | 43 | 155 | 62 | 215 | 80 |
| 36 | 21 | 96 | 43 | 156 | 62 | 216 | 80 |
| 37 | 21 | 97 | 43 | 157 | 62 | 217 | 81 |
| 38 | 21 | 98 | 43 | 158 | 62 | 218 | 81 |
| 39 | 21 | 99 | 44 | 159 | 63 | 219 | 82 |
| 40 | 22 | 100 | 45 | 160 | 64 | 220 | 82 |
| 41 | 22 | 101 | 45 | 161 | 64 | 221 | 82 |
| 42 | 23 | 102 | 45 | 162 | 64 | 222 | 82 |
| 43 | 24 | 103 | 45 | 163 | 64 | 223 | 82 |
| 44 | 24 | 104 | 46 | 164 | 64 | 224 | 83 |
| 45 | 24 | 105 | 47 | 165 | 64 | 225 | 84 |
| 46 | 24 | 106 | 47 | 166 | 64 | 226 | 84 |
| 47 | 25 | 107 | 47 | 167 | 64 | 227 | 85 |
| 48 | 25 | 108 | 47 | 168 | 64 | 228 | 85 |
| 49 | 25 | 109 | 47 | 169 | 65 | 229 | 86 |
| 50 | 26 | 110 | 47 | 170 | 65 | 230 | 86 |
| 51 | 26 | 111 | 48 | 171 | 66 | 231 | 87 |
| 52 | 27 | 112 | 49 | 172 | 67 | 232 | 87 |
| 53 | 27 | 113 | 50 | 173 | 67 | 233 | 87 |
| 54 | 27 | 114 | 50 | 174 | 68 | 234 | 87 |
| 55 | 27 | 115 | 50 | 175 | 68 | 235 | 87 |
| 56 | 27 | 116 | 51 | 176 | 69 | 236 | 87 |
| 57 | 27 | 117 | 51 | 177 | 69 | 237 | 87 |
| 58 | 28 | 118 | 52 | 178 | 70 | 238 | 88 |
| 59 | 28 | 119 | 52 | 179 | 70 | | |
| 60 | 29 | 120 | 52 | 180 | 70 | | |

Berechnet man nun das Type-Token-Verhältnis für diese Daten, so erhält man die Kurven aus Abschnitt 4.3 wie folgt

Herdan $V_i = L_i^{0.8193}$ $D = 0.9964$

Tuldava 1 $V_i = \frac{237.6324}{425.4590 + L_i}$ $D$ = 0.9957

Tuldava 2 $V_i = \frac{110.0397}{110.0397 - 1 + L_i}$ $D$ = 0.9228

Köhler-Martináková $V_i = \frac{0.5572 L_i}{1 - 0.002345 + 0.002345 L_i}$ $D$ = 0.9957

Brunet 1 $V_i = 0.1959(\ln\ L_i)^{3.5737}$ $D$ = 0.9938

Brunet 2 $V_i = 1 + 0.1656(\ln\ L_i)^{3.6653}$ $D$ = 0.9949

Maas $V_i = L_i(\ln\ L_i)^{-0.5358}$ $D$ = 0.9745

Dugast $V_i = L_i^{1-0.03623 \ln\ L_i}$ $D$ = 0.9851

Orlov $V_i = \frac{7000(\ln 7000 - \ln\ L_i)L_i}{(\ln\ 7000 + 0.7667)(7000 - L_i)}$ $D$ = 0.9982

Somers $V_i = \exp(\ln^{0.8781} L_i)$ $D$ = 0.9971

Sichel

$$V_i = \frac{2}{(-0.1230406)0.2083475}\left[1 - \exp(-0.1230406(1 - \sqrt{1 + 0.2083475 L_i})\right]$$

$D$ = 0.9986

Für diese Kurven gelten alle Kommentare aus Abschnitt 4.3. Wie man sieht, zeigt der Hreb-TTR-Verlauf ein reguläres Verhalten und erweist sich als fruchtbare textuelle Einheit. Wie oben angemerkt, fangen einige der theoretischen Kurven nicht bei $V_1 = 1$ an. Es ist nicht nötig zu entscheiden, welche von den Kurven die „beste“ ist, da die Anpassungsunterschiede sehr gering sind. Vorzuziehen ist vorläufig die Herdansche Kurve, und zwar nicht nur wegen ihrer Einfachheit und wegen ihrer Anpassungsgüte, sondern auch wegen ihrer Analogie zur selbstorganisierten Kritikalität, die in evolvierenden Systemen herrscht (vgl. Bak 1999) und dadurch der Textlinguistik helfen könnte, sich in eine allgemeinere Disziplin einzufügen. Die breite Geltung der Potenzkurve ist aus vielen Disziplinen

bekannt (vgl. z.B. Schroeder 1991; Glottometrics 3-5), aus der Linguistik besonders unter dem Namen Menzerathsches Gesetz besonders (vgl. Altmann, Schwibbe 1989).

## 5.7. Koinzidenz

Von den zahlreichen Koinzidenzarten werden wir hier nur die positionale Koinzidenz in Betracht ziehen. Zwei Hrebs betrachten wir als koinzident,

(a) wenn sie in einem Wort (in einer wortähnlichen Einheit) untrennbar verbunden sind, z.B. Verb + Personalendung als Subjekt oder Verb + Objektendung + Personalendung, wie z.B. im ungarischen „látlak“ *ich sehe dich*, wo sogar drei Hrebs verschmolzen sind, oder in einem Kompositum, das zerlegt worden ist. Unterschiedliche Sprachen nutzen hier sehr unterschiedliche Möglichkeiten. In Fällen der untrennbaren Verschmelzung sprechen wir von deterministischer Koinzidenz.

(b) Wenn sich zeigen lässt, dass die Wahrscheinlichkeit des gemeinsamen Vorkommens zweier Hrebs im gleichen Rahmen (z.B. im Satz oder im Vers) die vorgegebene α-Ebene unterschreitet, dann sprechen wir von stochastischer Koinzidenz. Die α-Ebene muss festgelegt werden, sie ist konventionell. Es wird sich wahrscheinlich irgendwann herausstellen, dass sie der Textgrösse angepasst werden muss, vorläufig können wir $\alpha = 0.05$ benutzen. Diese positionelle Koinzidenz wird mit Hilfe des Schemas I ausgewertet, in dem die Hrebs, als Zahlen kodiert, zeilenweise angegeben sind.

Die deterministische Koinzidenz lässt sich durch grammatische und durch semantische Analyse festlegen. So besteht „reitet“ in Position 2/3 aus dem Hreb [reiten] und dem Hreb [Vater], der durch die Endung „-t“ repräsentiert wird. Daher sind diese beiden Hrebs deterministisch verbunden.

Bei der stochastischen Koinzidenz müssen wir die entsprechende Wahrscheinlichkeit dafür finden, dass zwei Hrebs im gegebenen Rahmen gemeinsam so oft wie beobachtet oder noch häufiger vorkommen. Dazu sind folgende Größen notwendig:

- $N$ − Zahl der Rahmen (Sätze, Verse, Passagen vorbestimmter Länge usw.) im Text
- $M$ − Zahl der Rahmen, in denen Hreb A vorkommt (mehrfaches Vorkommen im Vers/Satz gilt als ein Vorkommen)
- $n$ − Zahl der Rahmen, in denen Hreb B vorkommt (mehrfaches Vorkommen im Vers/Satz gilt als ein Vorkommen)
- $x$ − Zahl der Rahmen, in denen A und B gemeinsam vorkommen.

Die Wahrscheinlichkeit, dass A und B unter diesen Bedingungen genau $x$ mal gemeinsam im gleichen Rahmen (Vers, Satz) vorkommen, berechnet sich aus folgender Überlegung: Die Zahl aller Möglichkeiten, $M$ A-s und $n$ B-s in $N$

Versen zu platzieren, beläuft sich auf $\binom{N}{M}\binom{N}{n}$. Die Zahl der günstigen Fälle ergibt sich daraus, dass man *M* A-s auf *N* Verse auf $\binom{N}{M}$ Weisen aufteilt; von diesen *M* Fällen kann man *x* Fälle so platzieren, dass sie in gleichen Versen vorkommen wie die B-s, und zwar auf $\binom{M}{x}$ Weisen. Die verbleibenden *N-M* Fälle von A verteilt man auf die *n-x* Verse, in denen kein B vorkommt, auf $\binom{N-M}{n-x}$ Weisen. Setzt man alle diese Möglichkeiten zusammen, so erhält man

$$P(X = x) = \frac{\binom{N}{M}\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{M}\binom{N}{n}} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}. \tag{5.18}$$

Da wir aber nicht nur die Wahrscheinlichkeit von genau *x* Koinzidenzen, sondern die von x oder mehr suchen, berechnen wir

$$P(X \geq x) = \sum_{i=x}^{\min(M,n)} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} . \tag{5.19}$$

Ist der Schweif der Verteilung zu lang, so kann man leichter mit dem Komplement rechnen, nämlich

$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} . \tag{5.20}$$

Beispiel: In Schema (I) des vorigen Abschnitts findet man Hreb 1 [Kind] in $M = 21$ Versen, Hreb 8 [mit] in $n = 6$ Versen, und insgesamt haben wir $N = 32$ Verse. Die beiden Hrebs kommen zusammen in $x = 4$ Versen vor. Die Wahrscheinlichkeit, dass sie in $x$ (= 4) oder in mehr Versen gemeinsam vorkommen, berechnet sich nach

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) =$$

$$= \frac{\binom{21}{4}\binom{32-21}{6-4} + \binom{21}{5}\binom{32-21}{6-5} + \binom{21}{6}\binom{32-21}{6-6}}{\binom{32}{6}} =$$

$$= \frac{5985(55) + 20349(11) + 54264(1)}{906192} = 0.6701.$$

Da diese Wahrscheinlichkeit größer ist als 0.05, besteht zwischen [Kind] und [mit] keine Koinzidenz.

Auf diese Weise berechnet man die Koinzidenzen aller Hrebs mit allen anderen, ausgehend von Schema (I). Das von R. Köhler geschriebene Programm für die mechanische Auswertung ist auf der Diskette zum Buch Ziegler, Altmann (2002) zu finden.

Betrachtet man die grammatischen Koinzidenzen als deterministisch und die restlichen positionalen als stochastisch und zeichnet man die Hrebs als Ecken und die Koinzidenzen als Kanten, dann bekommt man einen Graphen, der die **denotative Struktur** des Textes darstellt. Man kann sie auch als assoziative Struktur bezeichnen, denn das, was oft genug nah beieinander steht, assoziiert sich auf irgendeine Weise miteinander. Die Einbeziehung der deterministischen Koinzidenzen führt zu einer **erweiterten denotativen Struktur**. Zweifellos haben die Erweiterungen kein Ende, man kann sie in verschiedene Richtungen vorantreiben, z.B. in phonetischer, psycholinguistischer, soziologischer usw. Hinsicht. Auf jeden Fall ist hiermit ein objektiver Weg zur Erforschung der latenten Struktur eines Textes gegeben.

## 5.8. Der Graph des Textes

Die ermittelten Koinzidenzen kann man in Form eines Graphen darstellen. Die Ecken stellen Hrebs dar, die Kanten Koinzidenzen. Der Graph in Abbildung 5.3 stellt die erweiterte denotative Struktur des „Erlkönig“ dar. Einige der Kanten sind mit *d* gekennzeichnet und repräsentieren die deterministische Koinzidenz.

Zwei Ecken, die mit einer Kante verbunden sind, bezeichnet man in der Graphentheorie als adjazent. Im Text haben auch die im Graphen isolierten Ecken {4,9} Koinzidenzen, jedoch mit $P > \alpha$. Der resultierende Graph hat

$n =$ 88 Ecken, die den Hrebs entsprechen,
$m =$ 128 Kanten, die den Koinzidenzen entsprechen,
$k =$ 5 (Zusammenhangs)Komponenten (K1,…,K5).

Eine Zusammenhangskomponente ist ein (Sub)Graph, in dem es zwischen allen Ecken einen Weg (eine Folge von Kanten) gibt.

Ein Graph hat zahlreiche Eigenschaften, die man in der Textanalyse verwenden kann. Einige von ihnen werden weiter unten präsentiert.

Die Grenze $\alpha = 0.05$, die wir hier angenommen haben, ist nur eine von vielen möglichen. Man könnte die Grenze auch anders ansetzen. Bei der hier gewählten bleiben einige Ecken im „Erlkönig“ noch isoliert, besonders wenn man die deterministischen Verbindungen nicht berücksichtigt. Auch die Länge des Textes könnte Einfluss auf die Grenze haben. Daher könnte man noch eine andere Grenze $\beta$ definieren, bei der es keine isolierten Ecken gibt und jede mit mindestens einer anderen verbunden wäre. Auf dieser Ebene bleiben aber noch immer mehrere Komponenten bestehen. Daher kann man eine dritte Grenze $\gamma$ festlegen, bei der es nur eine einzige Komponente des Graphen gibt. Diese drei Grenzen bilden dann eine Art Kohärenzcharakteristik des Textes. Wir hätten

$\alpha$: signifikante Koinzidenzen (isolierte Ecken möglich)
$\beta$: keine isolierten Ecken (mehrere Komponenten möglich)
$\gamma$: nur eine einzige Komponente.

Auch diese Grenzenwahl hängt natürlich davon ab, wie man die Kriterien der Hrebbildung ansetzt. Und es wäre sehr interessant zu untersuchen, wie sich $\beta$ und $\gamma$ unter verschiedenen analytischen Kriterien verhalten.

### 5.8.1. Zusammenhang

Die Stärke des Zusammenhangs in einem Graphen lässt sich auf verschiedene Weisen messen. Maße dieser Art geben an, wie dicht das Netz der Kanten des Graphen ist oder, textlinguistisch interpretiert, wie stark die denotierte Realität konzentriert ist, wie stark im Text die Realitätsausschnitte zusammenhängen. In Ziegler, Altman (2002) wurden zwei Maße aus der Graphentheorie vorgeschlagen, nämlich das **relative Zusammenhangsmaß** ($n$ = Zahl der Ecken, $k$ = Zahl der Komponenten)

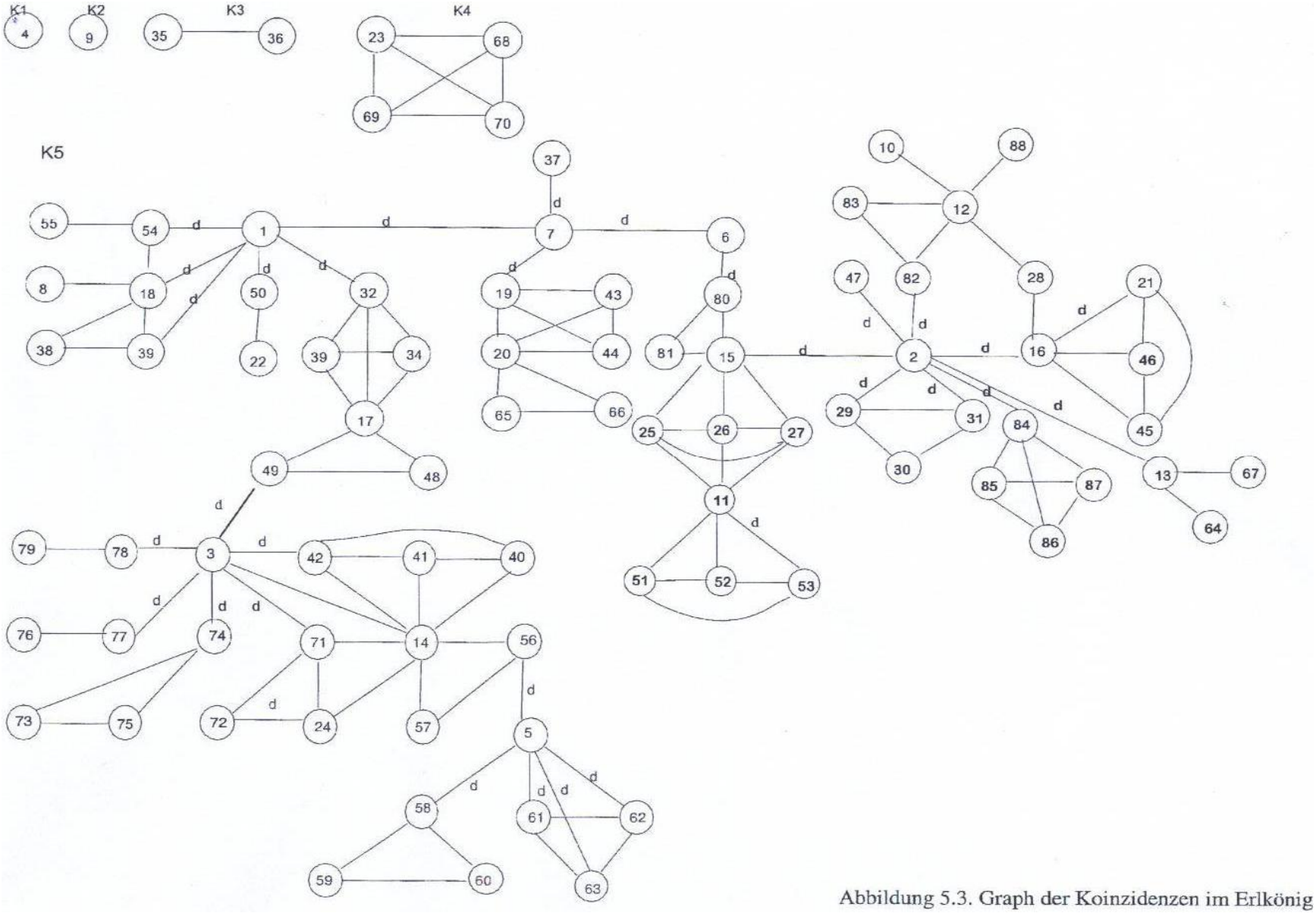


Abbildung 5.3. Graph der Koinzidenzen im Erlkönig

$$(5.21)\ \kappa_{rel} = \frac{n-k}{n-1},$$

das nur die Zahl der Ecken und die der Komponenten in Betracht zieht. Aus den oben genannten Zahlen (Abschnitt 5.8) erhält man für den Erlkönig

$$\kappa_{rel} = \frac{88-5}{88-1} = 0.9540.$$

Je mehr Komponenten es gibt, desto schwächer wird der Zusammenhang. Dieses Maß hängt direkt von der Grenze α ab, die man ausgewählt hat.

Ein anderes Maß, das auch die Zahl der Kanten in Betracht zieht, ist die relative zyklometrische Zahl (m = Zahl der Kanten)

$$(5.22)\ \mu_{rel} = \frac{2(m-n+k)}{(n-2)(n-2)}.$$

In unserem Fall erhalten wir

$$\mu_{rel} = \frac{2(128-88+5)}{87(86)} = 0.0120.$$

Diese Zahl kann man konventionell mit 100 multiplizieren, um sie etwas anschaulicher zu machen. Für verschiedene Textsorten wird es sicherlich unterschiedliche Zusammenhangsstärken geben, aufgrund derer man sie eventuell charakterisieren kann. Sie werden aber nur dann vergleichbar sein, wenn die Analyse unter gleichen Kriterien stattgefunden hat. Unter Auslassung der deterministischen Zusammenhänge erhielten Ziegler und Altmann (2002) für den „Erlkönig“ $\kappa_{rel} = 0.83$ und $\mu_{rel} = 0.0134$.

### 5.8.2. Eckengrad

Die Zahl der Kanten, die mit einer gegebenen Ecke inzidieren, heißt Eckengrad d(v). In Abbildung 5.3 hat die Ecke Nr. 4 den Grad d(4) = 0, Nr. 35 hat d(35) = 1, Nr. 23 hat d(23) = 3, usw. Der Grad einer Ecke zeigt, linguistisch gesehen, die **phraseologiche Potenz** des Hrebs, denn Hrebverbindungen sind die Quelle gegenwärtiger oder zukünftiger Phrasenbildung. Textanalytisch erzeugen Adjazenzen immer Konnotationen, die eine unterschiedliche Überlebensdauer haben. Der Grad eines Hrebs zeigt daher dessen **konnotative Potenz**. Die mit dem gegebenen Hreb adjazenten Hrebs zeigen, welche Assoziationen der Hreb hervorgerufen hat. Sie gehören nicht direkt zum Denotat, sie stehen mit ihm in konno-

tativer Relation. Die Grade einzelner Hrebs sind in Tabelle 5.6 aufgeführt.

Tabelle 5.6
Grade der Hrebs im „Erlkönig“

| Grad | Hreb |
|---|---|
| 0 | {und, so} |
| 1 | {mit, in, ruhig, Krone, Schweif, Nebelstreif, hören, dort, genau, jetzt, Leids, tot} |
| 2 | {es, wohl, liebes, leise, bleiben, feiner, warten, Reihn, nächtlich, düster, Ort, reizen, willig, Gewalt, anfassen, tun, geschwind, ächzend} |
| 3 | {Wind, sehen, Mutter, Weide, Gestalt, fassen, sicher, warmhalten, bang, Gesicht, kommen, gar, Spiel, bunt, Strand, wand, gülden, versprechen, dürr, Blatt, säuseln, wollen, sollen, führen, wiegen, tanzen, einsingen, scheinen, alt, grau, lieben, brauchen, grausen, halten, Hof, |
| 4 | Mühe, Not} |
| 5 | {sein, Blume, spät, durch, Nacht, bergen, erreichen} |
| 6 | {Tochter, Arm, haben, was, an} |
| 7 | {Kind, reiten, gehen} |
| 8 | {Erlkönig, schön} |
| | {Vater} |

Wie man sieht, weisen die Kernhrebs {Kind, Vater, Erlkönig} die höchsten Grade auf, jedoch lassen sich die anderen Hrebs wegen der Kürze des Textes linguistisch nicht einordnen. Ziegler und Altmann (2002) haben bei nichtdeterministischer Analyse für die Kernkrebs gerade umgekehrt die niedrigsten Grade gefunden.

Als Charakteristika kann man drei Maße benutzen, nämlich den **relativen Gesamtgrad** des Graphen

$$(5.23)\quad d_{rel} = \frac{2m}{n(n-1)},$$

der als Verhältnis der beobachteten zu der maximalen Kantenzahl definiert ist. In unserem Fall ergibt sich

$$d_{rel}(Erlkönig) = \frac{2(128)}{88(87)} = 0.0334.$$

Weiter gibt es den **durchschnittlichen Eckengrad** des Graphen

(5.24) $\bar{d} = \frac{2m}{n}$,

hier $\bar{d}$ = 2(128)/88 = 2.91, und den **maximalen Eckengrad**

(5.25) $\Delta = d_{\max}(G)$,

in unserem Fall Δ(Erlkönig) = 8. Formel (5.23) kann man als Maß der konnotativen Konzentration des Textes betrachten, denn je größer dieses Maß ist, desto stärker ist der abgebildete Ausschnitt der Realität miteinander sprachlich, assoziativ u.ä. verwoben.

Der Eckengrad spielt in der Netzwerktheorie eine bedeutende Rolle. Graphen werden auch danach beurteilt, wie sich bestimmte Aspekte der Eckengrade gestalten (vgl. Albert, Barabási 2002; Dorogovtsev, Mendes 2002). Es gibt darüber eine schier unübersichtliche Literatur.

### 5.8.3. Entfernungen

Die Entfernung zwischen zwei Ecken x und y ist der kürzeste Weg zwischen ichnen, d.h. die minimale Zahl der Kanten, die zwischen ihnen liegen. Wir bezeichnen sie als $d(x,y)$. So ist die Entfernung zwischen Hreb 1 und Hreb 2 in Abbildung 5.3 $d(1,2) = 5$, aber für $d(1,4) = \infty$, weil die Komponenten nicht verbunden sind. Entfernung hat also Sinn nur bei einer Zusammenhangskomponente.

Unter **Exzentrizität** der Ecke $x$, $e(x)$, bezeichnet man deren größte Entfernung in der Komponente, d.h.

(5.26) $e(x) = \max_{y \in G} d(x, y)$.

Die maximale Exzentrizität (maximale Entfernung) im Graph ist der Diameter $dm(G)$ des Graphen, d.h.

(5.27) $dm(G) = \max_{x \in G} e(x)$.

Der Diameter stellt eigentlich die **Denotationsbreite** dar, weil er numerisch die Breite des Ausschnitts aus der Realität ausdrückt, den der Text erfasst (vgl. Ziegler, Altmann 2001).

Das **Zentrum des Graphen** bilden diejenigen Ecken, deren Exzentrizitäten am kleinsten sind. Es können eine oder mehrere Ecken sein.

Die Summe aller Entfernungen in einer Komponente definieren wir als

(5.28) $z(K) = \sum_{x,y \in K} d(x,y)$.

Man sollte darauf verzichten, diese Indizes „von Hand“ zu rechnen, weil man dabei unweigerlich Fehler machen wird. Man benutze besser fertige Programme wie z.B. die Software MAPLE, die ein graphentheoretisches Modul hat, oder Pajek, Uci6 u.a., die man aus dem Internet herunterladen kann. Nur bei sehr kleinen Komponenten kann man ohne Fehlergefahr rechnen. So ist z.B. für jede Ecke der Komponente K4 in Abbildung 5.3 die Summe aller Entfernungen jeweils gleich 3, weil zwischen jeweils zwei Ecken $d(x,y) = 1$. Daher ergibt sich bei 4 Ecken $z(\text{K4}) = 4(3) = 12$. In dieser Komponente ist für alle Ecken die Exzentrizität $e(x) = 1$, so dass auch der Diameter $dm(\text{K4}) = 1$ und das Zentrum alle vier Hrebs bilden.

Die *mittlere Entfernung* im Graphen, d.h. in einer Komponente, drückt man mit

(5.29) $\bar{d} = \frac{1}{n_K(n_K - 1)} \sum_{x,y \in K} d(x,y) = \frac{z(K)}{n_K(n_K - 1)}$

aus, wo $n_K$ die Zahl der Ecken in der Komponente ist. Diesen Ausdruck bezeichnet man auch als den *zentralen Index* des Graphen. Die mittlere Entfernung drückt die **assoziative Stärke** aus, die zwischen den Hrebs einer Komponente besteht. Je stärker die Assoziation, desto kleiner wird die mittlere Entfernung, jedoch wurde beobachtet, dass die Größe der Graphen starken Einfluss auf dieses Maß ausübt, weil mit dem Anwachsen der Zahl der Ecken die Entfernungen ungleichmäßig zunehmen. Werden neue Ecken an der Peripherie des Graphen platziert, so vergrößert sich die mittlere Entfernung trotz z.B. der starken Verbundenheit mit den peripheren Ecken; werden sie „in der Mitte“ platziert, so verringert sie sich. Daher empfiehlt es sich eher, ein *relatives Zentralitätsmaß* zu benutzen, das sich nach Ziegler und Altmann (2002) als

(5.30) $Z = \frac{z_{\max} - z}{z_{\max} - z_{\min}}$

$= \frac{(n+1)n(n-1) - 3z}{n(n-1)(n-2)}, \quad \textit{für } n > 2$

ergibt. Hier ist $n$ wieder die Zahl der Ecken in der Komponente, und $Z$ hat nur dann Sinn, wenn $n > 2$ ist. So ergibt sich beispielsweise für die größte Komponente K5 mit $n_5 = 80$

$$Z(K5) = \frac{81(80)79 - 3(46410)}{80(79)78} = 0.7560.$$

Alle diese Maße findet man in Tabelle 5.7.

Tabelle 5.7
Entfernungsindizes im „Erlkönig“

| Komponente | $n_K$ | z(K) | dm(K) | $\overline{d}$ | Z(K) | Zentrum |
|---|---|---|---|---|---|---|
| K1 | 1 | 0 | 0 | 0 | - | {4} |
| K2 | 1 | 0 | 0 | 0 | - | {9} |
| K3 | 2 | 2 | 1 | 1 | - | {35, 36} |
| K4 | 4 | 12 | 1 | 1 | 1 | {23,68,69,70} |
| K5 | 80 | 46410 | 17 | 7.34 | 0.756 | {1,32} |

### 5.8.4. Schnittmengen und Cliquen

Betrachtet man den Graphen in Abbildung 5.3, so sieht man, dass einzelne Ecken einen unterschiedlichen Status haben. Einige von ihnen sind isoliert, z.B. Nr. 4 und 9, andere stehen an der Peripherie der Komponente K5, haben den Grad 1 und bilden Endecken, z.B. Nr. 76, 79. Einige stehen zwar zentraler, haben aber einen niedrigen Grad, z.B. Nr. 6, andere stehen zwar peripher, haben aber einen hohen Grad wie Nr. 11. Sie zeichnen sich also alle aus durch eine etwas umständlich ausdrückbare Assoziativität, deren Erforschung erst beginnt

Eine andere Qualität der Hrebs, die man am Graphen ablesen kann, ist deren Verbindungsrolle im Graphen, die man dann bewerten kann, wenn man die gegebene Ecke entfernt. Zerfällt der Graph nach der Entfernung einer Ecke in Komponenten, dann bezeichnet man die gegebene Ecke als Schnittecke oder Artikulationspunkt. Wenn man z.B. Hreb Nr. 1 beseitigt, dann zerfällt K5 in 4 kleinere Komponenten. Wir können sagen, dass Hreb Nr. 1 vier denotative Ausschnitte aus der Realität assoziativ verbindet, d.h., seine **assoziative Stärke** (A) ist 4.

Nicht alle Ecken sind Schnittecken. Die Schnittecken bilden eine Schnittmenge. Die Schnittecken und die Zahl der Komponenten (A), die sie verbinden, sind in Tabelle 5.8 dargestellt.

Wie man sieht, sind die Hrebs mit der größten konnotativer Potenz (ab Grad 5 in Tabelle 5.6) eine Teilmenge der Schnittmenge der Ecken. Offensichtlich besteht zwischen der konnotativer Potenz und der assoziativen Stärke eine positive Korrelation.

Tabelle 5.8
Schnittecken und ihre assoziative Stärke

| Nr. | Hreb | A | Nr | Hreb | A |
|---|---|---|---|---|---|
| 2 | Vater | 6 | 18 | gehen | 2 |
| 3 | Erlkönig | 5 | 19 | Blume | 2 |
| 1 | Kind | 4 | 20 | an | 2 |
| 7 | sein | 4 | 32 | bergen | 2 |
| 5 | Toch | 3 | 49 | versprechen | 2 |
| 12 | Arm | 3 | 50 | bleiben | 2 |
| 13 | sehen | 3 | 56 | sollen | 2 |
| 15 | reiten | 3 | 58 | führen | 2 |
| 6 | es | 2 | 74 | brauchen | 2 |
| 11 | Wind | 2 | 77 | anfassen | 2 |
| 14 | schön | 2 | 78 | tun | 2 |
| 16 | haben | 2 | 81 | geschwind | 2 |
| 17 | was | 2 | 84 | erreichen | 2 |

In dem Graphen sieht man auch, dass sich die Hrebs auf verschiedene Weisen „klumpen“. Hier werden wir nur die Cliquen in Betracht ziehen, d.h. solche Subgraphen, in denen alle Ecken miteinander adjazent sind. Nach dieser Definition lässt sich der ganze Graph in Cliquen zerlegen, wobei jede Ecke nur zu einer Clique gehören darf. Wir definieren hier eine Textclique unter der Zusatzbedingung, dass jede Ecke in der Clique die gleichen Adjazenzen mit dem restlichen Graphen haben muss. Eine **Textclique** stellt also eine Gruppe von Hrebs mit der stärksten inneren Assoziation dar. So stellen K3 und K4 sowohl Cliquen als auch Textcliquen dar. Die Menge {l8, 54, 55} ist zwar eine Clique, aber keine Textclique, weil jede ihrer Ecken noch andere, unterschiedliche Adjazenzen hat; die Menge {58, 59, 60} bildet eine Clique, aber nur {59, 60} ist eine Textclique. Auf diese Weise lassen sich folgende Textcliquen erkennen:

| | |
|---|---|
| {23, 68, 69, 70} | {Weide, scheinen, alt, grau} |
| {25, 26, 27} | {spät, durch, Nacht} |
| {61, 62, 63} | {wiegen, tanzen, einsingen} |
| {51, 52, 53} | {dürr, Blatt, säuseln} |
| {85, 86, 87} | {Hof, Mühe, Not} |
| {21, 45, 46} | {Mutter, Gewand, gülden} |
| {35, 36} | {Kron, Schweif} |
| {33, 34} | {bang, Gesicht} |
| {73, 75} | {willig, Gewalt} |

| | |
|---|---|
| {59, 60} | {Reihn, nächtlich} |
| {43, 44} | {bunt, Strand} |
| {65, 66} | {düster, Ort} |
| {40, 41} | {gar, Spiele} |
| {29, 31 } | {fassen, warmhalten} |

An den meisten Textcliquen erkennen wir die im Text vorkommenden Phrasen. Die Texcliquen sind eher ad hoc-linguistische Formationen, die sich als sekundäre Assoziationen in Distanz vom Textkern bilden.

# 6. Grammatik

Die grammatische Analyse eines ganzen Textes ist in der neuen Zeit besonders in der Korpuslinguistik in den Mittelpunkt geraten, weil man sich davon erhofft, auf induktivem Wege den „kompetenten Sprecher“ zu ersetzen und objektivere Resultate zu erzielen. Bei den riesigen Datenmengen, die dabei zu bewältigen sind, wäre manuelles Arbeiten völlig illusorisch. Die Ziele der Korpuslinguistik sind aber andere als die der Textanalyse, beide jedoch sind berechtigt und sowohl wissenschaftlich als auch praktisch von Bedeutung. Textanalytisch hat sich in diesem Bereich noch wenig getan, und erst weitere, besonders vergleichende Analysen werden zeigen, inwieweit grammatische Erscheinungen bei der Textanalyse eine Rolle spielen. Daher sind alle Ausführungen in diesem Kapitel cum grano salis zu nehmen.

Die grammatische Analyse kann man eventuell in eine morphologische und eine syntaktische Analyse aufteilen. Bei qualitativen Analysen ist das Ziel erreicht, wenn man Wörter und Sätze klassifiziert und deren Strukturen graphisch aufzeichnet oder auf eine algebraische Art darstellt, z.B. mit einem Klammersystem oder einem Graphen. Etwas unangenehm bei qualitativen Analysen ist die Tatsache, dass sie mit vielerlei Vagheiten behaftet sind, vom Typ der Grammatik abhängen und alternative Lösungen zulassen. Dieses notwendige Übel rührt aber von der Natur der Sprache selbst her und lässt sich mit keinerlei kategorischen Entscheidungen aus der Welt schaffen.

Bei der quantitativen Analyse ist das Resultat der qualitativen Analyse lediglich eine Voraussetzung, die erfüllt werden muss. Die weiteren Schritte sind Metrisierung und Messung, induktive Erfassung mit empirischen Formeln und schließlich deduktive Ableitung aus Theorien oder Einzelgesetzen, d.h. theoretische Systematisierung. Bei der poetologischen, literaturwissenschaftlichen, textologischen Analyse von Einzeltexten kann man natürlich nicht so weit gehen, weil es sich hier vor allem um die Eigentümlichkeiten, Idiosynkrasien des Textes handelt. Es ist selbstverständlich, dass man quantitative Resultate aus Einzeltexten auch für Urteile über die gegebene Sprache (mit Vorsicht) verwenden kann, jedoch ist dies eher ein Ziel von Linguisten, nicht eines von Textanalysten.

In den weiteren Abschnitten werden nur einige wenige Aspekte der grammatischen Phänomene erfasst. Die Entscheidung darüber, welche Indizes, Messungen usw. sinnvoller sind als andere, lässt sich induktiv und nur aufgrund der Untersuchung eines einzigen Textes nicht fällen. Das Gütekriterium ist nicht die Einfachheit oder die bessere Erfassung der Daten mit einer Kurve, sondern deren Ableitbarkeit, zumindest ihre Kompatibilität mit einer Theorie. Diese Art des Vorgehens ist bei den Textanalytikern vorläufig Mangelware, theoretische Pionierarbeit hat vor allem L. Hřebíček geleistet (z.B. 1993, 1995,1997, 2000).

## 6.1. Morphologische Eigenschaften

Die morphologischen Eigenschaften pflegt man mit Indizes zu erfassen. Die bekanntesten sind diejenigen von Greenberg (1960), es wurden jedoch schon vor Greenberg einige Indizes vorgeschlagen (vgl. Skalička 1935; Thorndike 1943; Hamp 1958).

Greenbergs Indizes wurden später modifiziert (vgl. Krupa 1965), und es wurde zum Glück festgestellt, dass man eine Eigenschaft auch auf andere Weisen operationalisieren kann (s. z.B. Kelemen 1970; Robins 1965; Slavíčková 1968; Krupa 1965), wodurch man einen großen Schritt zur Abschwächung der Absolutisierung unserer begrifflichen Konstrukte getan hat. Eine knappe Zusammenfassung findet man bei Altmann, Lehfeldt (1973) und Kempgen, Lehfeldt (2004). Ein Index gibt eine Eigenschaft nicht eindeutig wieder, jeder ist nur eine Approximation. Jedoch lässt sich in keinem Fall sagen, was der „wahre" Wert sein könnte. Hier werden wir exemplarisch nur den Synthetismus behandeln, in der Hoffnung, dass einiges aus diesem Bereich für die Textanalyse von Wert sein kann.

Unter Synthetismus versteht man die Eigenschaft, Wörter so komplex zu gestalten, dass sie im Extremfall mit einem ganzen Satz zusammenfallen, was man in einigen Indianersprachen des öfteren finden kann, oder – im anderen Extremfall – dass alle Begriffe durch Einzelwörter ausgedrückt werden, was einen totalen Analytismus bedeuten würde. Alle Sprachen liegen zwischen diesen beiden Extremen. Der Synthetismus kann sich in unterschiedlichen Bereichen einer Sprache unterschiedlich stark durchsetzen. So braucht man für das Analogon des Satzes „Ich lebe" im Ungarischen nur ein Wort, im Japanischen drei Wörter. Für das Analogon des Satzes „Ich lebe in meinem Haus" braucht man im Indonesischen 3 bis 4 Wörter (je nach Orthographie und Höflichkeitsstufe), im Ungarischen oder im Slowakischen 4, im Deutschen oder im Englischen 5, im Japanischen 7, usw.

Die als Synthetismus bezeichnete Eigenschaft lässt sich unterschiedlich operationalisieren. Bei der Aufstellung von Indizes sollte man Wert darauf legen, dass sie in einem kleineren Intervall, am besten zwischen 0 und 1, liegen und dass sie vergleichbar sind. Die erste Forderung lässt sich durch geeignete Transformation immer erreichen, bei der zweiten ist darauf zu achten, dass die sogenannte Stichprobenverteilung des Indexes bekannt ist. Dies lässt ich mathematisch zumindest asymptotisch immer erreichen, jedoch sollte man komplizierte Indizes vermeiden.

Benutzt man die Bezeichnungen

$W$ - Zahl der Wörter im Text
$W_1$ - Zahl der Wörter im Text, die aus einem Morphem bestehen
$M$ - Zahl der Morpheme im Text
$R$ - Zahl der Wurzelmorpheme im Text
$S$ - Zahl der Sätze im Text,

dann findet man in der Literatur vier Charakteristika des Synthetismus:

(a) Greenberg-Krupa: $GK = W/M$
(b) Kelemen: $Ke = W_1/W$
(c) Slavíčková: $Sl = R/M$
(d) Skalička: $Sk = S/W$.

Alle Indizes stellen Proportionen dar, d.h., alle liegen im Intervall <0,1>, und sie lassen sich als solche weiter verarbeiten. Die ersten drei beleuchten den gleichen Aspekt, so dass zur Charakterisierung eines Textes (a) und (d) ausreichen. Der einfachere von ihnen ist Skaličkas Index *Sk*, weil man für ihn keine morphologische Analyse durchführen muss. Er ist im Grunde der Kehrwert der durchschnittlichen Satzlänge (gezählt als Wortzahl) in einem Text, was eine vernünftige Textcharakteristik ist. Bei der Bestimmung der Morphemzahl gibt es leider immer sowohl Unsicherheiten als auch Bewertungsprobleme, denn betrachtet man das Morphem als eine formale Einheit, dann besteht „war" aus einem Morphem, „ist" aus einem oder zwei Morphemen, je nach dem, ob man „ist" per analogiam (zu „bin", „sind", usw.) für Suppletivismus hält oder für ein Verb in der 3. Person Sg. mit der Endung „-t". In anderen Sprachen entstehen zahlreiche andere Probleme. Hier ziehen wir die minimalistische Bewertung vor.

Im „Erlkönig" gibt es 225, Wörter und die Zahl der Morphe lässt sich mit 324 beziffern. Daher hat der Index (a) den Wert

GK = 225/324 = 0.69.

Jedoch ist auch die Segmentierung in Sätze nicht ohne Probleme. Auch bei diesem Problem muss man eine Entscheidung darüber treffen, ob man Doppelpunkt, Semikolon u.a. für Grenzsignale hält (vgl. Niehaus 1997). Auch hier halten wir uns an die minimalistische Bewertung und betrachten alle Grenzsignale außer dem Komma als Satzendesignale. So erhalten wir insgesamt 25 Sätze, und der Index (d) hat den Wert

*Sk* = 25/225 = 0.11.

Das Maß des Analytismus ist das Komplement zu dem Index, d.h.

*Analytismus* = 1 – *I*.

In der Textanalyse hat man diese Indizes (bis auf die Satzlänge) bisher noch nicht benutzt, weil ihre Werte für Texte einer Sprache vermutlich sehr nah beieinander liegen würden. Nur Skaličkas Index verspricht eine größere Variabilität. Je variabler er jedoch in einer Sprache ist, desto weniger ist er für die Typologie geeignet.

## 6.2. Syntax

Die Erfassung syntaktischer Eigenschaften hängt größtenteils davon ab, welche Art von Grammatik man zugrundelegt. Grundsätzlich lassen sich aber alle syntaktischen Eigenschaften quantifizieren.

R. Köhler (1999) analysiert in seiner synergetischen Betrachtung der Syntax syntaktische Konstrukte und quantifiziert folgende Eigenschaften:

- Frequenz: Vorkommen des Konstrukts im Korpus
- Länge: Zahl der Endknoten (Wörter) in der Konstituente
- Komplexität: Zahl der unmittelbaren Konstituentenin der Konstituente
- Position: Stelle in der übergeordneten Konstituente von links nach rechts
- Tiefe der Einbettung: Zahl der Produktionsschritte vom Startsymbol aus
- Information: Gedächtnisraum, der für die temporäre Speicherung der grammatischen Relationen in der Konstituente notwendig ist
- Polyfunktionalität: Zahl unterschiedlicher Funktionen der Konstruktion
- Synfunktionalität: Zahl der unterschiedlichen Funktionen, mit der sich die gegebene Funktion an der syntaktischen Repräsentation beteiligt.

Ausgehend von der graphentheoretischen Darstellung der Phrasenstruktur, quantifizieren R. Köhler und G. Altmann (1999) einige dieser Eigenschaften und suchen zunächst nach Wahrscheinlichkeitsverteilungen dieser Variablen im Korpus. Einige andere Eigenschaften werden in Altmann, Lehfeldt (1973) gezeigt. Auch einige Greenbergsche Indizes (1960) sind syntaktischer Natur (vgl. Kempgen, Lehfeldt 2005).

Man sieht also, dass die Syntax der Quantifizierung nicht entgehen kann, auch wenn sie als Wissenschaft von Relationen betrieben wird. Im Gegenteil, erst dadurch kann man hier zu Selbstregulations- und Verteilungsmodellen gelangen. Textanalytiker sind kaum an der syntaktischen Strukturierung des Textes interessiert, weil man hier oft zahlreiche Randbedingungen und „Ausnahmen“ berücksichtigen muss. Der Textanalytiker interessiert sich für globale Eigenschaften, die messbar sind und sich im Text auf eine bestimmte Weise ausprägen. Die Erfassung lässt sich so bewerkstelligen, dass man die Satzkonstruktion auf eine abstraktere Ebene überführt und die Quantifizierung dort durchführt.

In mehreren Grammatikmodellen hat sich die Darstellung der abstrakten Struktur des Satzes in Form eines baumartigen Graphen bewährt. Die Dependenzgrammatik entfernt sich langsam von der baumartigen Graphenstruktur und baut auch Zyklen in die Graphen ein. Andere Grammatikmodelle bewahren aber die Baumstruktur mit *S* als Quelle, die syntaktischen Nexus als Zwischenecken und Wörtern als Endecken. Ein typischer Baum dieser Art ist

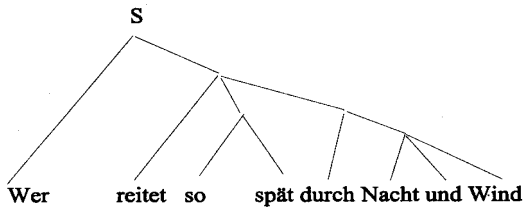


Abbildung 6.1

Je nach Grammatikmodll kann der Satz auch anders analysiert werden, z.B.

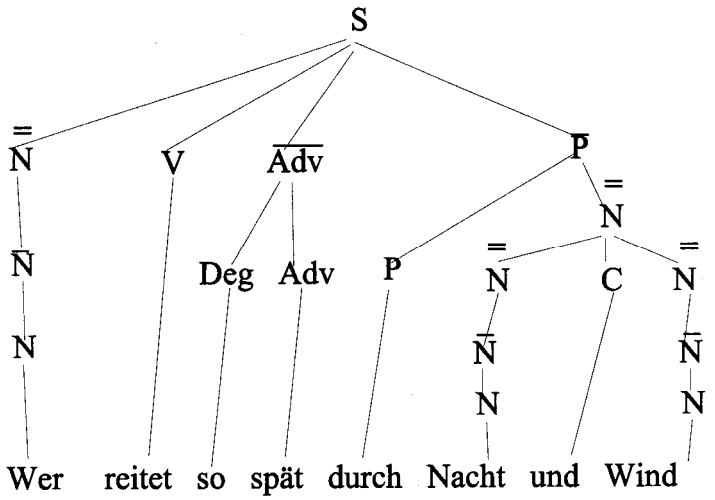


Abbildung 6.2.

oder, mit Auslassung der Zwischenecken ohne Abzweigung, wie in Abb. 6.3.

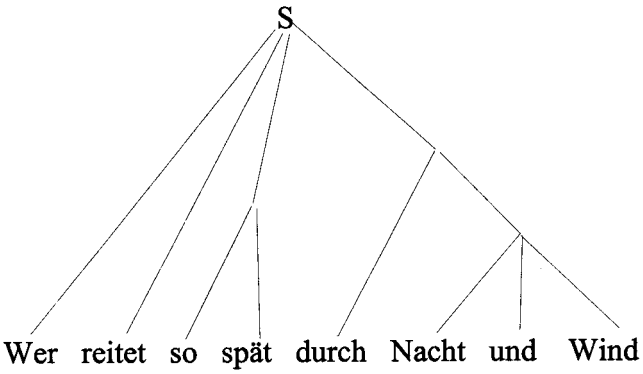


Abbildung 6.3.

Die obigen Graphen stellen sogenannte Bäume dar, deren graphentheoretische Eigenschaften für die Textologie von Nutzen sein könnten. Hier werden wir

exemplarisch nur eine einzige Eigenschaft, nämlich den binären Kode des Textes in Form von Bäumen untersuchen.[1]

### 6.2.1. Der binäre Kode

Zeichnen wir nochmals Abbildung 6.3, diesmal numerieren wir aber die Ecken, und zwar von oben nach unten und von links nach rechts. Dann bekommen wir Abbildung 6.4.

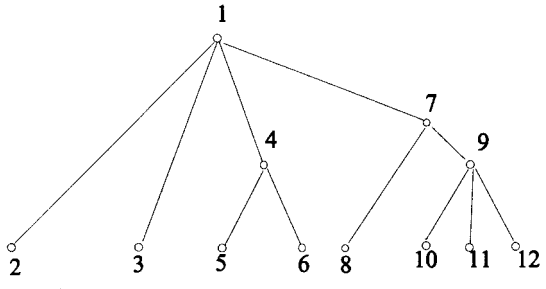


Abbildung 6.4

Einen Graphen kann man immer in Form einer Adjazenzmatrix kodieren, in der die Zeilen und Spalten die Ecken (hier $v_1, v_2, \ldots, v_{12}$) darstellen. Die Zahlen in der Matrix sind

$$a_{ij} = \begin{cases} 0, & \text{wenn die Ecken } i \text{ und } j \text{ nicht adjazent sind} \\ 1, & \text{wenn die Ecken } i \text{ und } j \text{ adjazent, d.h. mit einer Kante verbunden sind} \end{cases}$$

Der Graph in Abbildung 6.4 kann demnach als eine symmetrische Matrix in Tabelle 6.1 dargestellt werden. Die Diagonale ist für unsere Zwecke irrelevant.

Diese Matrix lässt sich mit einer einzigen Zahl so kodieren, dass man sie aus dieser Zahl wieder rekonstruieren kann, wenn man weiß, wie viele Ecken es in der Matrix gibt.

Als den *binären Kode* (*BK*) bezeichnen wir die Zahl (s. Balakrishnan 1997: 22)

$$(6.1) BK = a_{12}2^0 + a_{13}2^1 + \ldots + a_{1n}2^{n-1} + a_{23}2^n + \ldots + a_{2n}2^{2n-3} + \ldots + a_{n-1,n}2^{k-1},$$

wo $k = n(n\text{-}1)/2$ ($n$ ist die Zahl der Ecken). Wie man sieht, braucht man zur Berechnung nur die obere Dreiecksmatrix, weil die Diagonale irrelevant und die Matrix des ungerichteten Graphen symmetrisch ist. Für die Matrix in Tabelle 6.1 erhalten wir explizit

[1] Wir bedanken uns bei Reinhard Köhler für die syntaktische Analyse des Gedichtes.

Tabelle 6.1
Adjazenzmatrix des Graphen 6.4

| *v* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | - | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 1 | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 0 | 0 | 0 | 0 | 0 | - | 1 | 1 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | - | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | - | 1 | 1 | 1 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | - | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | - | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | - |

1. Zeile: $1(2^0) + 1(2^1) + 1(2^2) + 0(2^3) + 0(2^4) + 1(2^5) + 0(2^6) + 0(2^7) + 0(2^8) + + 0(2^9) + 0(2^{10})$
2. Zeile: $0(2^{11}) + ... + 0(2^{20})$
3. Zeile: $0(2^{21}) + ...+ 0(2^{29})$
4. Zeile: $1(2^{30}) + 1(2^{31}) + 0(2^{32}) +…+ 0(2^{37})$
5. Zeile: $0(2^{38}) +... + 0(2^{44})$
6. Zeile: $0(2^{45}) + ... + 0(2^{50})$
7. Zeile: $1(2^{51}) + 1(2^{52}) + ...+ 0(2^{53}) + ...+ 0(2^{55})$
8. Zeile: $0(2^{56}) + ... + 0(2^{59})$
9. Zeile: $1(2^{60}) + 1(2^{61}) + 1(2^{62})$
10. Zeile: $0(2^{63}) + 0(2^{64})$
11. Zeile: $0(2^{65})$.

Da alle Summanden mit $a_{ij} = 0$ entfallen, ergibt sich die Summe dieser Zahlen als

$$BK = 2^0 + 2^1 + 2^2 + 2^5 + 2^{30} + 2^{31} + 2^{51} + 2^{52} + 2^{60} + 2^{61} + 2^{62} =$$
$$= 8077205934910210087.$$

Diese Zahl ist jedoch so groß, dass sie kaum eine Assoziation hervorruft. Da wir keine Rücktransformation zur Rekonstruktion des Baumes brauchen, relativieren wir sie einfach, indem wir sie durch den maximalen Wert von *BK* dividieren. Diesen Wert würden wir erhalten, wenn alle $a_{ij} = 1$ wären. Dann wäre

$$(6.2) \quad BK_{\max} = \sum_{i=0}^{\frac{n(n-1)}{2}-1} 2^i = 2^{\frac{n(n-1)}{2}} - 1.$$

In unserem Fall (mit $n$ = 12) würde sich $BK_{max}$ = 73786976294838206463 ergeben. Die relative Zahl erhielten wir als

$$(6.3) \quad BK_{rel} = \frac{BK}{BK_{\max}},$$

was in unserem Fall 0.1095 ergeben würde. Es ist zu bemerken, dass im Falle $a_{ij}$ = 1 für alle $i, j$ unser Graph kein Baum mehr wäre, sondern ein vollständiger Graph.

Wenn man auf die oben angegebene Weise die einzelnen syntaktischen Konstrukte des „Erlkönig“ der Reihe nach misst, dann bekommt man eine Sequenz von Zahlen, die man auf Regularitäten hin untersuchen kann. Für den „Erlkönig“ ergaben sich folgende Zahlen, die auf eine bestimmte Weise die Folge von Satzkonstruktionen charakterisieren:

0.1095; 03779; 0.3779; 0.0147; 0.0147; 0.4286; 0.3751; 0.0469; 1.0000; 0.1095; 0.4286; 0.3752; 0.3783; 1.0000; 0.3783; 0.3750; 0.3799; 0.3779; 0.4286; 0.4286; 0.0009; 0.4286; 0.4286; 0.4286; 0.3750; 0.0469; 0.4286; 0.0000001; 0.4286; 0.4286; 0.3779; 0.4286; 0.0147; 0.3751; 0.1111; 0.0146; 0.4286; 0.4286; 0.0009; 0.0469; 0.09557; 0.1111; 0.1095; 0.0029.

Für diese Zahlen lassen sich unterschiedliche charakteristische Maße berechnen, und es lassen sich auch charakteristische Maße des gesamten Verlaufs dieser Zahlen ermitteln, da es sich um eine Zeitreihe handelt, die Satzstrukturen numerisch wiedergibt. Da wir aber keine Vergleichsmöglichkeiten haben, verlegen wir die Auswertung auf später. In Abbildung 6.5 geben wir zumindest den Graphen dieses Verlaufs an, der ein Fraktal darstellt.

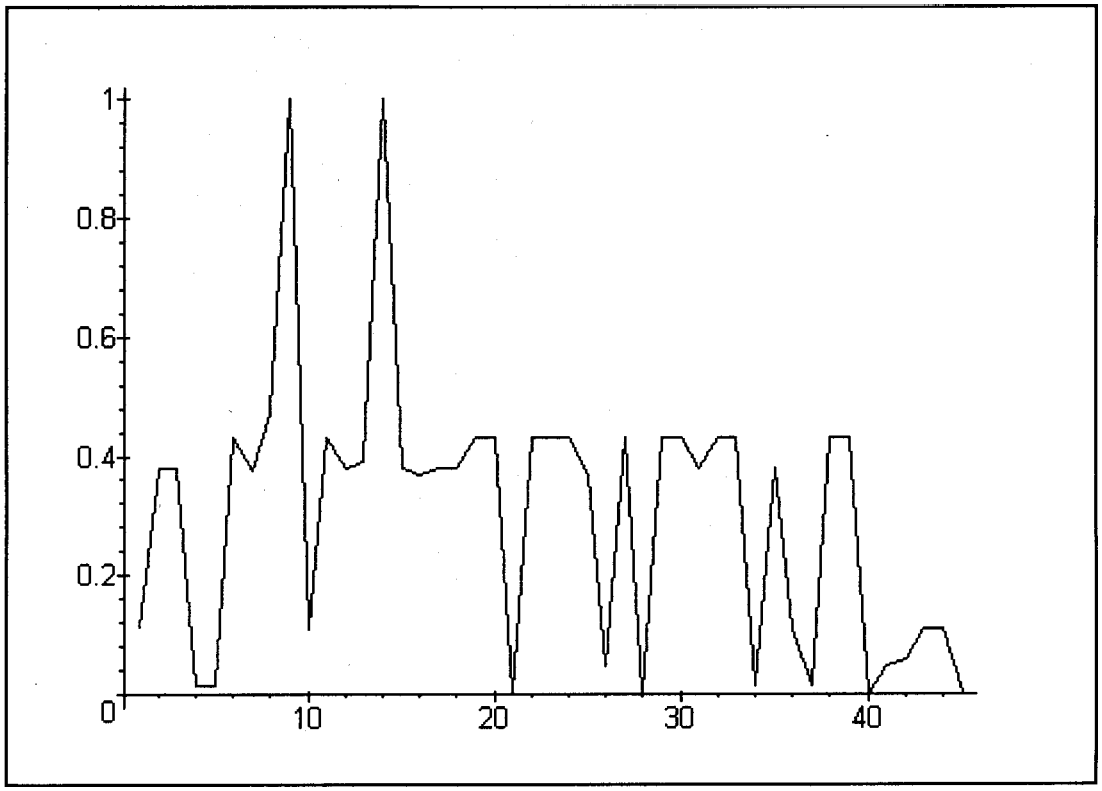


Abbildung 6.5. Verlauf des relativen binären Kodes für den „Erlkönig“

Wie man sieht, wiederholen sich in der Folge einige Zahlen bzw. liegen in einigen wenigen Intervallen. Aber bei so kurzen Texten ist die Erforschung der Verteilung nicht effektiv, und es ist auch nicht rational, Hypothesen aufzustellen. Das Problem muss zurückgestellt werden, liefert aber vielversprechende Perspektiven für Zeitreihen- und Fraktalforscher.

# 7. Schlusswort

Das vorliegende Buch stellt nur eine bescheidene Auswahl von Methoden dar, die die Untersuchung von dichterischen Texten zu objektivieren helfen. Die Zahl der Aspekte eines Textes ist wortwörtlich unendlich, und zu jedem Aspekt ist es möglich und nützlich, mathematische Methoden anzuwenden. Ist man am Theorieaufbau interessiert, so ist dies sogar unumgänglich.

Wir haben hier nur einige wenige Bereiche erörtert, die an sich alle umfangreiche Forschungsfelder darstellen. Die Untersuchung des sequentiellen Verhaltens des Textes haben wir nur teilweise berührt, vor allem weil der Text des „Erlkönig“ etwas kurz ist. Die Untersuchung der sequentiellen Aspekte ist zwar gut entwickelt, aber bisher wenig verbreitet. Besondere Aufmerksamkeit verdienen die Arbeiten von L. Hřebíček (1993a, 1996, 1997, 1997a, 1997b, 2000), die sich mit Chaos, Fraktalen, Persistenz, Lokation, Zeitreihen, Hierarchien usw. im Text beschäftigen und sicherlich als Anstoß zur Entwicklung umfangreicher Forschungsfeldern dienen werden.

Statistische Methoden kann man nicht nur auf Texte, sondern auch heuristisch verwenden, und auch hier stehen viele Türen offen, wenn man sich unterschiedlicher statistischer Disziplinen bedient.

In der Textanalyse wurde bisher am meisten die Graphentheorie vernachlässigt, die uns sowohl quantitative Möglichkeiten des Ausdrucks textlinguistischer Begriffe als auch die Möglichkeit, Relationen graphisch darzustellen und quantitativ auszuwerten, liefert (vgl. vor allem Skorochod´ko 1981), ganz zu schweigen von der heuristischen Potenz, die eine graphentheoretischen Darstellung mit sich bringt. Die hier verwendeten Methoden stellen nur einen bescheidenen Anfang dar, der weitergeführt werden sollte.

Die graphentheoretische denotative Analyse, eingeführt von A. Ziegler und G. Altmann (2002), ist ziemlich neu und verzichtet vorläufig auf standardisierte Verfahren z.B. zur Etablierung von Hrebs, die wir hier etwas anders gestaltet haben als in dem obengenannten Werk. Zahlreiche Texte in unterschiedlichen Sprachen müssen untersucht werden, bis man sich auf eine Batterie von Kriterien geeinigt haben wird. Die Übertragung graphentheoretischer Begriffe in die Textanalyse ist vielversprechend. Solche Begriffe haben den Vorteil, dass sie eindeutig definiert sind und ihre Messbarkeit bereits gewährleistet oder leicht herbeizuführen ist.

Es ist zu hoffen, dass der Leser die hier dargestellten Methoden in seiner Forschung erfolgreich einsetzen kann.

# Anhang I

Der **Determinationskoeffizient** wird berechnet als

$$D = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2},$$

wo $y_i$ die gemessenen Werte der Variablen sind ($i$ = 1,2,…,$n$), $\hat{y}_i$ sind die theoretischen (berechneten) Werte und $\overline{y}$ ist der Durchschnitt aller $y_i$-Werte. Der Zähler enthält die Summe der Abweichungsquadrate der beobachteten von den berechneten Werten, d.h. die „nichterklärten“ Abweichungen, der Nenner enthält die Summe der Quadrate aller Abweichungen. D drückt daher den Anteil der durch die angepasste Kurve erklärten Abweichungen. Je größer D, das im Intervall <0, 1> liegt, desto besser die Anpassung.

# Anhang II

**Die Methode der kleinsten Quadrate**

Einfachkeitshalber nur für eine Gerade (viele Kurven lassen sich auf eine Gerade trasformieren). Sei $y = a + bx$ eine Gerade. Man stellt auf die Gleichung

$$W = \sum_i (y_i - a - bx_i)^2$$

und berechnet die Parameter $a$ und $b$ so, dass $W$ minimal ist. Dies lässt sich durch Ableitungen bewerkstelligen, nämlich

$$\frac{\partial W}{\partial a} = 2\sum (y - a - bx)(-1) = 0 \qquad \text{(den Laufindex kann man auslassen)}$$

$$\frac{\partial W}{\partial b} = 2\sum (y - a - bx)(-x) = 0.$$

daraus ergeben sich nach Ausmultiplikation die Gleichungen

$$-\sum y + na + b\sum x = 0$$
$$-\sum xy + a\sum x + b\sum x^2 = 0$$

deren Lösung für $a$ und $b$ die Schätzer

$$\hat{a} = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}, \qquad \hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

ergibt.

## Literatur


**Albert, R., Barabási, A.-L.** (2002). Statistical mechanics of complex networks. *Review of Modern Physics 74, 47-97.*

**Altmann, G.** (1963). Phonic structure of Malay pantun. *Archiv orientální 31, 274-286.*

**Altmann, G**. (1978). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.*

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

**Altmann, G.** (1987). Tendenzielle Vokalharmonie. *Glottometrika 8, 104-112.*

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

**Altmann, G.** (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 287-298.*

**Altmann, G.** (1993). Phoneme counts. *Glottomerika 14, 54-68.*

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

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

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

**Altmann, G., Štukovský, R.** (1963). Analýza náhleho klimaxu. *Litteraria 6, 62-72.*

**Altmann, G., Štukovský, R.** (1965). The climax in Malay pantun. *Asian and African Studies 1, 13-20.*

**Altmann-Fitter** (1997). *Iterative fitting of probability distributions*. Lüdenscheid: RAM-Verlag (Software).

**Antić, G., Kelih, E., Grzybek, P.** (2002). Conference Report. „Word Length in Texts. An International Symposium on Quantitative Text Analysis." Institute for Slavic Sudies, Graz University, June 21-23, 2002. *J. of Quantitative Linguistics 9, 275-279.*

**Antosch, F.** (1969). The diagnosis of literary style with the verb-adjective ratio**.** In: Doležel, L., Bailey, R.W. (eds), *Statistics and Style: 57-65.* New York: Elsevier.

**Arapov, M.V.** (1981). Sistemnyj analiz leksičeskoj struktury tekstov. Sistemnye issledovanija. *Ežegodnik 1980, 372-403.*

**Baayen, R.H.** (2001). *Word frequency distributions.* Dordrecht: Kluwer.

**Bak, P.** (1999). *How nature works*. New York: Copernicus-Springer-Verlag.

**Balakrishnan, V.K.** (1997). *Graph theory*. New York: McGraw-Hill.

**Barton, D.E., David, F.N.** (1957). Multiple runs. *Biometrika 44, 168-178.*

**Best, K.-H.** (1994). Word class frequencies in contemporary German short prose texts. *J. of Quantitative Linguistics 1, 144-147.*

**Best, K.-H.** (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. *Glottometrika 16, 276-285.*

**Best, K.-H. (ed.)** (1997). *Glottometrika 16. The distribution of word and sentence length.* Trier: WVT.

**Best, K.-H.** (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4, 37-51*.

**Best, K.-H.** (2001). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt.

**Best, K.-H.** (2001a). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26.*

**Best, K.-H.** (2003). *Quantitative Linguistik. Eine Annäherung*. Göttingen: Peust & Gutschmidt.

**Bock, H.H.** (1974). *Automatische Klassifikation*. Göttingen: Vandenhoek & Rupprecht.

**Boder, D.P.** (1940). The adjective-verb quotient: a contribution to the psychology of language. *Psychological Revue 3, 309-343.*

**Boroda, M.G., Zörnig, P.** (1990). Zipf-Mandelbrot´s law in a coherent text: towards the problem of validity. *Glottometrika 12, 41-60*

**Bortz, J., Lienert, G.A., Boehnke, K.** (1990). *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.

**Bradley, J.V.** (1968). *Distribution-free statistical tests*. Englewood Cliffs: Prentice Hall.

**Brainerd, B.** (1972). On the relation between types and tokens in literary texts. *Journal of Applied Probability 9, 507-518.*

**Brainerd, B.** (1982). On the relation between the type-token and the species-area problem. *J. of Applied Probability 19, 785-793.*

**Brunet, E.** (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution.* Genéve: Slatkine.

**Busemann, A.** (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena.

**Carroll, J.B.** (1969). Vectors of prose style. In: Doležel, L., Bailey, R. (Hrsg.), *Statistics and style: 147-155.* New York: Elsevier.

**Chitashvili, R.J., Baayen, R.H.** (1993). Word frequency distributions of texts and corpora as large number of rare event distributions. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis 54-135*. Trier: WVT.

**Cochran, W.G.** (1954). Some methods for the strengthening of the common $\chi^2$-tests. *Biometrics 10, 417-451.*

**Cohen, A., Mantegna, R.N., Havlin, S.** (1997). Numerical analysis of word frequentcies in artificial and natural language texts. *Fractals 6(1), 95-104.*

**Condon, E.U.** (1928). Statistics of vocabulary. *Science 67, 300.*

**Cox, D.R.** (1958). The regression analysis of binary sequences. *J. of the Royal Statistical Society B 20, 215-232.*

**Cox, D.R., Stuart, A.** (1955). Some quick sign tests for trend in location and dispersion. *Biometrika 42, 80-95.*

**Daneš, F.** (1970). Zur linguistischen Analyse der Textstruktur. *Folia Linguistica 4, 72-78.*

**Dorogovtsev, S.N., Mendes, J.F.F.** (2002). Evolution of networks. *Advances in Physics 51, 1079-1187.*

**Drobisch, M.V.** (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft, Philosophisch-historische Klasse 18, 73-139.*

**Drobisch, M.V.** (1868a). Über die Formen des Hexameters bei Klopstock, Voss und Goethe. *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft, Philosophisch-historische Klasse 20, 138-160.*

**Drobisch, M.V.** (1868b). Weitere Untersuchungen über die Formen des Hexameters des Vergil, Horaz und Homer. *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft, Philosophisch-historische Klasse 20, 16-53.*

**Dugast, D.** (1978). Sur quoi se fonde la notion d´étendue théorique du vocabulaire? *Le français modern 1, 25-32.*

**Dugast, D.** (1979). *Vocabulaire et stylistique. 1. Théâtre et dialogue*. Genève: Slatkine.

**Dugast, D.** (1979a). *Vocabulaire et discours. Essai de léxicométrie quantitative.* Genéve: Slatkine

**Dugast, D.** (1980). La mesure de la richesse lexicale: une esquisse historique. *Verbum 3, 115-134.*

**Eddington, E.S.** (1961). Probability table for the number of runs of sign for first difference in ordered series. *Journal of the American Statistical Association 56, 156-159.*

**Ejiri, K., Smith, A.** (1993). Proposal of a new constraint measure for text. In: Köhler, R., Rieger, B. (eds.), *Contributions to quantitative linguistics: 195-211*. Dordrecht: Kluwer.

**Estoup, J.B.** (1916). *Les Gammes Sténographiques*. Paris, Institut Sténographique

**Fischer, H.** (1969). Entwicklung und Beurteilung des Stils. In: Kreuzer, H., Gunzenhäuser, R. (Hrsg.), *Mathematik und Dichtung: 171-183.* München: Nymphenburger Verlag.

**Fucks, W.** (1955a). Theorie der Wortbildung. *Mathematisch-Physikalische Semesterberichte. Bd. 4, 195-212.*

**Fucks, W.** (1955b). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln - Opladen, Westdeutscher Verlag.

**Fucks, W.** (1956). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Forschungsberichte. Bd. 3, 7-21.*

**Gani, J.** (1975). Stochastic models for type counts in a literary text. In**:** Gani, Joseph (ed.): *Perspectives in probability and statistics. Papers in honour of M.S. Bartlett.* London: Academic Press.

**Glottometrics 3, 4, 5** (2002). *To honor G.K. Zipf.* Lüdenscheid: RAM-Verlag.

**Gottman, J.M., Roy, A.K.** (1990). *Sequential analysis. A guide for behavioral researchers*. Cambridge: Cambridge University Press.

**Greenberg,** J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

**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.*

**Grotjahn, R., Altmann, G.** (1993). Modelling the distribution of word length: some methodological problems. In. Köhler, R., Rieger, B. B. (eds.), *Contributions to quantitative linguistics: 141-153.* Dordrecht u.a.: Kluwer.

**Grzybek, P. (ed.)** (2005). *Word length studies and related issues.* Boston/ Dordrecht: Kluwer.

**Guiter, H., Arapov, M.V.** (Hrsg.) (1982). *Studies on Zipf´s law*. Bochum: Brockmeyer

**Gumenyuk, A.**, **Kostyshin, A., Borisov, K., Salnikova, O.** (2004). On the acoustic elements of a poem and on the formal procedures of their segmentation. *Glottometrics 8, 42-67*

**Haight, F.A., Jones, R.B.** (1974). A probabilistic treatment of qualitative data with special reference to word association tests. *J. of Mathematical Psychology 11, 237-244*.

**Haitun, S.D.** (1983). *Naukometrika – sostojanie i perspektivy*. Moskva: Nauka.

**Hájek, J.** (1969). *Nonparametric statistics*. San Franciso: Holden-Day.

**Hammerl, R.** (1990). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

**Hamp, E.** (1958). The calculation of parameters of morphological complexity. *Proceedings of the Eight International Congress of Linguists, Oslo: 134-142*.

**Herdan, G**. (1964). *Quantitative linguistics.* London: Butterworths.

**Herdan, G.** (1966). *The advanced theory of language as choice and chance.* Berlin, Springer.

**Hřebíček, L.** (1993). Text as a strategic process. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 136-150*. Trier: WVT.

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

**Hřebíček, L.** (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law*. Trier: WVT.

**Hřebíček, L.** (1996). Word frequency and word location in a text. *Archiv orientální 64, 339-347.*

**Hřebíček, L.** (1997). *Lectures on text theory*. Prague: Oriental Institute.

**Hřebíček, L.** (1997a). Persistence and other aspects of sentence-length series. *J. of Quantitative Linguistics 4, 103-109.*

**Hřebíček, L.** (1997b). Hurst´s indicator and text. In: Altmann, G., Koch, W.A. (eds.), *Systems. New paradigms for the human sciences: 572-588*. Berlin: de Gruyter.

**Hřebíček, L.** (2000). *Variation in sequences*. Prague: Oriental Institute.
**Jakubaitis, T.O.** (1981). *Časti reči i tipy tekstov.* Riga: Zinatne.
**Judt, B.** (1995). *Wortartenhäufigkeiten im Deutschen und Französischen.* Staatsexamensarbeit, Göttingen.
**Kelemen, J.** (1970). Sprachtypologie und Sprachstatistik. In: Dezsö, L., Hajdú, P. (eds.), *Theoretical problems of typology and the northern Eurasian languages*: 53-63. Amsterdam.
**Kendall, M.G., Stuart, A.** (1967). *The advanced theory of statistics*. London: Griffin.
**Klein, H.** (1999). *INTEXT*. Version 4.1.
**Kločkova, E.A.** (1968). O raspredelenii klassov slov v nekotorych funkcional´nych stiljach russkogo jazyka. *Voprosy jazykoznanija: Saratov, SGU, 109-118.*
**Köhler, R.** (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
**Köhler, R.** (1995). *Bibliography of Quantitative Linguistics*. Amsterdam: Benjamins.
**Köhler, R.** (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6, 46-57.*
**Köhler, R.** (Hrsg.) (2002). Korpuslinguistische Untersuchungen zur quantitativen und synergetischen Linguistik. http://ubt.opus.hbz-nrw.de/volltexte/2004/
**Köhler, R.**, **Altmann, G.** (1999). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7, 189-200.*
**Köhler, R.**, **Martináková-Rendeková, Z.** (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W.A. (eds.), *Systems. New paradigms for the human sciences: 514-546.* Berlin: de Gruyter.
**Krappe, A.H.** (1921). *Alliteration in the Chanson de Roland and in the Carmen de Prodicione Guenosis.* Iowa City, Iowa.
**Krupa, V.** (1965). On quantification of typology. *Linguistics 12, 31-36.*
**Lánský, R.**, **Radil-Weiss, T.** (1980). A generalization of the Yule-Simon model, with spezial reference to word association tests and neural cell assembly formation. *J. of Mathematical Psychology 21, 53-65.*
**Levickij, V.V., Hikow, L.** (2004). Zum Gebrauch der Wortarten im Autorenstil. *Glottometrics 8, 12-22.*
**Maas, H.-D.** (1972). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zs. für Literaturwissenschaft und Linguistik 8, 73-96.*
**Mann, H.B., Whitney, D.R.** (1947). On a test whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics 18, 50-60.*
**Maxwell, A.E.** (1961). *Analysing qualitative data*. London: Methuen.
**McIntosh, R.P.** (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology 48, 392-404.*

**Mood, A.M.** (1940). Distribution theory of runs. *The Annals of Mathematical Statistics 11, 367-392.*

**Müller, W.** (1971). Wortschatzumfang und Textlänge. Eine kleine Studie zu einem vielbehandelten Problem. *Muttersprache 81, 266-276.*

**Nešitoj, V.V.** (1975). Dlina teksta i ob'em slovarja. Pokazateli leksičeskogo bogatstva teksta. In: *Metody izučenija leksiki: 110-118.* Minsk, BGU.

**Niehaus, B.** (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. *Glottometrika 16, 213-275.*

**Orlov, Ju.*K*., Boroda, M.G. Nadarejšvili, I.Š.** (1982). *Text, Sprache, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

**Pawlowski, A**. (1994). Ein Problem der klassischen Stilforschung: Die Stabilität einiger Indikatoren des Lexikonumfangs (anhand stilistisch homogener Stichproben unterschiedlicher Länge*). ZeT – Zeitschrift für empirische Sprachwissesnchaft 1, 67-72.*

**Ratkowsky, D.A., Halstead, M.H., Hantrais, L.** (1980). Measuring vocabulary richness in literary works: A new proposal and re-assessment of some earlier measures. *Glottometrika 2, 125-147.*

**Robins, R.H.** (1965). Some typological observation on Sundanese morphology. *Lingua 15, 435-450.*

**Schlissmann, A.** (1948). Sprach- und Stilanalyse mit einem vereinfachten Aktionsquotienten. *Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik 2.*

**Schmidt, P. (ed.)** (1996). *Glottometrika 15. Issues in general linguistic theory and the theory of word length.* Trier: WVT.

**Schroeder, M.** (1991). *Fractals, chaos, power laws: minutes from an infinite paradise*. New York: Freeman.

**Schulz, K.P., Altmann, G.** (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrika 9, 1-48.*

**Schweers, A., Zhu, J.** (1991). Wortartenklasssfizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (Hrsg.), *Diversification processes in language: grammar: 157-165.* Hagen: Rottmann Medienverlag.

**Shewan, A.** (1925). Alliteration and assonance in Homer. *Classical Philology 20, 193-210.*

**Sichel, H.S.** (1986). Word frequency distributions and type-token characteristics. *The Mathematical Scientist 11, 45-72.*

**Simon, H.A.** (1995). On a class of skew distribution functions. *Biometrika 42, 425-440.*

**Skalička, V.** (1935*). Zur ungarischen Grammatik*. Prague.

**Skinner, B.F**. (1939). The alliteration in Shakespeare´s sonnets. A study in literary behavior. *Psychological Record 3, 186-192.*

**Skorochod´ko, E.F.** (1981). *Semantische Relationen in der Lexik und in Texten.* Bochum: Brockmeyer.

**Slavíčková, E.** (1968). Towards a typological evaluation of related languages. *Travaux Linguistiques de Prague 7, 73-100.*

**Somers, H.H.** (1959). *Analyse mathématiquwe du langage: Lois générales et mesures statistiques*. Louvain : Nauwelaerts.

**Somers, H.H.** (1966). Statistical methods in literary analysis. In: Leed, J. (ed.), *The computer and the literary style: 128-140*. Kent, Ohio.

**Strauss, U., Sappok, Ch., Diller, J.H., Altmann, G.** (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7, 73-100.*

**Štukovský, R., Altmann, G.** (1965). Vývoj otvoreného rýmu v slovenskej poézii. *Litteraria 8, 156-161.*

**Štukovský, R., Altmann, G.** (1966). Die Entwicklung des slowakischen Reims im XIX. und XX. Jahrhundert. In: *Teorie verše I, 258-261.*

**Thorndike, E.L.** (1943). Derivation ratios. *Language 19, 27-37.*

**Tiščenko, W.** (1970). Častota častyn movy v riznych funkcionalnych styljach sučasnojiukrainkoji movy. In: Perebyjnis, V.S., Muravycka, M.P. (Hrsg.), *Pytannja strukturnoji leksykolohii: 215-224.* Kyjiv: Naukova dumka.

**Tuldava, J.** (1995*). Methods in quantitative linguistics*. Trier: WVT.

**Tuldava, J.** (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: WVT.

**Tweedie, F.J., Baayen, R.H.** (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32, 323-352.*

**Uhlířová, L., Wimmer, G**. (2003). A contribution to word length theory. In: Kempgen, S., Schweier, U., Berger, T. (eds.*), Festschrift für Werner Lehfeldt zum 60. Geburtstag: 524-530*. München: Sagner.

**Viehweger, D**. (1978). Struktur und Funktion nominativer Ketten im Text. *Studia Grammatica XVII, 149-169*. Berlin: Akademie-Verlag.

**Wallis, W.A., Moore, G.H.** (1941). A significance test for time series analysis. *J. of the American Statistical Association 20, 257-267.*

**Wimmer, G., Altmann, G.** (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133.* Trier: WVT.

**Wimmer, G., Altmann, G**. (1999). Review article: On vocabulary richness. *J. of Quantitative Linguistics 6, 1-9.*

**Wimmer, G. Altmann, G.** (1999a). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

**Wimmer, G**., **Altmann, G.** (2001). Some statistical investigations concerning word classes. *Glottometrics 1, 109-123.*

**Wimmer, G., Altmann, G.** (2005). Unified derivation of some linguistic laws. In: Grzybek, P. (Hrsg.), *Word length studies and related issues: 307-316*. Boston/ Dordrecht: Kluwer.

**Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S.** (2003). *Úvod do analýzy textov.* Bratislava: Veda.

**Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G.** (1994). Towards a theory of word length distribution. *J. of Quantitative Linguistics 1, 98-106.*

**Wimmer, G., Šidlík, P., Altmann, G**. (1999). A new model of rank-frequency distribution. *J. of Quantitative Linguistics 6, 188-193.*

**Wimmer, G., Witkovský, V., Altmann, G.** (1999). Modification of probability distributions applied to word length research. *J. of Quantitative Linguistics 6, 257-268.*

**Ziegler, A.** (1998). Word class frequencies in Brazilian-Portuguese press texts. *J. of Quantitative Linguistics 5, 269-280.*

**Ziegler, A.** (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L. et al. (Hrsg.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in Honor of Luděk Hřebíček: 295-312.* Trier: WVT.

**Ziegler, A.**, **Altmann, G.** (2002). *Denotative Textanalyse*. Wien: Praesens.

**Ziegler, A.**, **Altmann, G.** (2003). Text stratification. *J. of Quantitative Linguistics 10, 275-292.*

**Ziegler, A., Best, K.-H., Altmann, G.** (2001). Nominalstil. *ETC-Empirische Text- und Kulturforschung 2, 72-85.*

**Ziegler, A., Best, K.-H., Altmann, G.** (2001a). A contribution to text spectra. *Glottometrics 1, 97-108.*

**Ziegler, A., Jüngling, R., Altmann, G.** (2005). Latente konnotative Textstruktur. *Festschrift B. Rieger (im Druck)*

**Zörnig, P.** (1984). The distribution of distances between like elements in a sequence I. *Glottometrika 7, 1-15;* II. *Glottometrika 7, 1-14.*

**Zörnig, P.** (1987). A theory of distance between like elements in a sequence. *Glottometrika 8, 1-22.*

**Zörnig, P., Altmann, G.** (1983). The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 5, 205-211.*

**Zörnig, P., Altmann, G.** (1984). The entropy of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 6, 41-47.*

**Zörnig, P., Boroda, M.** (1992). The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent texts. *Glottometrika 13, 205-218.*

# Namensregister

# Sachregister