Reinhard Köhler (Hg.)
Korpuslinguistische Untersuchungen
zur quantitativen und systemtheoretischen Linguistik
2002, S. 127 - 177

Untersuchungen zur chinesischen Sprache und Schrift

Hartmut Bohn

1. Einleitung

Die quantitativen Eigenschaften der chinesischen Sprache und Schrift sind bisher weitgehend unerforscht. Dabei findet man in der chinesischen Sprache und insbesondere in der chinesischen Schrift eine Reihe von im Vergleich zu anderen Sprachen extremen Merkmalsausprägungen, die eine nähere Untersuchung gerade aus quantitativ-linguistischer Perspektive besonders lohnenswert erscheinen lassen.

Im Folgenden sollen verschiedene Aspekte der modernen chinesischen Sprache und Schrift unter quantitativen Gesichtspunkten betrachtet werden. Grundlage aller Untersuchungen bilden Sprachdaten aus der modernen chinesischen Hochsprache, wie sie heute in der Volksrepublik China verwendet wird. Die Untersuchungen zur modernen chinesischen Schrift beziehen sich dabei auf die auf dem chinesischen Festland gebräuchlichen vereinfachten Schriftzeichen.

Den Schwerpunkt der Arbeit bilden Untersuchungen zum Menzerathschen Gesetz. Hier werden die Konstrukt-Konstituentenbeziehungen auf den verschiedenen Ebenen der chinesischen Schrift untersucht. Zusätzlich werden die Komplexitäts- und Längenverteilungen der Elemente auf den untersuchten Ebenen genauer erörtert.

Anhand eines Einzeltextes soll anschließend die Rang-Frequenzverteilung von Wörtern in abgeschlossenen Texten nach Zipf bzw. Zipf-Mandelbrot im Chinesischen überprüft werden.

Schließlich soll ein vermuteter Zusammenhang zwischen Schriftzeichenfrequenz und dem Vorhandensein phonetisch wirksamer Schriftzeichenkomponenten nachgewiesen werden.

2. Das Menzerathsche Gesetz und die chinesische Schrift

Das Menzerathsche Gesetz wurde inzwischen an einer Vielzahl typologisch unterschiedlicher Sprachen und auf verschiedenen sprachlichen Ebenen überprüft (siehe z.B. Prün, 1994:151).

Bisher liegt jedoch keine Untersuchung zur chinesischen Sprache vor. Dabei stellt das Chinesische - und hier vor allem die chinesische Schrift - ein besonders

lohnenswertes Untersuchungsobjekt zur Überprüfung des Menzerath-schen Gesetzes dar. Zunächst verkörpern die chinesischen Schriftzeichen, wie schon Prün (1994:149) feststellt, ein ideales Beispiel für Informationsverarbeitung auf begrenztem Raum, da jedes Schriftzeichen, ungeachtet seiner Komplexität, denselben Raum einnimmt. Doch auch die die Schriftzeichen konstituierenden Elemente folgen ihrerseits vermutlich der von Menzerath beobachteten Regel. Die Besonderheiten der chinesischen Wortbildung lassen darüber hinaus eine Untersuchung der Wortebene und höherer Ebenen interessant erscheinen.

In den folgenden Abschnitten sollen daher verschiedene Ebenen der modernen chinesischen Schrift - ausgehend von den Einzelstrichen bis hin zur Satzebene - zur Validierung des Menzerathschen Gesetzes im Chinesischen herangezogen werden.

Als Grundlage der Untersuchungen zum Menzerathschen Gesetz im Chinesischen wird auf der graphischen Ebene der chinesischen Schrift folgende Hierarchie angenommen:

- Strichebene
- Komponentenebene
- Schriftzeichenebene
- Wortebene
- Teilsatzebene
- Satzebene

Nun lassen sich die Arbeitshypothesen für die zu untersuchenden Ebenen gemäß dem Menzerathschen Gesetz wie folgt formulieren (vgl. auch Altmann, 1980:8ff.; Altmann & Schwibbe, 1989:8ff.):

Hypothese 1 (Komponentenebene): *Je komplexer eine Komponente, gemessen in der Anzahl der Einzelstriche, desto einfacher die Striche.*

Hypothese 2 (Schriftzeichenebene): *Je komplexer ein Schriftzeichen, gemessen in der Zahl seiner Komponenten, desto einfacher die Komponenten, gemessen in der Zahl ihrer Striche.*

Hypothese 3 (Wortebene): *Je länger ein Wort, gemessen in der Anzahl der Schriftzeichen, desto einfacher die Schriftzeichen, gemessen in der Zahl ihrer Komponenten.*

Hypothese 4 (Teilsatzebene): *Je länger ein Teilsatz, gemessen in der Anzahl der Wörter, desto kürzer die Wörter, gemessen in der Zahl der Schriftzeichen.*

Hypothese 5 (Satzebene): *Je länger ein Satz, gemessen in der Anzahl der Teilsätze, desto kürzer die Teilsätze, gemessen in der Zahl der Wörter.*

Diese fünf Arbeitshypothesen sollen in den folgenden Abschnitten operationalisiert und anhand von Sprachdaten empirisch überprüft und bewertet werden.

2.2 Das Menzerathsche Gesetz auf Komponentenebene

Bevor die Hypothese 1 überprüft werden kann, müssen die Begriffe Strich und Komponente geklärt werden.

Mit Strich ist die minimale graphische Einheit der chinesischen Schrift gemeint, wie sie etwa zur Klassifizierung in Strichzahlindizes chinesischer Nachschlagewerke verwendet wird.

Unter Komponenten sollen im Folgenden jene Elemente der chinesischen Schrift verstanden werden, die - zwischen Schriftzeichen- und Einzelstrichebene liegend - die einzelnen Schriftzeichen konstituieren.

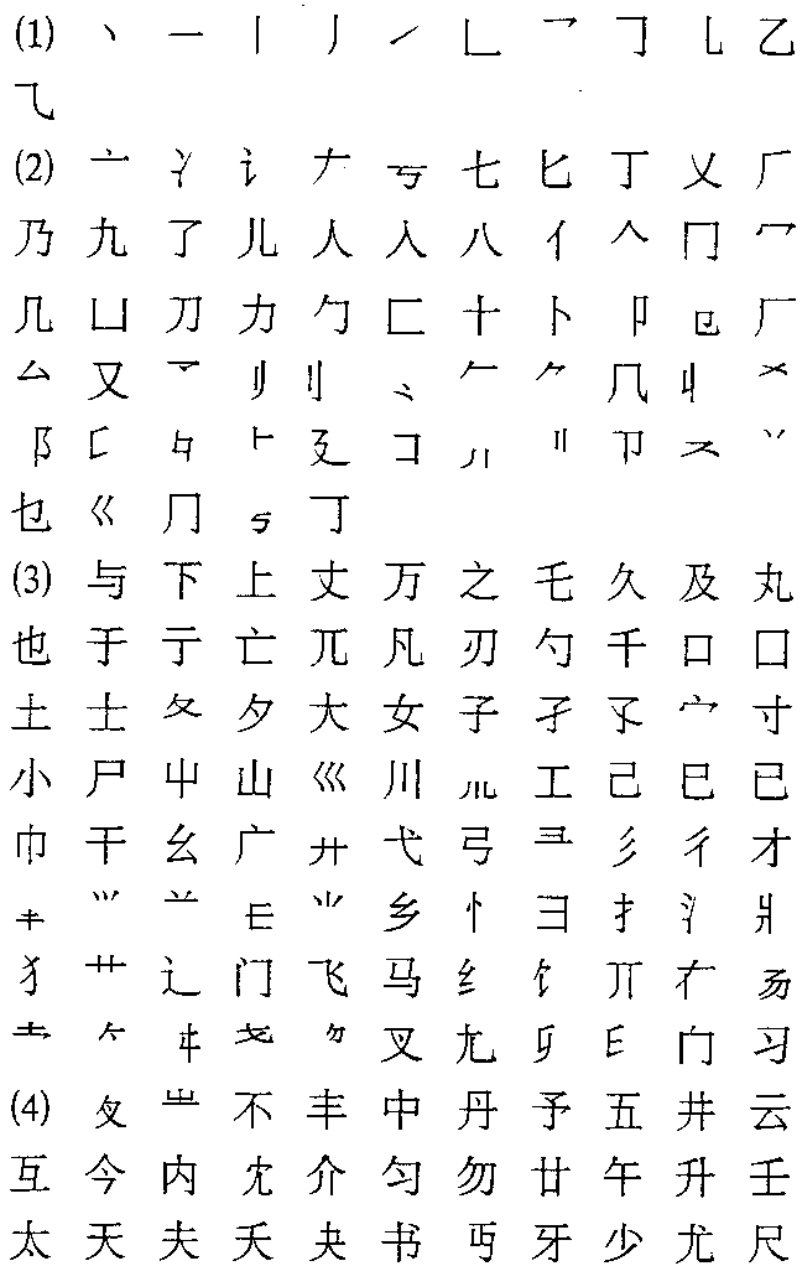
Die Komponentenebene ist nicht problemlos von der Strichebene einerseits und der Schriftzeichenebene andererseits abzugrenzen. Einige Komponenten bestehen aus nur einem einzelnen Strich. Viele andere wiederum finden auch als eigenständiges Schriftzeichen Verwendung. Hier wird die empirische Überprüfung des Menzerathschen Gesetzes Aufschluss darüber geben können, ob es gerechtfertigt ist, von der Komponentenebene als linguistisch relevanter Ebene des chinesischen Schriftsystems zu sprechen.

2.2.1 Operationalisierung

Benötigt wird zum ersten ein Inventar der in modernen chinesischen Schriftzeichen vorkommenden Komponenten. Zum zweiten ist ein Strichinventar, d.h. eine Liste von in chinesischen Schriftzeichen vorkommenden Strichtypen erforderlich. Zusätzlich ist ein Aufwandsmaß zu entwickeln, mit dem es möglich ist, den einzelnen Strichtypen Werte für ihre ‚Einfachheit' zuzuordnen.

Ausgehend von der Liste der 485 Kanjigrapheme von Stalph (1989:73f.) wurde ein Komponenteninventar erarbeitet. Stalphs Liste wurde systematisch mit den 6.763 Schriftzeichen des volksrepublikanischen Computerstandards GB 2312-80, wie er im HSZ (1988) dokumentiert ist, abgeglichen. Die Minimalpaaranalyse wurde jedoch in einigen Fällen weiter geführt, als dies bei Stalph der Fall ist. Der Schriftzeichenvergleich fand ausschließlich auf der graphischen Ebene statt, wobei die Druckformen im HSZ (1988) maßgebend waren. Zeichenetymologische Erwägungen wurden bei der Analyse nicht berücksichtigt.

Das Ergebnis dieser Graphemanalyse wird in Abbildung 2.1 wiedergegeben. Der vorläufige Charakter der Analyse muss an dieser Stelle ausdrücklich betont werden: wie auch bei Stalphs Grapheminventar bleiben eine Reihe von Inkonsistenzen bestehen.


(1)
(2)
(3)
(4)

尹 屯 巴 心 戈 户 手 攵 文 文 斗
斤 方 日 曰 月 木 欠 止 歹 毌 毛
氏 气 水 火 灬 ⺕ 朩 毋 父 片 牛
犬 王 开 廿 圭 牛 辵 丷 ⺧ 仒 尸
小 罒 礻 耂 皿 ⺁ 丹 户 白 ⺟ 丰
乌 以 巨 旡 瓦 内 车 长 为 贝 办
月 木 专 卄 正 乌 爿 市 无 丐 见
韦 乒 丑 爪 区 办
(5) 世 且 目 丙 丘 主 乎 乍 击 出
凸 凹 北 半 史 四 央 失 平 弗 必
戊 戉 斥 本 末 未 母 民 永 丞 玉
甘 生 用 甪 田 甲 申 由 疒 癶 白
皮 皿 矛 矢 石 示 禾 立 罒 业 兴
阝 串 自 夫 六 电 术 氺 礻 册 正
甩 丝 丰 冎 东 鸟 尔 龙 瓜 钅 冊
戋 匆 头 艮 夬 东 乐 令 丰 手 产
冉 阝 禾
(6) 豕 争 亦 亥 兆 吏 州 戍 曲 束
朱 竹 米 糸 缶 羊 而 耒 耳 聿 肉
自 至 臼 舟 艮 虍 虫 血 西 自 亚
关 兴 产 产 类 夹 豸 臣 页 伞 关
声 襾 处 戋 羊 良 与 夷 肙 曳 血
农 兴 告 单 网 甚 产 且

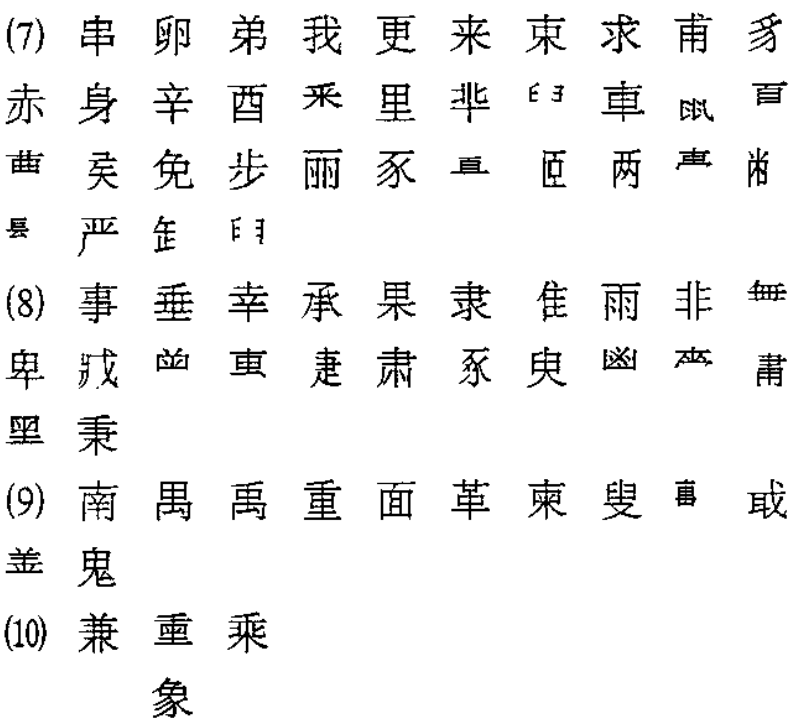


Abb. 2.1: Das Komponenteninventar, nach Strichzahl geordnet

Abbildung 2.2 zeigt das der Untersuchung zugrunde liegende Einzelstrichinventar sowie das jedem Strich zugeordnete Aufwandsmaß. Als Maß für den Strichaufwand wurde die Anzahl der Schreibrichtungsänderungen innerhalb des jeweiligen Striches gewählt.

Für jede Komponente des Inventars wurde die Zahl ihrer Striche sowie die Aufwandsmaße nach Abbildung 2.2 erfasst. Auf dieser Grundlage konnte nun für jede Komponente der durchschnittliche Schreibaufwand als arithmetisches Mittel der Aufwandsmaße ihrer einzelnen Striche errechnet werden. Im zweiten Schritt wurde das arithmetische Mittel des Schreibaufwands aller Komponenten gleicher Strichzahl berechnet.

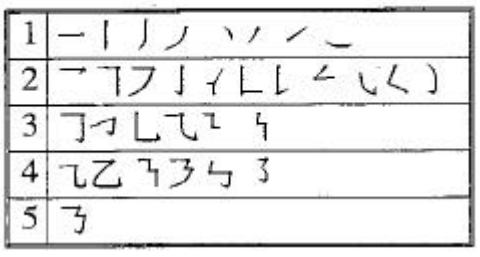


Abb. 2.2: Das Strichinventar, nach Aufwand geordnet

2.2.2 Ergebnis und Bewertung

Die Ergebnisse der Berechnung sind in Tabelle 1 im Anhang aufgeführt; Abbildung 2.3 zeigt die graphische Darstellung der empirischen Daten. Allein der optische Eindruck bestätigt die Grundaussage von Hypothese 1.

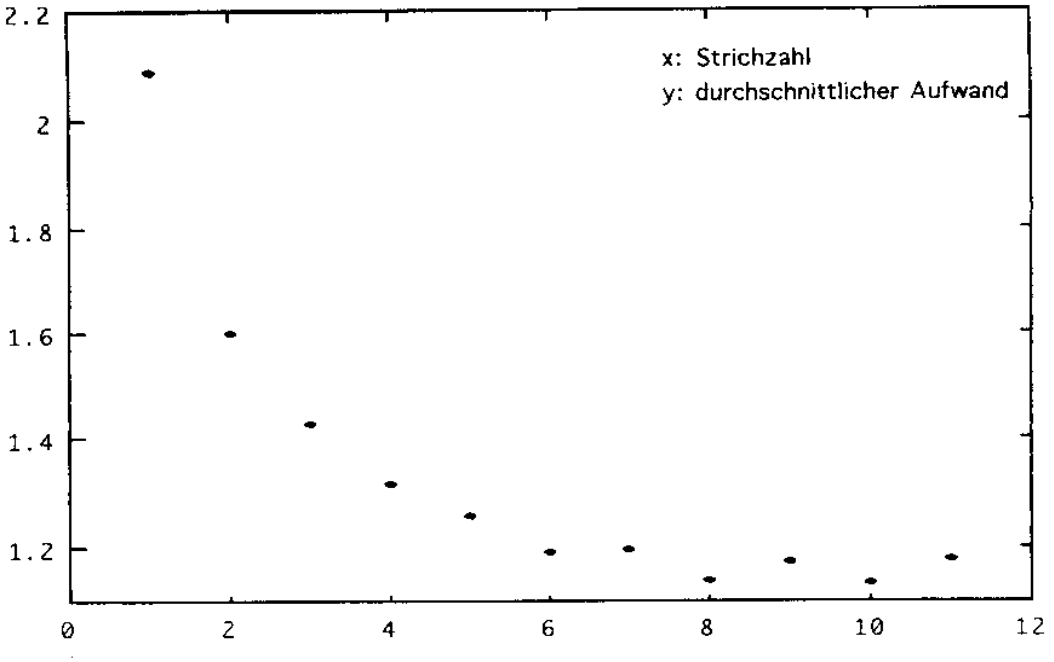


Abb. 2.3: Das Menzerathsche Gesetz auf Komponentenebene - empirische Werte

Zur statistischen Überprüfung des Menzerathschen Gesetzes wurde die Funktionsgleichung $y = ax^b$ (b < 0) an die empirischen Daten angepasst. Um Verzerrungen des Ergebnisses aufgrund zu weniger Ausgangsdaten zu vermeiden, flossen in die Anpassung jedoch nur die empirischen Werte ein, in deren Berechnung zehn oder mehr Belege eingegangen waren. Als Maß der Güte der Anpassung wurde der Determinationskoeffizient verwendet. Die Anpassung ergab für die Parameter folgende Werte: a = 2.0184; b = -0.2831 bei einem Determinationskoeffizienten von D = 0.9673.

Die den empirischen Werten entsprechenden, aus den Anpassungsparametern errechneten theoretischen Werte sind in Tabelle 1 im Anhang aufgeführt. Abbildung 2.4 stellt den empirisch ermittelten Datenpunkten die theoretische Kurve $y = 2.0184\ x^{-0.2831}$ gegenüber.

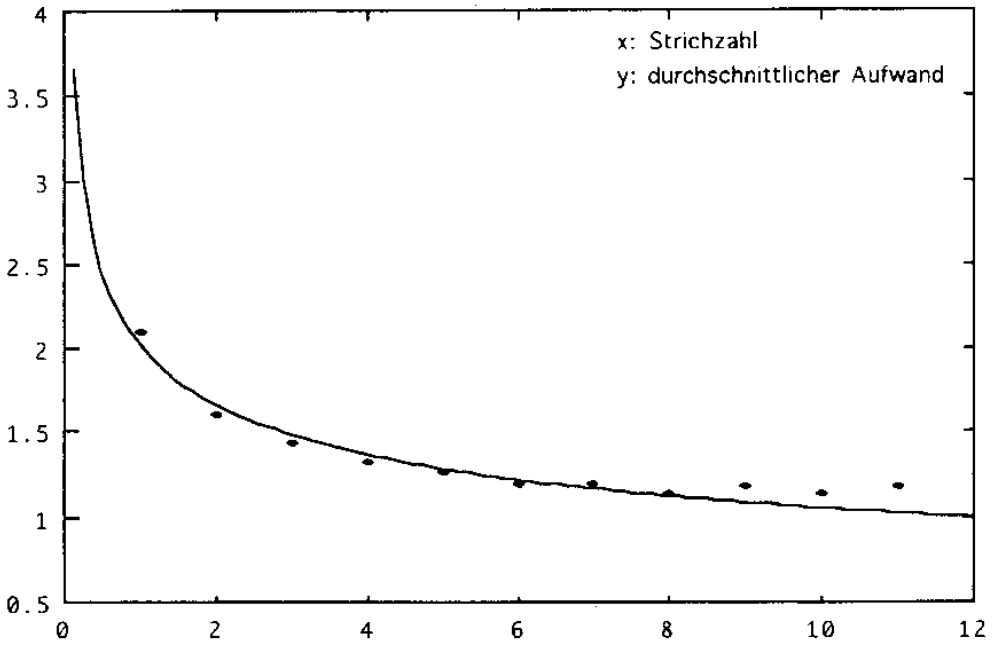


Abb. 2.4: Das Menzerathsche Gesetz auf Komponentenebene - Gegenüberstellung der empirischen Werte und der theoretischen Kurve

Auf der Basis des Determinationskoeffizienten kann die Gültigkeit des Menzerathschen Gesetzes im Falle der Komponenten der chinesischen Schrift angenommen werden.

Darüber hinaus erlaubt dieses Ergebnis den Schluss, dass es sich bei der ‚Komponentenebene' um eine linguistisch relevante Ebene der chinesischen Schrift handelt.

2.3 Das Menzerathsche Gesetz auf Schriftzeichenebene

Hypothese 2 zur Schriftzeichenkomplexität wurde von Prün (1994) für die in Japan gebräuchlichen chinesischen Schriftzeichen (Kanji) überprüft und bestätigt. Grundlage dieser Untersuchung bildeten die 1.945 Joyokanji - die offizielle Liste ‚allgemein gebräuchlicher Kanji' - sowie die Liste der 485 Kanjigrapheme von Stalph (1989).

Im Folgenden soll diese Hypothese auf Grundlage einer umfangreicheren Datenmenge auch an der modernen chinesischen Schrift überprüft werden.

2.3.1 Operationalisierung

Zur Durchführung der Untersuchung ist eine für die moderne chinesische Schriftsprache repräsentative Schriftzeichenliste sowie ein Komponenteninventar erforderlich.

Als Schriftzeicheninventar wurden die Schriftzeichen des Computerstandards der Volksrepublik China, GB 2312-80, gewählt. Der Standard enthält 6.763 Zeichen einschließlich einiger Zeichenkomponenten, die nicht selbständig verwendet werden. Von den 6.763 Zeichen des Computerstandards werden nur diejenigen ausgewählt, denen in HSZ (1988) eine Aussprache zugeordnet ist, d.h. Zeichen, die im Gegensatz zu den im Standard enthaltenen Komponenten selbständig verwendet werden können. Somit verblieben für die Untersuchung noch 6.724 Zeichen.

Als Komponenteninventar diente das in Abbildung 2.1 dargestellte Inventar, das auf der Grundlage der Zeichen des für die Untersuchung verwendeten Computerstandards erarbeitet wurde.

Ausgehend von Schriftzeichen- und Komponenteninventar wurde zunächst jedes Zeichen in seine Komponenten zerlegt. Somit konnte für jedes Zeichen die Anzahl seiner Komponenten bestimmt werden. Über den Abgleich mit der Komponentenliste konnte für jedes Zeichen die Strichzahlen seiner Komponenten ermittelt und seine durchschnittliche Komponenten-Strichzahl errechnet werden. Daraufhin wurde das arithmetische Mittel der Komponenten-Strich-zahlen aller Schriftzeichen gleicher Komponentenzahl berechnet.

2.3.2 Ergebnis und Bewertung

Das Ergebnis der Berechnungen ist in Tabelle 2 im Anhang dargestellt. In Abbildung 2.5 werden die empirischen Ergebnisse graphisch dargestellt. Auch auf der Schriftzeichenebene wird die Ausgangshypothese (Hypothese 2) durch die empirischen Ergebnisse gestützt.

Für die Anpassung der Funktionsgleichung $y = ax^b$ ($b < 0$) an die Daten wurden - wie in der vorhergehenden Untersuchung - nur diejenigen Werte berücksichtigt, zu deren Berechnung zehn oder mehr Datenbelege beitrugen. Für die Parameter wurden folgende Werte ermittelt: a = 4.8513, b = -0.2915 bei einem Determinationskoeffizienten von D = 0.9575.

In Tabelle 2 im Anhang werden die aus diesen Parametern errechneten theoretischen Werte den empirischen Werten gegenübergestellt. Abbildung 2.6 zeigt die empirisch gewonnenen Datenpunkte zusammen mit der theoretischen Kurve $y = 4.8513\ x^{-0.2915}$.

Auf der Basis des guten Determinationskoeffizienten kann die Gültigkeit des Menzerathschen Gesetzes für die Zeichen-Ebene der chinesischen Schrift angenommen werden.

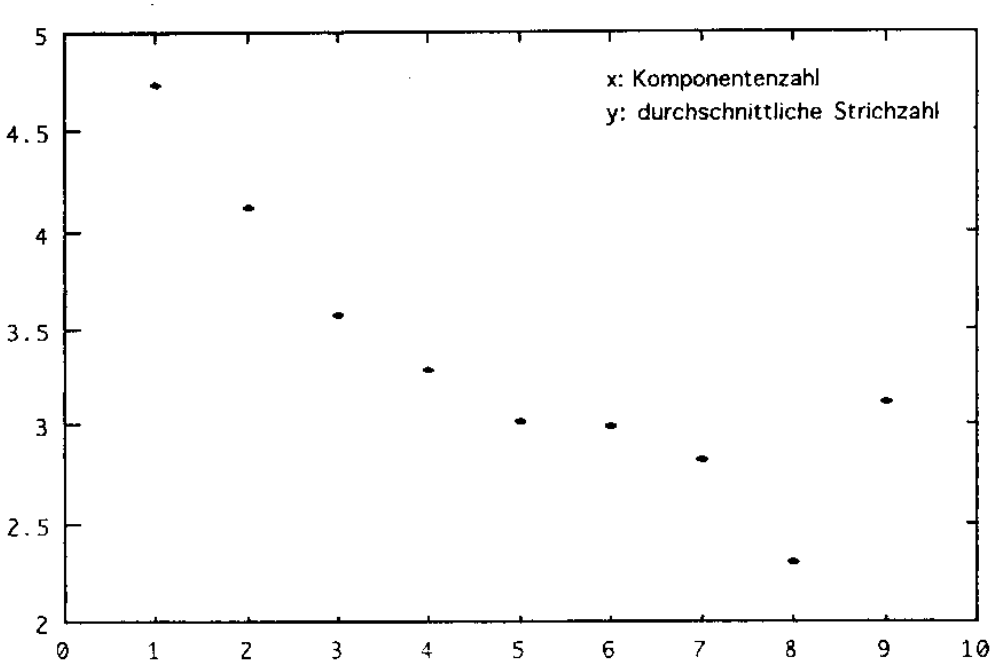


Abb. 2.5: Das Menzerathsche Gesetz auf Schriftzeichenebene - empirische Werte

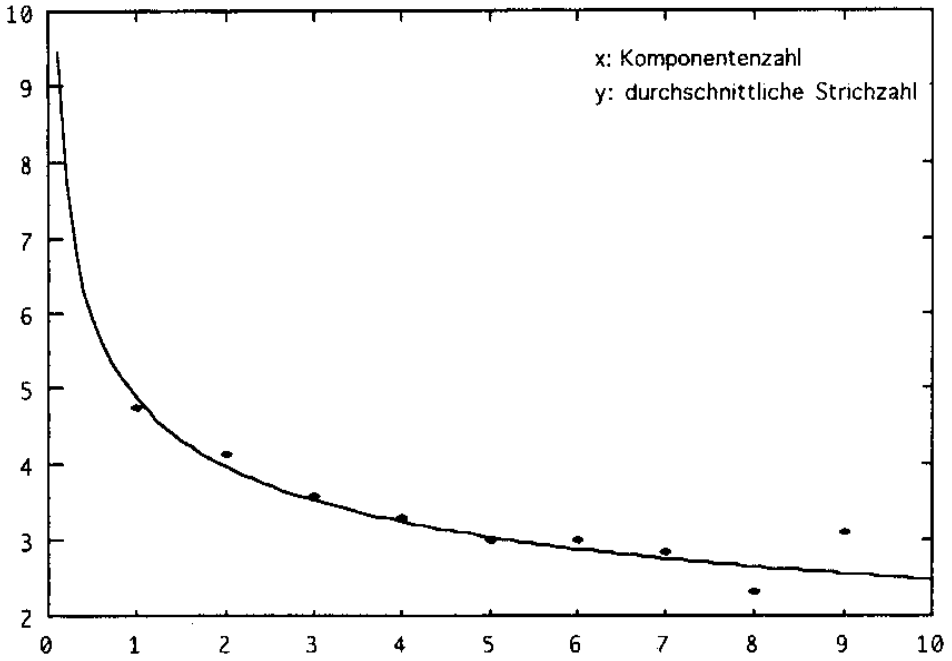


Abb. 2.6: Das Menzerathsche Gesetz auf Schriftzeichenebene - Gegenüberstellung der empirischen Werte und der theoretischen Kurve

2.4 Das Menzerathsche Gesetz auf Wortebene

Zum Menzerathschen Gesetz auf der Wortebene liegen eine Reihe von Untersuchungen an verschiedenen Sprachen vor, so etwa zum Verhältnis von Wortlänge und Silbenlänge (Altmann, 1980: Englisch, Bachka-Deutsch), Wortlänge und Morphlänge (Gerlach, 1982: Deutsch; Krott, 1994: Deutsch, Englisch, Niederländisch), Wortlänge und Bedeutungskomplexität (Altmann, Beöthy & Best, 1992: Deutsch, Slowakisch, Ungarisch; Rothe, 1983: Französisch, Portugiesisch, Spanisch; Fickermann, Markner-Jäger & Rothe, 1984: Englisch, Schwedisch, Indonesisch; Sambor, 1984: Polnisch, Russisch; Schwibbe, 1984: Deutsch).

Im Folgenden soll der Zusammenhang zwischen Wortlänge und Zeichenkomplexität auf der graphischen Ebene im Chinesischen untersucht werden.

2.4.1 Operationalisierung

Als Datenbasis für die Überprüfung von Hypothese 3 wurde eine systematische Wörterstichprobe aus dem „Neuen Chinesisch-Deutschen Wörterbuch“ (XHDC 1985) erhoben. Es resultierte eine Liste von 1105 Wörtern mit Wortlängen zwischen einem und fünf Schriftzeichen.

Jedes Wort der Liste wurde Schriftzeichenweise mit der für die vorhergehende Untersuchung erstellten Zeichen-Komponentenzahl-Liste abgeglichen, so dass für jedes Wort die durchschnittliche Komponentenzahl pro Zeichen errechnet werden konnte. Anschließend wurden die arithmetischen Mittel der durchschnittlichen Komponentenzahl aller Wörter gleicher Wortlänge gebildet.

2.4.2 Ergebnis und Bewertung

In Tabelle 3 im Anhang wird das Ergebnis der Berechnungen wiedergegeben. In Abbildung 2.7 werden die empirischen Werte graphisch dargestellt. Auch auf der Wortebene bestätigt der optische Trend die Grundaussage der Ausgangshypothese.

Bei der Anpassung der Funktionsgleichung $y = ax^b$ (b < 0) an die empirischen Werte der Wortlängen-Untersuchung wurden nur die Wortlängen berücksichtigt, für die mindestens zehn Belege vorlagen. Die Anpassung lieferte für die Parameter folgende Werte: a = 2.6960, b = -0.1539 bei einem Determinationskoeffizienten von D = 0.9659.

In Tabelle 3 im Anhang werden die aus den Parametern errechneten theoretischen Werte den empirischen Werten gegenübergestellt. Abbildung 2.8 zeigt die empirisch gewonnenen Datenpunkte zusammen mit der theoretischen Kurve $y = 2.6960\, x^{-0.1539}$.

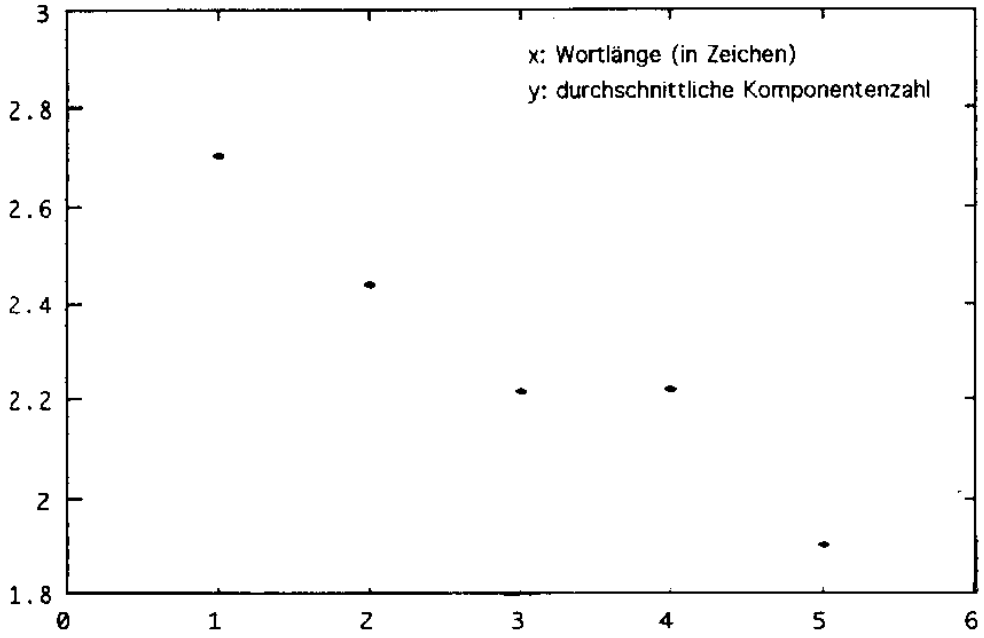


Abb. 2.7: Das Menzerathsche Gesetz auf Wortebene - empirische Werte

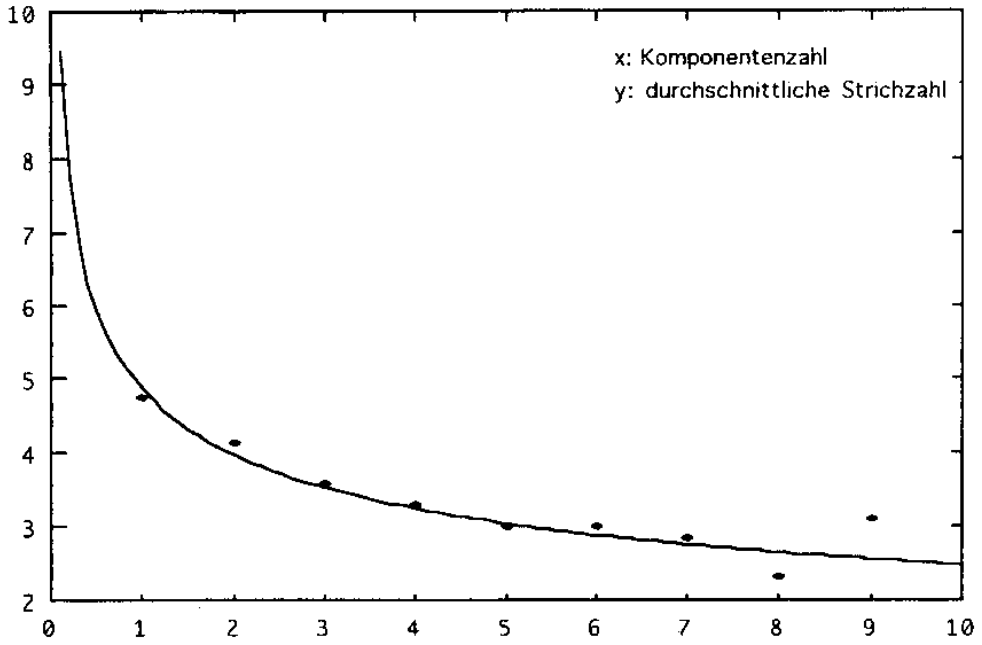


Abb. 2.8: Das Menzerathsche Gesetz auf Wortebene - Gegenüberstellung der empirischen Werte und der theoretischen Kurve

Auch auf der Wortebene erlaubt der gute Determinationskoeffizient die Annahme der Gültigkeit des Menzerathschen Gesetzes für die chinesische Schrift.

2.5 Das Menzerathsche Gesetz auf Teilsatzebene

Der Untersuchung zum Menzerathschen Gesetz auf Teilsatzebene soll eine graphisch motivierte Einheit Teilsatz zugrunde gelegt werden, die sich an der im modernen Chinesisch verwendeten Interpunktion orientiert. Als satz- und teilsatzabschließende Zeichen sind dabei der Punkt, das Fragezeichen und das Ausrufezeichen zu berücksichtigen, als teilsatzabschließende Zeichen das Komma, das Semikolon, das Kolon, nicht jedoch das chinesische Aufzählungskomma.

2.5.1 Operationalisierung

Arbeitshypothese 4 soll im Folgenden an einem großen Textkorpus sowie an einem Einzeltext überprüft werden.

Als Textkorpus wurde das PH-Korpus von Guo und Lui (1994) herangezogen, das aus 7.907 Meldungen der volksrepublikanischen Nachrichtenagentur Xinhua von 77 Tagen im Zeitraum von Januar 1990 bis März 1991 besteht und etwa vier Millionen Schriftzeichen umfasst. Das Korpus ist maschinenlesbar aufbereitet und wurde automatisch wortsegmentiert. Leider wird von den Autoren keine Angabe über die durchschnittliche Genauigkeit der Segmentierung gemacht. Es ist jedoch davon auszugehen, dass auf dieser Grundlage durchgeführte Untersuchungsergebnisse nicht unerheblich verfälscht sind.

Da sich eine Nachsegmentierung des kompletten Korpus nicht realisieren ließ, wurde aus dem Korpus ein längerer Einzeltext eher narrativen Charakters („Rensheng de jiazhi zaiyu fengxian“) gewählt und von Hand nachbearbeitet, wobei offensichtliche Segmentierungsfehler beseitigt wurden. Für die Untersuchung standen also sowohl das gesamte Textkorpus als auch ein handsegmentierter Einzeltext zur Verfügung.

Nun konnten - getrennt für Gesamtkorpus und Einzeltext - die Teilsatzlängen und ihre entsprechenden durchschnittlichen Wortlängen berechnet werden. Schließlich wurde das arithmetische Mittel der Wortlängen aller Teilsätze gleicher Wortzahl gebildet.

2.5.2 Ergebnis und Bewertung

In Tabelle 4 und Tabelle 5 im Anhang sind die empirischen Ergebnisse der Untersuchungen aufgeführt. Abbildung 2.9 zeigt die graphische Darstellung der empirischen Untersuchung des gesamten Korpus, Abbildung 2.10 die des Einzeltextes.

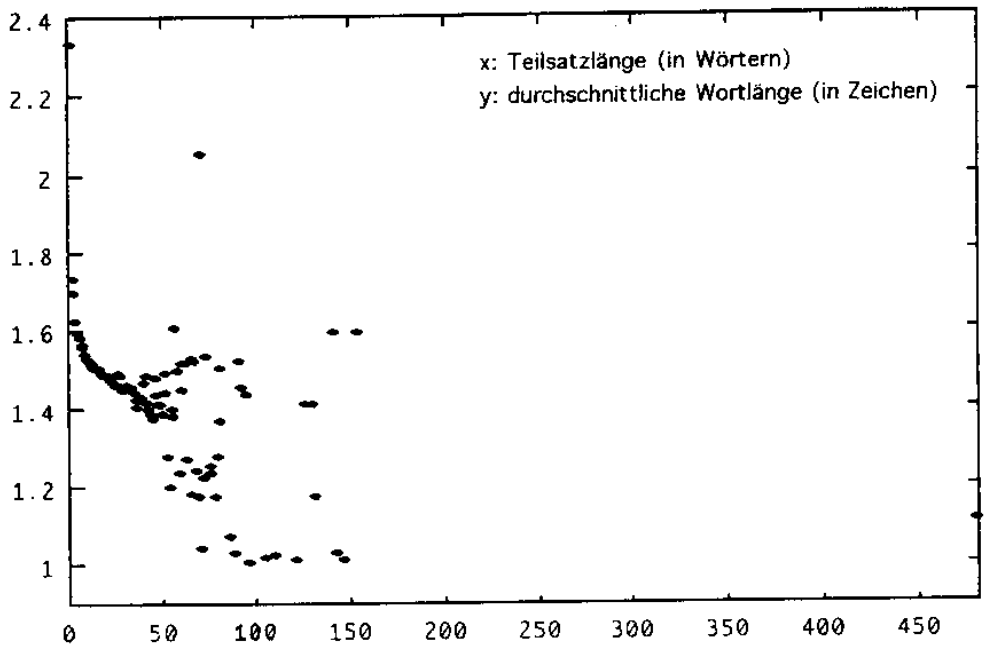


Abb. 2.9: Das Menzerathsche Gesetz auf Teilsatzebene: Korpus - empirische Werte

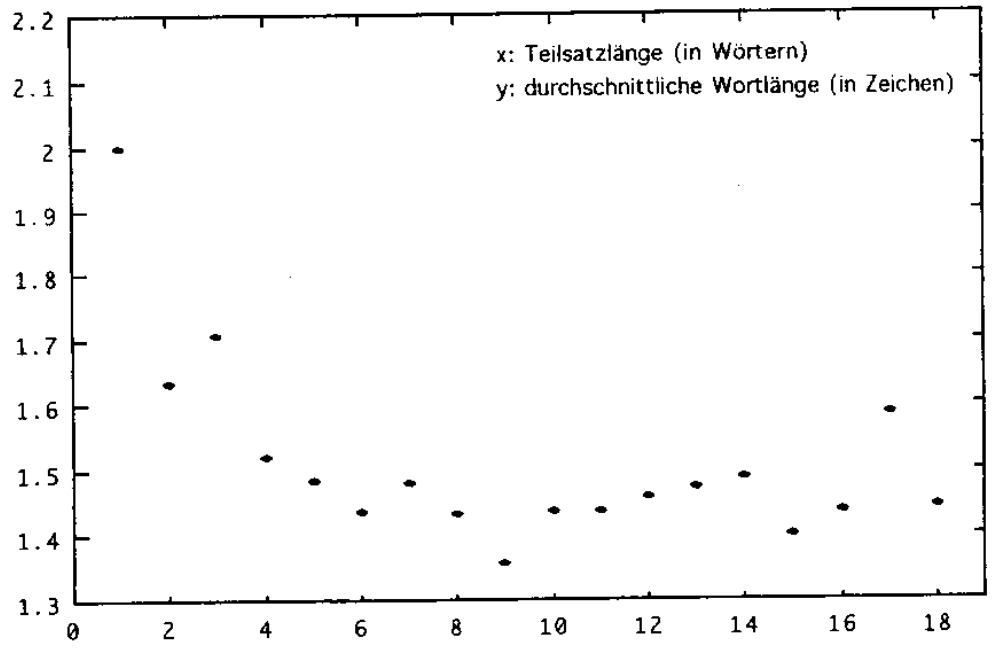


Abb. 2.10: Das Menzerathsche Gesetz auf Teilsatzebene: Einzeltext - empirische Werte

Bei Abbildung 2.9 fällt zunächst die große Zahl von Teilsätzen mit sehr hoher Wortzahl auf. Diese extremen Teilsatzlängen lassen sich auf Texte mit langen Aufzählungen, etwa von Konferenzteilnehmern, zurückführen.

Dem optischen Eindruck nach ist die Tendenz zu fallenden Wortlängen bei steigender Teilsatzlänge im Anfangsbereich deutlich zu erkennen. Ab einer Teilsatzlänge von etwa 30 Wörtern zeigt sich jedoch eine starke Streuung der Ergebnisse.

Auch für den Einzeltext bestätigt Abbildung 2.10 den erwarteten Trend, allerdings mit erheblichen Abweichungen einzelner Punkte der Ergebnismenge.

Die Anpassung der Funktionsgleichung $y = ax^b$ (b < 0) an die empirischen Ergebnisse, bei der wiederum lediglich Datenpunkte mit mindestens zehn Belegen berücksichtigt wurden, erbrachte für die Korpus-Untersuchung die folgenden Funktionsparameter: a = 1.9377, b = -0.0864 bei einem Determinationskoeffizienten von D = 0.7379. Die entsprechenden Werte der Einzeltext-Untersuchung lauten: a = 1.9207, b = -0.1427 bei einem Determinationskoeffizienten von D = 0.8789.

Die theoretischen Werte der Teilsatzuntersuchungen sind in den Tabellen 4 und 5 im Anhang aufgeführt. In Abbildung 2.11 und 2.12 werden die empirischen Datenpunkte den theoretischen Kurven gegenübergestellt.

In beiden Fällen kann die Gültigkeit des Menzerathschen Gesetzes aufgrund der guten bis befriedigenden Determinationskoeffizienten vorläufig angenommen werden.

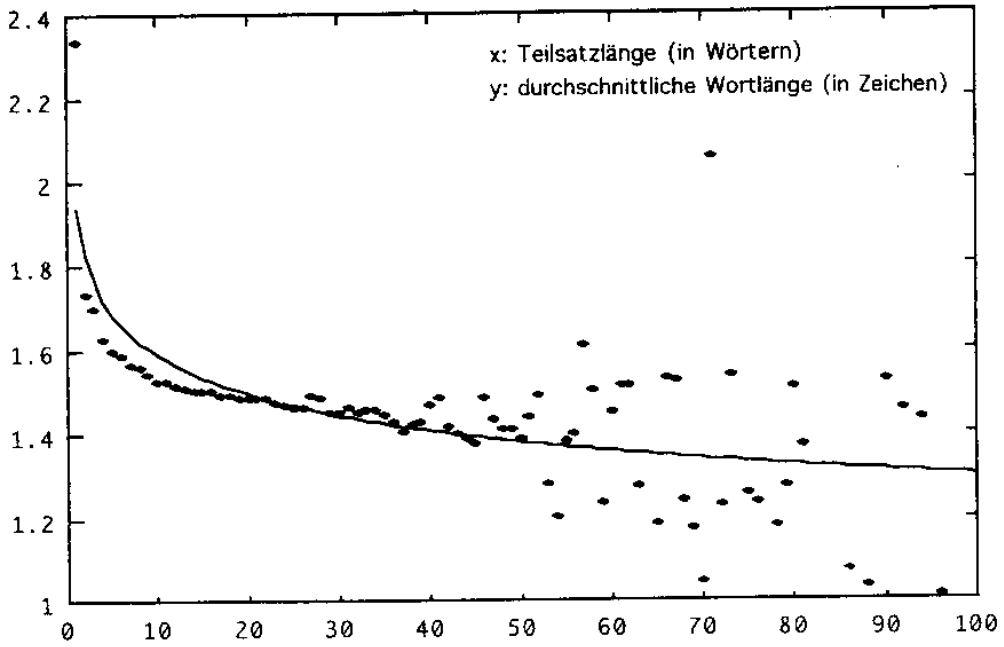


Abb. 2.11: Das Menzerathsche Gesetz auf Teilsatzebene: Korpus - Gegenüberstellung der empirischen Werte und der theoretischen Kurve (Ausschnitt).

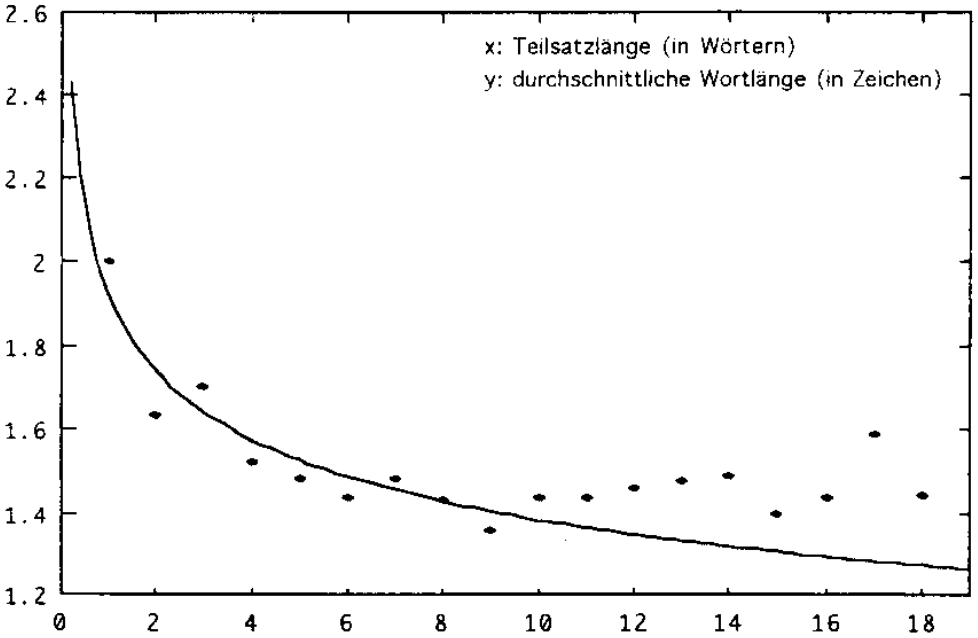


Abb. 2.12: Das Menzerathsche Gesetz auf Teilsatzebene: Einzeltext-Gegenüberstellung der empirischen Werte und der theoretischen Kurve.

2.6 Das Menzerathsche Gesetz auf Satzebene

Zum Menzerathschen Gesetz auf Satzebene gibt es bereits mehrere Untersuchungen an verschiedenen Sprachen, so etwa von Köhler (1982: Englisch), Heups (1983: Deutsch) und Teupenhayn und Altmann (1984: Deutsch, Englisch, Französisch, Schwedisch, Ungarisch, Slowakisch, Tschechisch, Indonesisch). Im Folgenden soll die Gültigkeit des Gesetzes an chinesischen Sprachdaten nachgewiesen werden.

2.6.1 Operationalisierung

Wie in der vorangegangenen Untersuchung zur Teilsatz-Wort-Beziehung wird auch hier von einer graphischen Teilsatz- bzw. Satzdefinition ausgegangen (vgl. auch Heups, 1983:114).

Als Datengrundlage der Satz-Teilsatz-Beziehung werden das PH-Korpus und der für die Teilsatz-Untersuchung handsegmentierte Einzeltext herangezogen.

Getrennt für Korpus und Einzeltext wurden für jeden Satz die Satzlänge in der Anzahl seiner Teilsätze sowie die durchschnittlich Teilsatzlänge in Anzahl der Wörter berechnet. Im zweiten Schritt wurde für alle Sätze gleicher Länge das arithmetische Mittel der Teilsatzlängen berechnet.

2.6.2 Ergebnis und Bewertung

Die Ergebnisse der Untersuchungen sind in Tabelle 6 (Korpus) und Tabelle 7 (Einzeltext) im Anhang wiedergegeben. Abbildung 2.13 stellt die empirischen Werte der Korpusuntersuchung, Abbildung 2.14 die der Einzeltextuntersuchung graphisch dar.

Im Falle des Korpus scheint der optische Trend die Ausgangshypothese gut zu bestätigen. Erst ab einer Satzlänge von etwa 10 Teilsätzen macht sich eine deutliche Streuung bemerkbar. Weniger deutlich ist dagegen das Ergebnis für den Einzeltext. Zwar ist ein eindeutig fallender Trend erkennbar, die Abweichung von einer gedachten Idealkurve scheint jedoch erheblich. Die Anpassung der Funktionsgleichung $y = ax^b$ (b < 0) an die empirischen Werte, die aus mindestens zehn Belegen hervorgegangen waren, ergab für die Korpus-Untersuchung die folgenden Funktionsparameter: a = 10.8069, b = -0.2301 bei einem Determinationskoeffizienten von D = 0.9750.

Die entsprechenden Werte der Einzeltext-Untersuchung lauten: a = 6.1360, b = -0.1264 bei einem Determinationskoeffizienten von D = 0.5628.

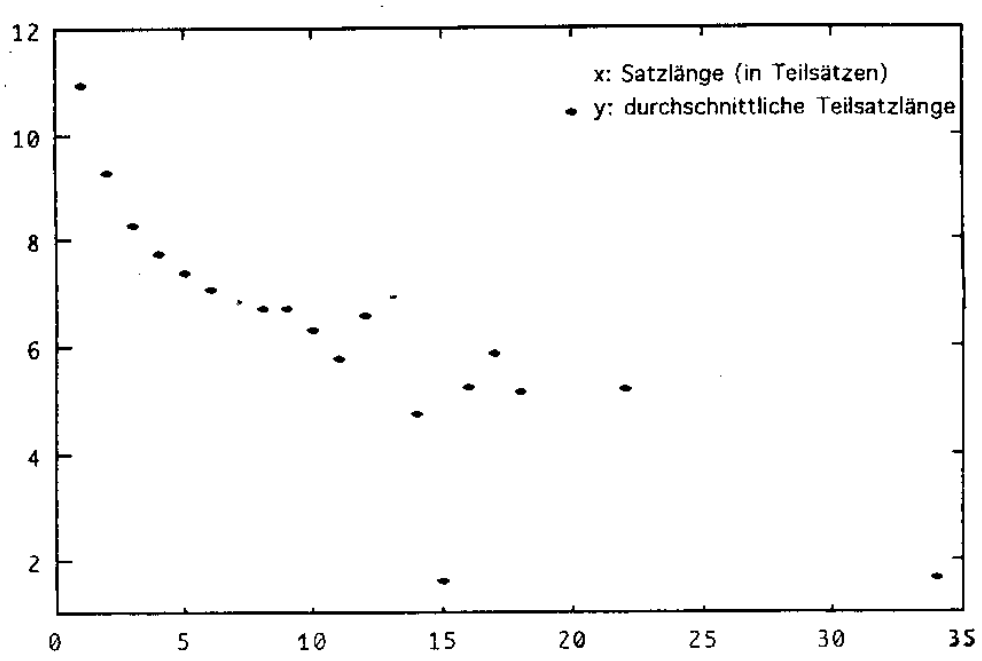


Abb. 2.13: Das Menzerathsche Gesetz auf Satzebene: Korpus - empirische Werte

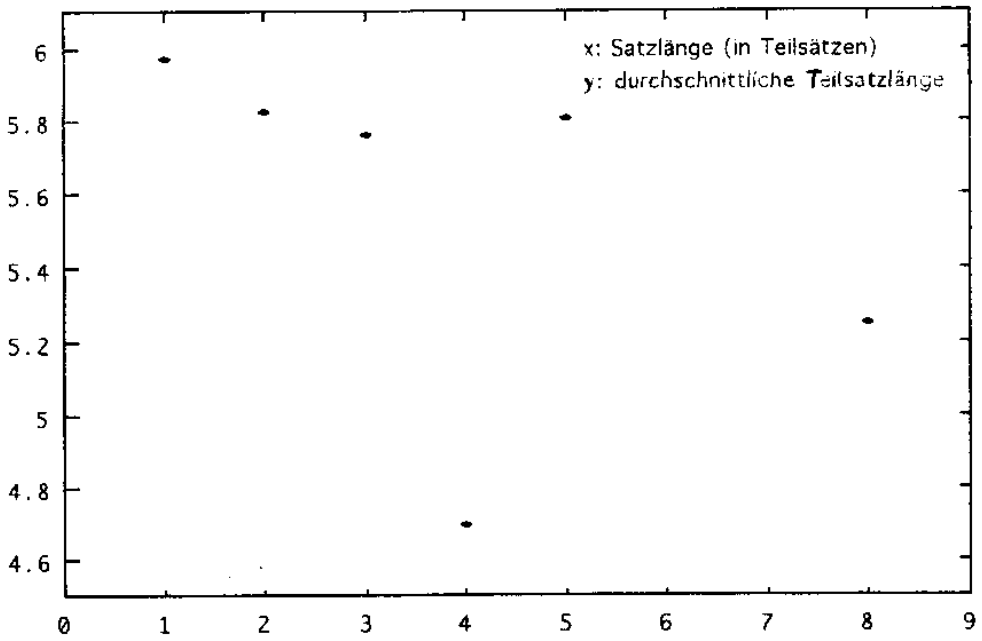


Abb. 2.14: Das Menzerathsche Gesetz auf Satzebene: Einzeltext - empirische Werte

Die den empirischen Ergebnissen entsprechenden theoretischen Werte sind in Tabelle 6 für die Korpusuntersuchung bzw. in Tabelle 7 für den Einzeltext im Anhang aufgeführt. Abbildung 2.15 zeigt die empirischen Werte der Korpusuntersuchung gemeinsam mit der theoretischen Kurve $y = 10.8069\ x^{-0.2301}$. Abbildung 2.16 zeigt das entsprechende Ergebnis der Einzeltextuntersuchung, zusammen mit der theoretischen Kurve $y = 6.1360\ x^{-0.1264}$.

Dem für die Korpusdaten errechneten sehr guten Determinationskoeffizienten steht der eher schlechte Determinationskoeffizient der Einzeltextanpassung gegenüber. Das schlechte Ergebnis im Falle der Einzeltextuntersuchung lässt sich aber leicht durch den zu geringen Stichprobenumfang von nur 180 Sätzen erklären. Daher kann an dieser Stelle die Validität des Menzerathschen Gesetzes auf der Satzebene für das moderne Chinesisch vorläufig angenommen werden.

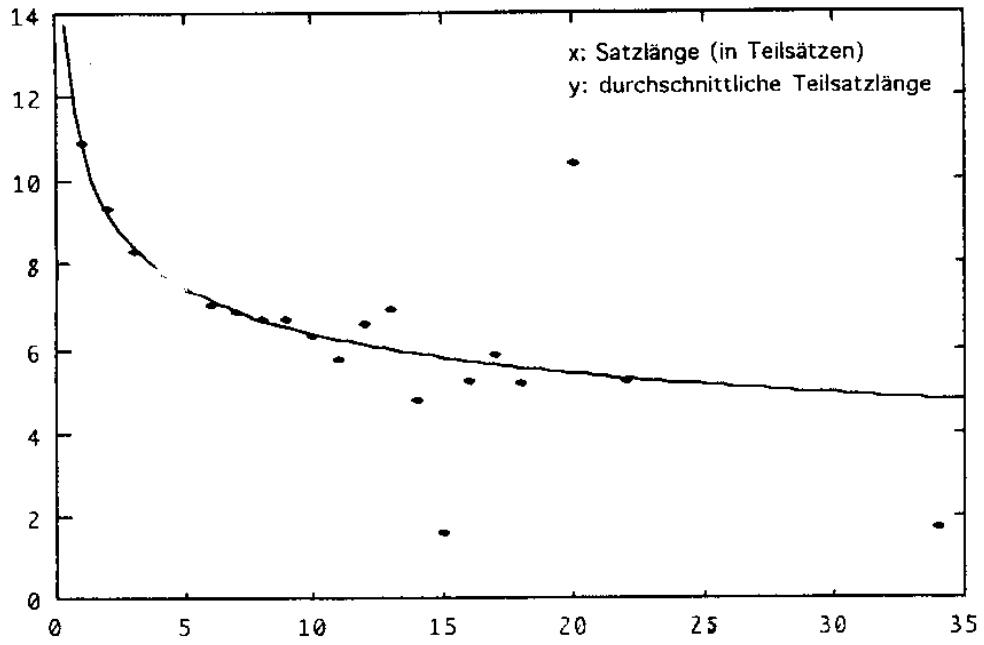


Abb. 2.15: Das Menzerathsche Gesetz auf Satzebene: Korpus - Gegenüberstellung der empirischen Werte und der theoretischen Kurve

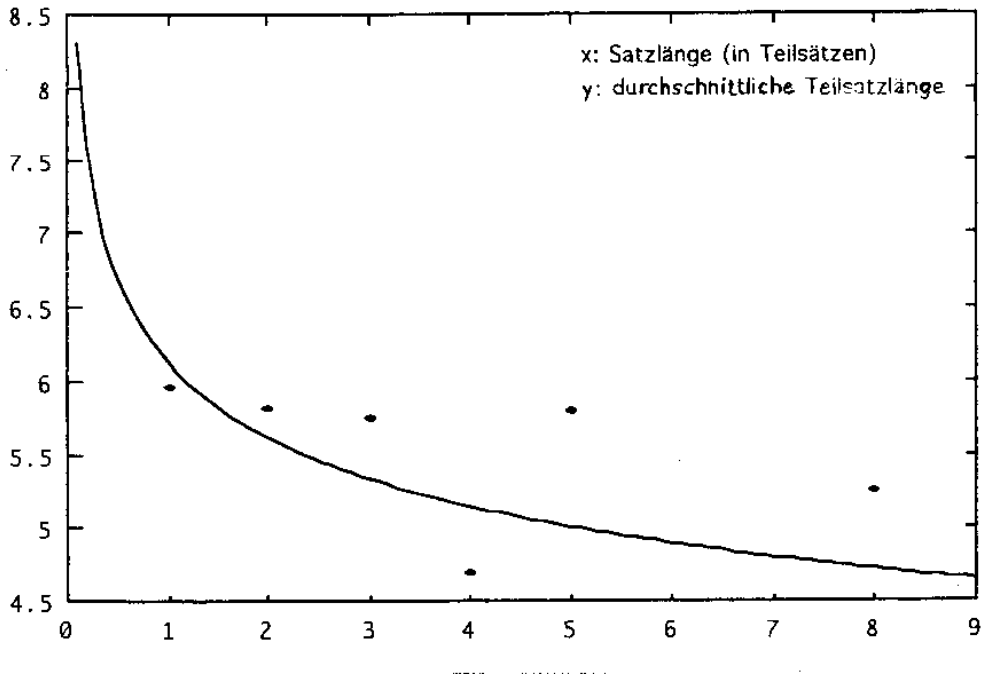


Abb. 2.16: Das Menzerathsche Gesetz auf Satzebene: Einzeltext - Gegenüberstellung der empirischen Werte und der theoretischen Kurve

2.7 Zusammenfassung der Untersuchungsergebnisse

Durch die empirische Überprüfung der Ausgangshypothesen auf der Komponenten-, Schriftzeichen-, Wort-, Teilsatz- und Satzebene konnte die Validität des Menzerathschen Gesetzes für die chinesische Schrift auf all diesen Ebenen vorläufig bestätigt werden. Dabei stimmen die empirischen Werte auf der Komponentenebene, der Zeichenebene und der Wortebene besonders gut mit den erwarteten theoretischen Werten überein, was auch in den guten Determinationskoeffizienten zum Ausdruck kommt. Das Resultat spricht für die Annahme der Gültigkeit des Menzerathschen Gesetzes auf diesen Ebenen der chinesischen Schrift.

Die Ergebnisse auf der Teilsatzebene sind aufgrund der ermittelten Determinationskoeffizienten - sowohl bei der Untersuchung des Textkorpus als auch bei der Einzeltextuntersuchung - als gut bis befriedigend zu werten. Auch für diese Ebene kann die Gültigkeit des Menzerathschen Gesetzes vorläufig angenommen werden, sollte aber nochmals an größeren handsegmentierten Korpora und längeren Einzeltexten überprüft werden.

Auf der Satzebene erzielte die Korpus-Untersuchung ein sehr gutes Resultat, das im starken Gegensatz zu dem der Einzeltextuntersuchung steht. Das relativ schlechte Ergebnis der Einzeltextuntersuchung kann jedoch leicht auf den zu geringen Stichprobenumfang zurückgeführt werden, so dass das Menzerathsche Gesetz trotzdem auch auf der Satzebene als vorläufig bestätigt betrachtet werden kann, wenngleich auch hier weitere Untersuchungen an längeren, handsegmentierten Einzeltexten wünschenswert bleiben.

3. Häufigkeitsverteilungen

Im folgenden sollen die bei den Untersuchungen zum Menzerathschen Gesetz erhobenen Daten zur Häufigkeitsverteilung der Elemente in den untersuchten Inventaren, Wörterbuchstichproben, Texten und Korpora auf den verschiedenen Ebenen des chinesischen Schriftsystems noch einmal gesondert betrachtet werden.

3.1 Komponentenkomplexität

Abbildung 3.1 zeigt die Verteilung der Komponenten des in Kapitel 2.2 beschriebenen Komponenteninventars nach ihrer Strichzahl (Daten in Tabelle 8 im Anhang). Mit zunehmender Strichzahl wächst die Zahl der Komponenten der jeweiligen Strichzahl im Inventar, fällt jedoch ab einer gewissen Komplexität wieder langsam ab.

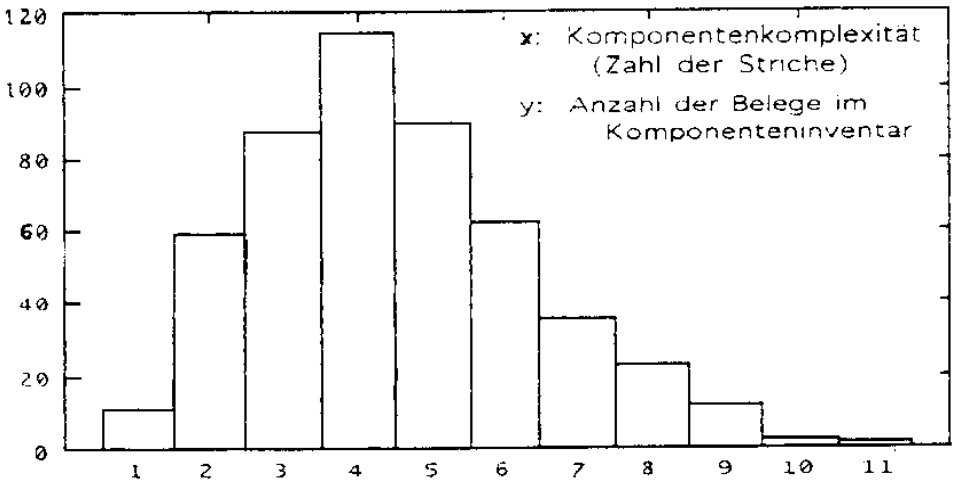


Abb. 3.1: Verteilung der Komponenten nach Strichzahl

Der Verlauf der Verteilung lässt sich möglicherweise auf einen Ausgleich zwischen den kombinatorischen Möglichkeiten und dem Leser-Bedürfnis nach Distinktivität einerseits (je mehr Striche, desto mehr Komponenten lassen sich bilden, und desto leichter ist eine Komponente von anderen zu unterscheiden) und dem Ökonomiebedürfnis des Schreibers andererseits (mit steigender Strichzahl erhöht sich der Produktionsaufwand) zurückführen. Eine theoretische Ableitung der Verteilungsfunktion konnte jedoch im Rahmen der vorliegenden Untersuchung nicht durchgeführt werden.

Die automatische Anpassung an die empirische Häufigkeitsverteilung führte zu einer guten Annäherung an die Poisson-Verteilung bei einem Wert für den Parameter *?* von *?* =3,5426.

Als Maß der Güte der Anpassung wurden dabei folgende als gut zu bewertende Werte ermittelt: X^2 = 5,3360 (FG = 9), $P(X^2)$ = 0,8141, C = 0,0107 ($C = \sqrt{(X^2 / N)}$.

In Tabelle 8 im Anhang werden den empirischen Werten die gemäß der Poisson-Verteilung zu erwartenden Werte gegenübergestellt.

3.2 Zeichenkomplexität

Auch die in Abbildung 3.2 dargestellte Verteilung der selbständigen Zeichen des volksrepublikanischen Computerstandards nach Komponentenzahl (Daten in Tabelle 9 im Anhang) zeigt bei zunehmender Komponentenzahl zunächst einen starken Anstieg, dann einen langsamen Abfall in den Zeichenhäufigkeiten.

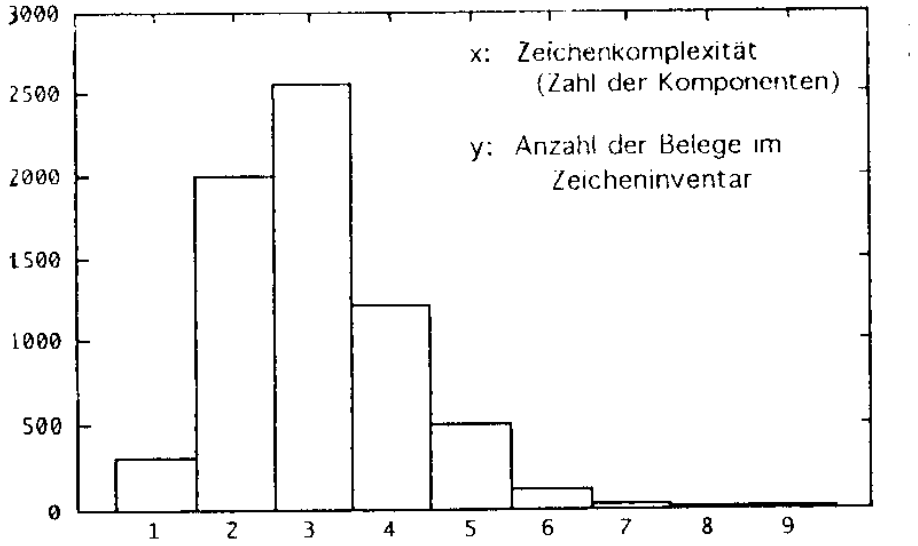


Abb. 3.2: Verteilung der Schriftzeichen nach Komponentenzahl

Analog zur Komponentenkomplexität ließe sich auch der Verlauf dieser Verteilung auf einen Ausgleich zwischen den kombinatorischen Möglichkeiten und dem Bedürfnis nach Distinktivität (je mehr Komponenten, desto mehr Zeichen lassen sich bilden und desto leichter ist ein Schriftzeichen von anderen zu unterscheiden) und dem Ökonomiebedürfnis (mit steigender Komponentenzahl erhöht sich der Produktionsaufwand) zurückführen. Wie im Falle der Komponentenkomplexität konnte jedoch keine theoretische Ableitung der Verteilungsfunktion durchgeführt werden.

Geht man davon aus, dass für die Verteilung der Schriftzeichenkomplexität im Zeicheninventar ähnliche Faktoren wie für die Verteilung der Komponentenkomplexität im Komponenteninventar wirksam sind, so liegt auch für die Daten zur Schriftzeichenhäufigkeit eine Poisson-Verteilung nahe. Die Anpassung der Poisson-Verteilung lieferte jedoch keine befriedigenden Ergebnisse.

Die automatische Anpassung an die Daten mit Hilfe des Altmann-Fitters ergab jedoch gute Werte für die Güte der Anpassung (*C*-Werte) für die Dacey-Poisson-Verteilung ($C = 0{,}0092$), die Verallgemeinerte Dacey-Poisson-Verteilung ($C = 0{,}0095$) und die Hyperpoisson-Verteilung ($C = 0{,}0108$), so dass anzunehmen ist, dass auf der Zeichenebene zusätzliche Faktoren wirksam sind. Tabelle 9 im Anhang stellt den empirischen Häufigkeiten die gemäß der Dacey-Poisson-Verteilung zu erwartenden Häufigkeiten gegenüber.

3.3 Wortlängenverteilung

Abbildung 3.3 zeigt die Wortlängenverteilung der Wörterbuchstichprobe aus Abschnitt 2.4 (Daten in Tabelle 10 im Anhang). Besonders auffällig ist der große Anteil von aus zwei Zeichen bestehenden Wörtern.

Die Anpassung der von Wimmer, Köhler, Grotjahn und Altmann (1994) abgeleiteten Wortlängenverteilungen - Conway-Maxwell-Poisson, Hyperpoisson, Hyperpascal, Negativ-Binomial, Palm-Poisson und Consul-Jain-Poisson - ergab keine befriedigenden Resultate.

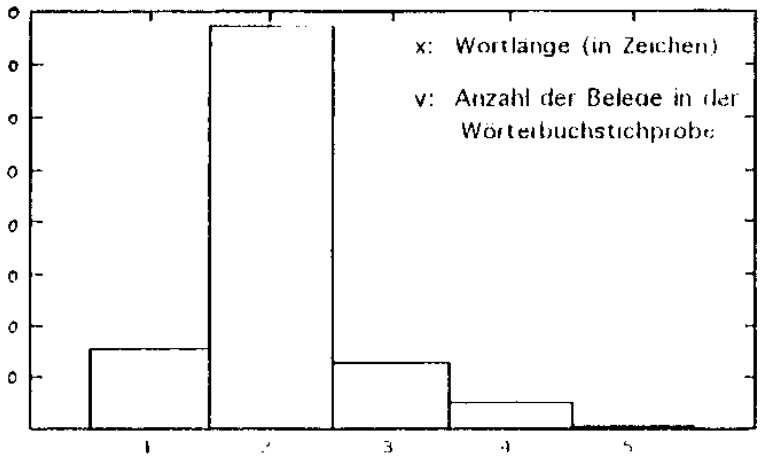


Abb. 3.3: Wortlängenhäufigkeiten

Dieses Ergebnis spricht für die Besonderheit der Wortbildung im Chinesischen im Vergleich zu anderen Sprachen, wie sie auch bei den Textstichprobenuntersuchungen von Zhu und Best (1992:51ff.) sowie Best und Zhu (1993) festgestellt und mit der Rolle des Sprechrhythmus in der chinesischen Sprache erklärt wird.

Die automatische Anpassung führte jedoch zu einer guten Anpassung für die Positive Cohen-Negative Binomialverteilung (siehe Tabelle 10).

3.4 Teilsatzlängenverteilung

Abbildung 3.4 zeigt die Verteilung der Teilsatzlängen, gemessen in Wörtern, des untersuchten Korpus (Daten in Tabelle 11 im Anhang), Abbildung 3.5 die entsprechende Verteilung im Einzeltext (Daten in Tabelle 12 im Anhang).

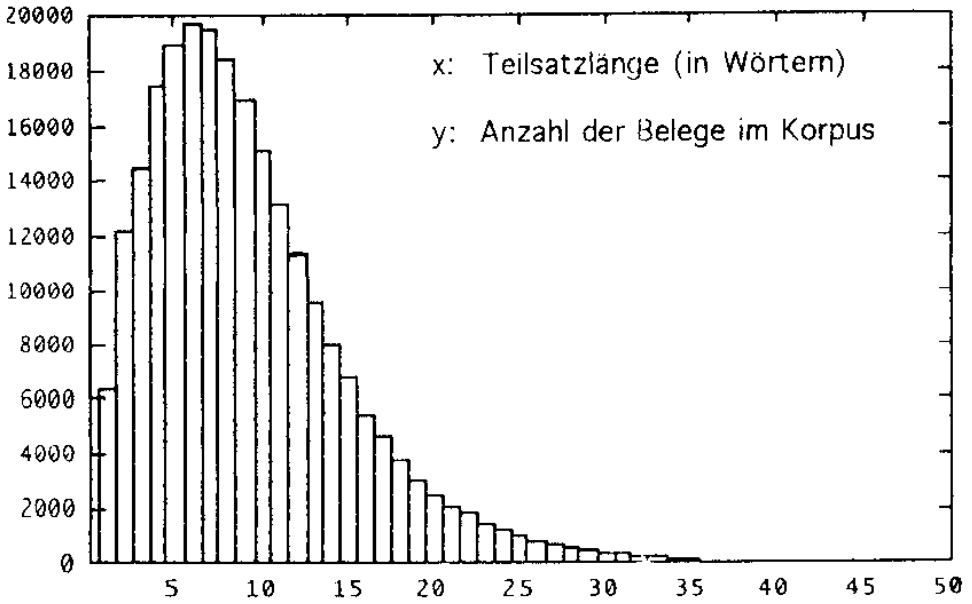


Abb. 3.4: Teilsatzlängen im Korpus

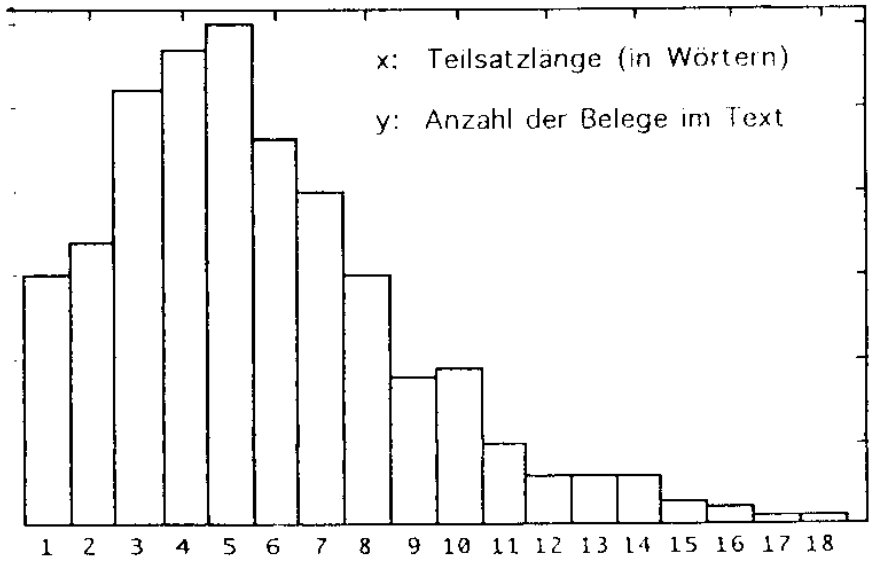


Abb. 3.5: Teilsatzlängen im Einzeltext

Dem Modell zur Satzlängenverteilung von Altmann (1988) folgend wurde eine Anpassung der negativen Binomialverteilung an die empirischen Ergebnisse vorgenommen.

Für die Teilsatzlängenverteilung im Korpus wurden die Parameter $k = 2.3457$, $p = 0.2195$ bei einem C-Wert von $C = 0.0134$ errechnet. Die Gegenüberstellung der empirischen Werte und der Erwartungswerte findet sich in Tabelle 11 im Anhang. Die Anpassung im Falle des Einzeltextes führte zu den Werten $k = 3.5831$

p = 0.4311 bei einem C-Wert von C = 0.0253. Beide Ergebnisse können als gute bis befriedigende Anpassung an das Modell gewertet werden.

3.5 Satzlängenverteilung

Abbildung 3.6 zeigt die Verteilung der Satzlängen, gemessen in Teilsätzen, des untersuchten Korpus (Daten in Tabelle 13 im Anhang), Abbildung 3.7 die entsprechende Verteilung im Einzeltext (Daten in Tabelle 14 im Anhang).

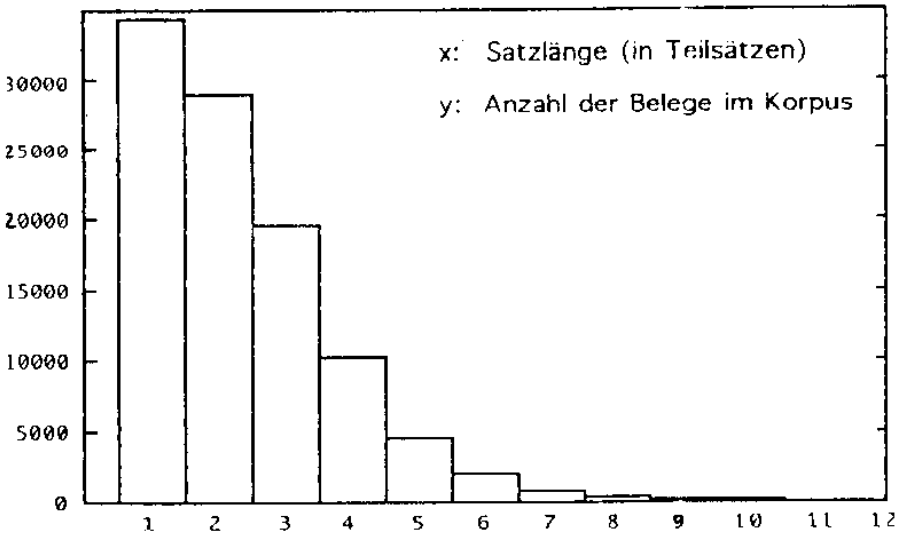


Abb. 3.6: Satzlängen im Korpus

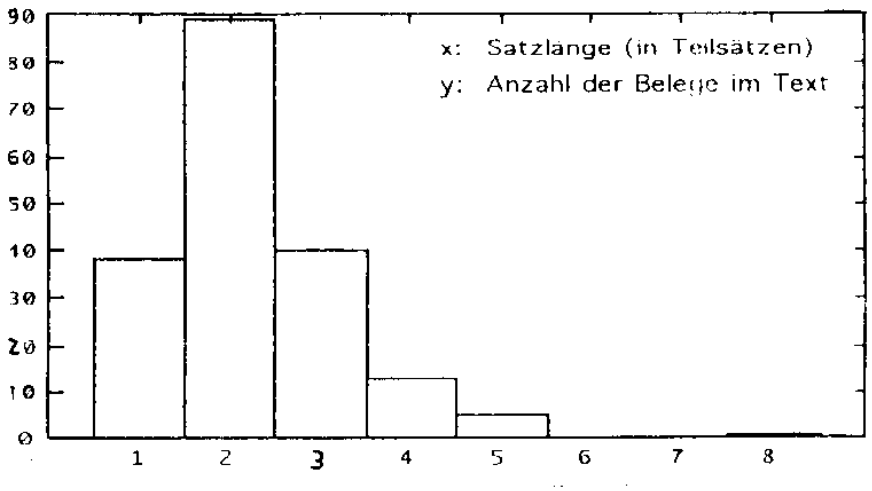


Abb. 3.7: Satzlängen im Einzeltext

Auffällig ist hier zunächst, dass im Falle des Einzeltextes die Zahl der aus zwei Teilsätzen bestehenden Sätze die der aus einem Teilsatz bestehenden Sätze bei weitem übersteigt, dann aber mit zunehmender Satzlänge ein kontinuierlicher Abfall zu verzeichnen ist. Dagegen überwiegen im Falle der Korpusuntersuchung

die aus einem Teilsatz bestehenden Sätze. Vermutlich ergibt sich dieser Unterschied aus der für das Korpus untypischen Textsorte des untersuchten Einzeltextes: Während die Mehrzahl der Texte im Korpus aus Kurzmeldungen besteht, handelt es sich bei dem Einzeltext um einen Text narrativen Charakters.

Altmann (1988) leitet für die Satzlängenverteilung - gemessen in Clauses - die negative Binomialverteilung ab und stellt bei der Überprüfung an Textstichproben aus zehn Texten typologisch unterschiedlicher Sprachen fest, dass alle untersuchten Texte dieser Verteilung folgen.

In der Tat liefert auch die Anpassung der negativen Binomialverteilung an die Satzlängenverteilung im Korpus ein sehr gutes Ergebnis (Gegenüberstellung der empirischen Häufigkeitsverteilung und der theoretischen Werte in Tabelle 13 im Anhang): $k = 2.5310$, $p = 0.6478$ bei einem C-Wert von $C = 0.0020$.

Die Anpassung der negativen Binomialverteilung an die Daten des untersuchten Einzeltextes führte dagegen zu einem weniger befriedigenden Resultat (Gegenüberstellung der empirischen Häufigkeitsverteilung und der theoretischen Werte in Tabelle 14 im Anhang): $k = 23.9328$, $p = 0.9508$ bei einem C-Wert von $C = 0.0838$.

4. Rang-Frequenzverteilung von Wörtern im Einzeltext

Anhand der Rang-Frequenzverteilung des Vokabulars des halbautomatisch wortsegmentierten, dem PH-Korpus von Guo und Lui (1994) entnommenen Einzeltextes, der schon zur Überprüfung des Menzerathschen Gesetzes herangezogen wurde, soll im folgenden die Gültigkeit des Zipfschen Gesetzes bzw. des Zipf-Mandelbrotschen Gestezes im Chinesischen überprüft werden.

Zur Untersuchung der Rang-Frequenzverteilung wurde zunächst das Vorkommen jedes Wortes im Text ausgezählt. Anschließend wurde die so ermittelte Wortliste nach absoluter Häufigkeit geordnet und daraus eine Liste mit Rangnummer und absoluter Frequenz erzeugt.

Tabelle 15 im Anhang zeigt die Rang-Frequenzverteilung des untersuchten Textes. Die Verteilung wird in Abbildung 4.1 graphisch dargestellt.

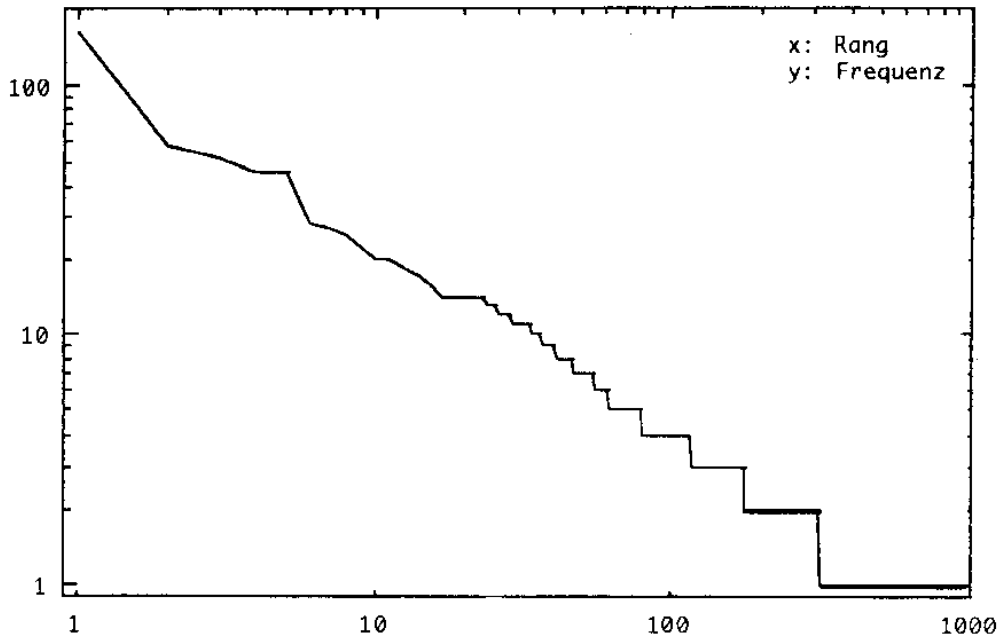


Abb. 4.1: Rang-Frequenz-Verteilung der Wörter im Einzeltext (doppelt logarithmisches Koordinatensystem)

Vergleicht man die graphische Darstellung der empirischen Verteilung mit Untersuchungen aus anderen Sprachen, so entspricht ihr Verlauf im großen und ganzen den Erwartungen. Der unregelmäßige Verlauf zu Beginn der Kurve lässt sich vermutlich auf die gewählte Wortsegmentierung - etwa im Falle von Eigennamen (z.B. Tabelle 15, Rang 3) und Affixen (z.B. Tabelle 15, Rang 5) – zurückführen. Die Anpassung der dem Zipfschen bzw. dem Zipf-Mandelbrot-schen Gesetz entsprechenden Funktionen bestätigt jedoch die Konformität der chinesischen Daten.

Die Anpassung der dem Zipfschen Gesetz entsprechenden Funktion (vgl. Arapov, 1982:36) an die Daten mit Hilfe nichtlinearer Regression ergab für den Parameter g den Wert g = -0.8067; für die Normierungskonstante C wurde der Wert C = 0.006889 errechnet.

Als Gütetest diente der Determinationskoeffizient, der mit D = 0.9713 als sehr gut zu bewerten ist.

Auch die Gegenüberstellung der theoretischen Funktion und der empirischen Daten in Abbildung 4.2. zeigt die gute Übereinstimmung der Verteilungen.

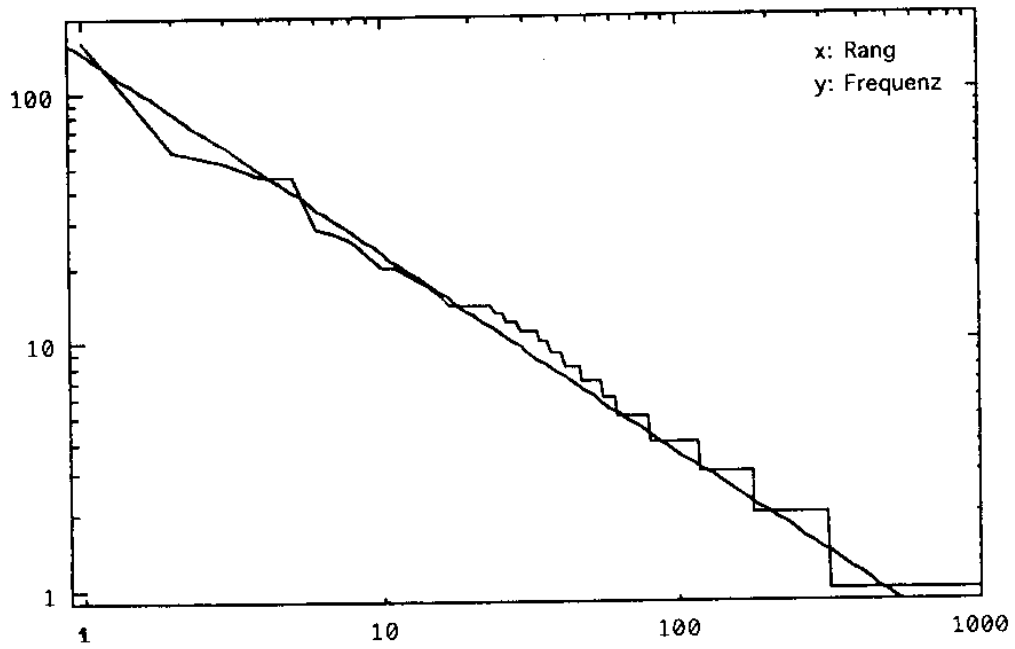


Abb. 4.2: Gegenüberstellung der empirischen Werte und der theoretischen Wahrscheinlichkeitsfunktion des *Zipfschen* Gesetzes (doppelt logarithmisches Koordinatensystem)

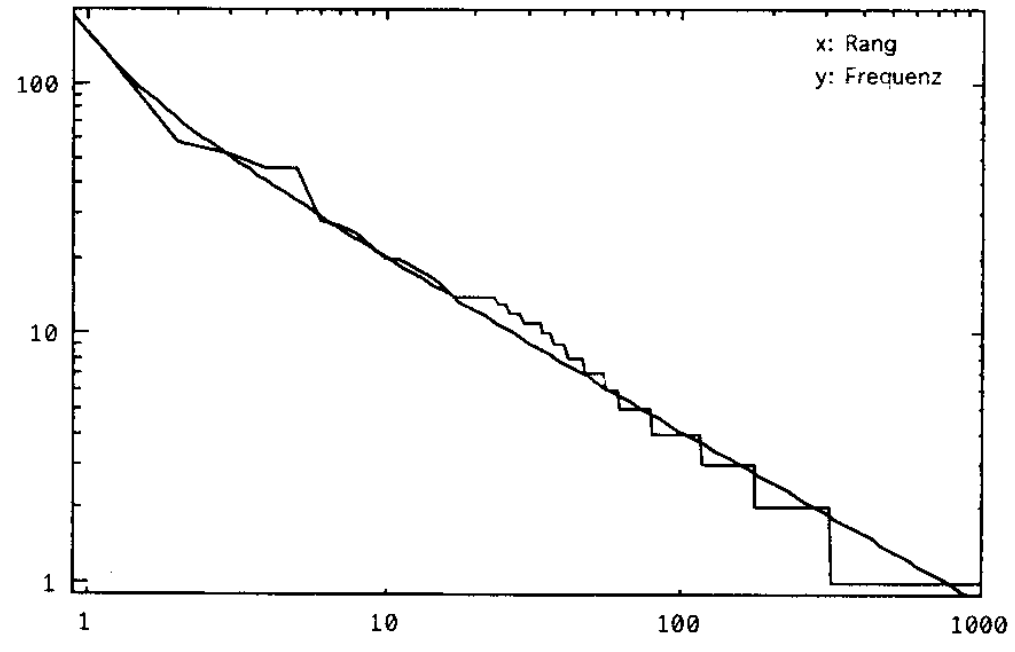


Abb. 4.3: Gegenüberstellung der empirischen Werte und der theoretischen Wahrscheinlichkeitsfunktion des *Zipf-Mandelbrotschen* Gesetzes (doppelt logarithmisches Koordinatensystem)

Noch besser scheinen die empirischen Daten und theoretische Funktion nach Zipf-Mandelbrot in Abbildung 4.3 übereinzustimmen.

Die Anpassung der dem Zipf-Mandelbrotschen-Gesetz entsprechenden Funktion (vgl. Arapov, 1982:36; Altmann, 1994:92) an die Daten ergab für die Parameter folgende Werte: b = -0.5385, $?$ = -0.6828. Für die Normierungskonstante C wurde der Wert C = 0.01063 ermittelt, bei einem als sehr gut zu bewertenden Determinationskoeffizienten von D = 0.9868.

Die Gültigkeit des Zipfschen Gesetzes bzw. des Zipf-Mandelbrotschen Gesetzes kann somit für den untersuchten chinesischen Text angenommen werden.

5. Schriftzeichenfrequenz und Phonetizität

Die Mehrzahl aller chinesischen Schriftzeichen enthält phonetische Elemente, anhand derer ihre Aussprache mehr oder minder genau erschlossen werden kann. Diese phonetischen Elemente lassen die Aussprache der zusammengesetzten Schriftzeichen, in denen sie verwendet werden, in unterschiedlichem Maße erkennen. In vielen Fällen vererben sie ihre Aussprache vollständig, Silbe wie Ton, auf das komplexe Zeichen, dessen Bestandteil sie sind. Häufig stimmt zwar die Silbe, nicht jedoch der Ton des phonetischen Elements mit dem des zusammengesetzten Zeichens überein. Zum Teil sind die phonetischen Gemeinsamkeiten nur noch in segmentalen Phonemen - etwa im Anlaut oder Auslaut - zu erkennen. In einigen Zeichen sind die phonetischen Elemente sogar völlig unwirksam (vgl. DeFrancis, 1984:101ff.). Die Phonetizität der einzelnen Schriftzeichen ist also unterschiedlich stark ausgeprägt.

Ein von DeFrancis (1984:109) durchgeführter Abgleich der 895 Elemente umfassenden Silbentabelle von Soothill mit einer Liste 4.719 hochfrequenter Zeichen - es handelt sich um traditionelle, nicht vereinfachte Zeichen - ergab, dass 14% der untersuchten Zeichen selbst phonetische Elemente darstellten, 87% der Zeichen phonetische Elemente enthielten und lediglich 3% keinen phonetischen Schlüssel enthielten. Stichproben von Zeichen unterschiedlicher Frequenzklassen ergaben, dass zwischen 60 und 81% der Zeichen nützliche phonetische Elemente enthielten (DeFrancis, 1984:108).

Eine von Zhou (1978) durchgeführte Untersuchung der modernen, vereinfachten Schriftzeichen eines weit verbreiteten modernen Wörterbuchs führte zu einer Liste von 1.348 phonetischen Elementen, die von Zhou (1980) ausführlich dokumentiert wurden. Zhous Untersuchung ergab, dass 17% der Zeichen phonetische Elemente darstellten, 81% der Zeichen enthielten phonetische Elemente und nur 2% der Zeichen enthielten keine phonetischen Elemente. Fasst man die Zahl der Zeichen mit nützlichen phonetischen Elementen und die Zahl der phone-

tischen Elemente selbst zusammen, so enthalten etwa 65% der Zeichen des Wörterbuchs nützliche phonetische Elemente (Zhou, 1978:172f.; DeFrancis, 1984:110).

Den Zusammenhang zwischen der Phonetizität und der Häufigkeit von Schriftzeichen untersucht DeFrancis (1984:105ff.) anhand einer Stichprobe von traditionellen Schriftzeichen verschiedener Frequenzklassen. Er kommt zu dem Ergebnis, dass in den hohen Frequenzklassen phonetische Elemente überdurchschnittlich häufig als eigenständige Zeichen vorkommen, während bei abnehmender Frequenz auch die Zahl eigenständiger phonetischer Elemente abnimmt. Dagegen steigt mit abnehmender Frequenz die Zahl zusammenge-setzter Schriftzeichen, die phonetische Elemente enthalten - zunächst ohne Berücksichtigung, ob diese Elemente auch tatsächlich phonetisch wirksam sind.

Die Untersuchung über die Anzahl der ‚nützlichen' phonetischen Elemente, also der Schriftzeichen, in denen die phonetischen Elemente tatsächlich Aufschluss über die Aussprache geben, ergab für die verschiedenen Frequenzklassen ein uneinheitliches Bild. Es lassen sich hier keine Aussagen über einen Zusammenhang zwischen Frequenz und Phonetizität machen. Aufgrund des relativ geringen und weitgestreuten Stichprobenumfangs sowie des stochastischen Charakters des möglichen Zusammenhangs lässt die Untersuchung von DeFrancis nicht den generellen Schluss zu, es bestehe keinerlei Verbindung zwischen den Größen Frequenz und Phonetizität. Es gibt vielmehr Faktoren, die einen Zusammenhang zwischen Zeichenfrequenz und Zeichenphonetizität vermuten lassen.

Geht man von der Voraussetzung aus, die chinesische Schrift sei in erster Linie ein phonetisches Schriftsystem, so sind aus systemtheoretischen Erwägungen folgende Zusammenhänge zwischen Phonetizität und Frequenz zu erwarten:

(a) Aus der Sicht des Schreibers

Sind die dem Schreiber zu Gebote stehenden Schriftzeichen im Gedächtnis primär nach phonetischen Schlüsseln abrufbar, so verwendet er diese phonetischen Zeichen vermutlich mit größerer Häufigkeit als Zeichen ohne nützlichen phonetischen Hinweis. Die Phonetizität müsste demnach mit fallender Zeichenfrequenz abnehmen.

Das Bedürfnis des Schreibers nach Minimierung des Produktionsaufwandes führt dazu, dass hochfrequente Schriftzeichen möglichst ‚einfach' sind, das heißt aus möglichst wenigen Strichen bestehen. Da aus phonetischem Element und nichtphonetischem Element zusammengesetzte Zeichen komplexer sind als die als Schriftzeichen auftretenden selbständigen phonetischen Elemente, ist der von DeFrancis beobachtete hohe Anteil der selbständigen phonetischen Elemente unter den hochfrequenten Zeichen leicht zu erklären.

(b) Aus der Sicht des Lesers

Auch der Leser ist gemäß seinem Bedürfnis nach Minimierung des Dekodierungsaufwandes daran interessiert, in möglichst vielen der auftretenden Schriftzeichen phonetische Hinweise zu finden. Es ist also auch für die dekodierende Seite im Kommunikationsprozess von Vorteil, wenn gerade hochfrequente Zeichen ein hohes Maß an Phonetizität besitzen.

Eine nähere Untersuchung des Zusammenhangs zwischen Schriftzeichenfrequenz und Schriftzeichenphonetizität an der modernen chinesischen Schrift scheint also durchaus gerechtfertigt.

5.1 Operationalisierung

Grundlage der Untersuchung bilden die Schriftzeichen des volksrepublikanischen Computerstandards GB 2312-80, wie sie in HSZ (1988) dokumentiert sind, die Zeichenfrequenzdaten des PH-Korpus(Guo & Lui, 1994) sowie die Silbentabelle von Zhou (1980).

Die Silbentabelle von Zhou enthält 1.348 phonetische Elemente, die auf der Grundlage aller Zeichen des Wörterbuchs Xinhua Zidian in der Ausgabe von 1971 gewonnen wurden, basiert also auf vereinfachten Zeichen (Zhou, 1978:172; Zhou, 1980:2). Die phonetischen Elemente sind nach ihrer Aussprache in der Lautumschrift Hanyu pinyin geordnet. Zusätzlich zur Silbe des phonetischen Elements wird bei selbständig auftretenden phonetischen Elementen auch der Ton angegeben. Einigen Elementen sind mehrere Aussprachen zugeordnet, die aber nur an einer Stelle der Tabelle eingeordnet wurden - Querverweise sind nicht vorhanden.

Zunächst wurde Zhous Silbentabelle um Querverweise erweitert, so dass Elemente mit mehr als einer Aussprache unter jeder dieser Aussprachen nachgeschlagen werden konnten. Elemente mit mehreren Aussprachen wurden außerdem als solche markiert. Unselbständige phonetische Elemente, denen von Zhou kein Ton zugewiesen wurde, denen aber aufgrund der von ihm angeführten Wörterbuchbelege eindeutig ein Ton zugeordnet werden kann, wurden mit der entsprechenden Tonangabe versehen.

Die selbständigen Schriftzeichen des GB-2312-80-Computerstandards aus dem HSZ (1988) wurden mit den phonetischen Elementen abgeglichen. Bei Zeichen mit mehr als einer Aussprache wurde nur die am häufigsten verwendete Aussprache berücksichtigt. Ausgehend von der Aussprache jedes Zeichens wurde in der modifizierten Silbentabelle nach phonetischen Elementen gleicher oder ähnlicher Aussprache gesucht. Jedem Zeichen wurde so ein Phonetizitätsmaß von '4' (Silbe und Ton eindeutig erkennbar), '3 (Silbe eindeutig erkennbar), '2 (mehrdeutiges phonetisches Element), '1 (ähnliche Silbe) oder '0' (kein phonetischer Hinweis) zugeordnet.

Im Einzelnen wurden die Phonetizitätswerte wie folgt vergeben:

Phonetizitätsmaß	Anzahl	Anteil (gerundet) in %
4	2.494	37,09%
3	1.301	19,35%
2	407	6,33%
1	642	9,55%
0	1.880	27,96%
3	6.724	100%

Schließlich wurde die Liste der GB-Zeichen, zusammen mit ihren Phonetizitätswerten, gemäß den absoluten Häufigkeiten ihres Vorkommens im PH-Korpus in absteigender Reihenfolge geordnet.

5.4 Untersuchungsergebnisse

In den Abbildungen 5.1, 5.2 und 5.3 sind Ausschnitte von jeweils 200 Datenpunkten unterschiedlicher Häufigkeitsklassen dargestellt: Abbildung 5.1 zeigt die Phonetizitätswerte der häufigsten 200 Schriftzeichen, Abbildung 5.2 die Werte für die Schriftzeichen der Ränge 3.000 bis 3.200, Abbildung 5.3 die Werte für die Schriftzeichen der Ränge 6.000 bis 6.200.

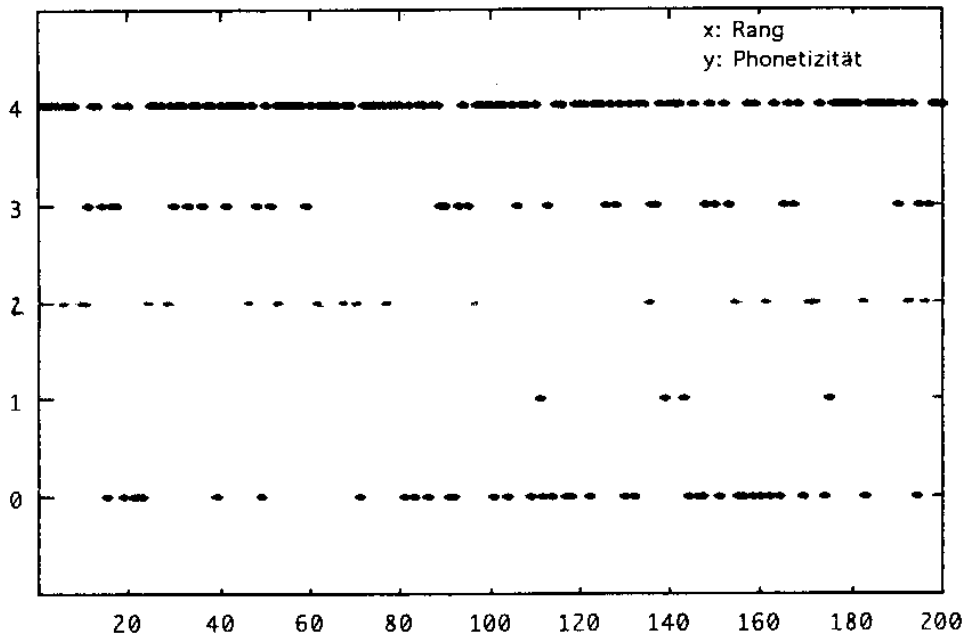


Abb. 5.1: Phonetizitätswerte der Ränge 0 bis 200 aller GB-Zeichen

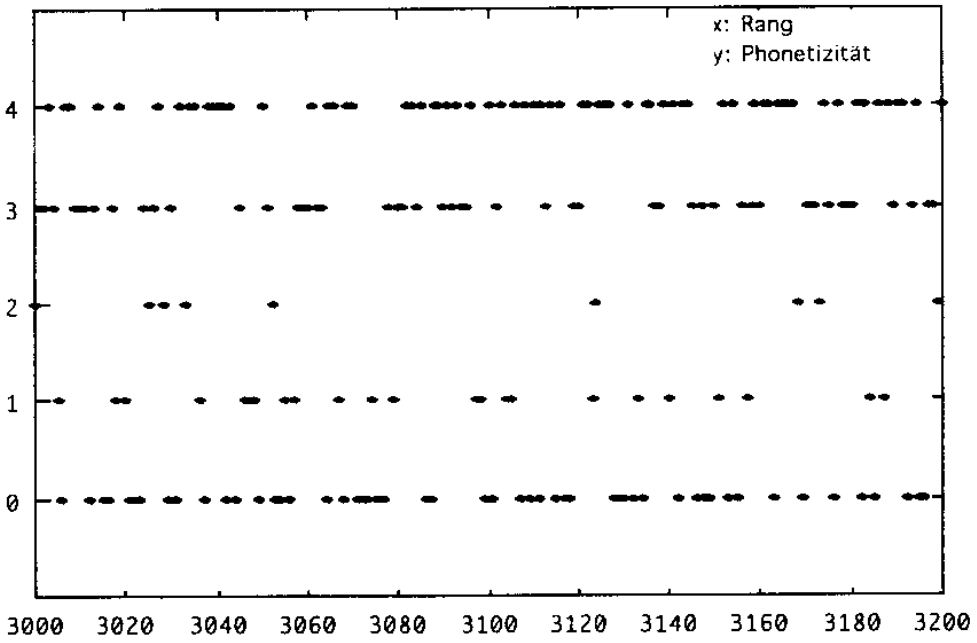


Abb. 5.2: Phonetizitätswerte der Ränge 3.000 bis 3.200 aller GB-Zeichen

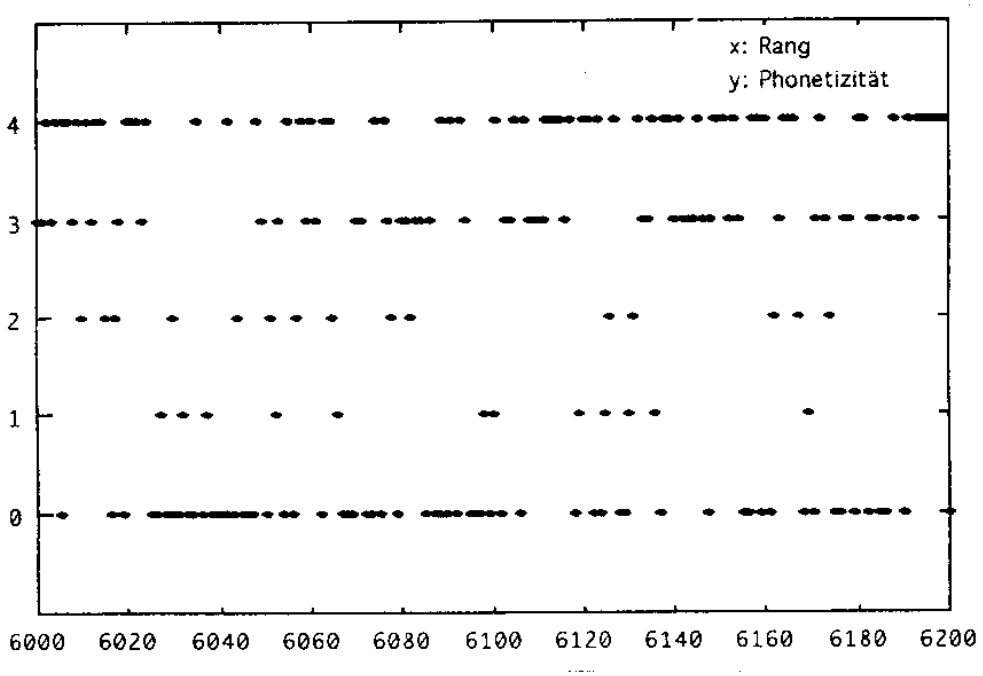


Abb. 5.3: Phonetizitätswerte der Ränge 6.000 bis 6.200 aller GB-Zeichen

Aus dem Vergleich dieser Abbildungen ist zunächst nur zu erkennen, dass in der Klasse der häufigsten 200 Zeichen der höchste Phonetizitätswert '4' häufiger auftritt als in den beiden anderen Häufigkeitsklassen. Dies kann als erster Hinweis

auf die Richtigkeit der Annahme, hochfrequente Zeichen seien ‚phonetischer' als niederfrequente, gewertet werden.

In Abbildung 5.4 wurde versucht, diesen eventuell in den Daten verborgenen Trend durch die Berechnung gleitender Mittelwerte (Fensterbreite: 200) sichtbar zu machen.

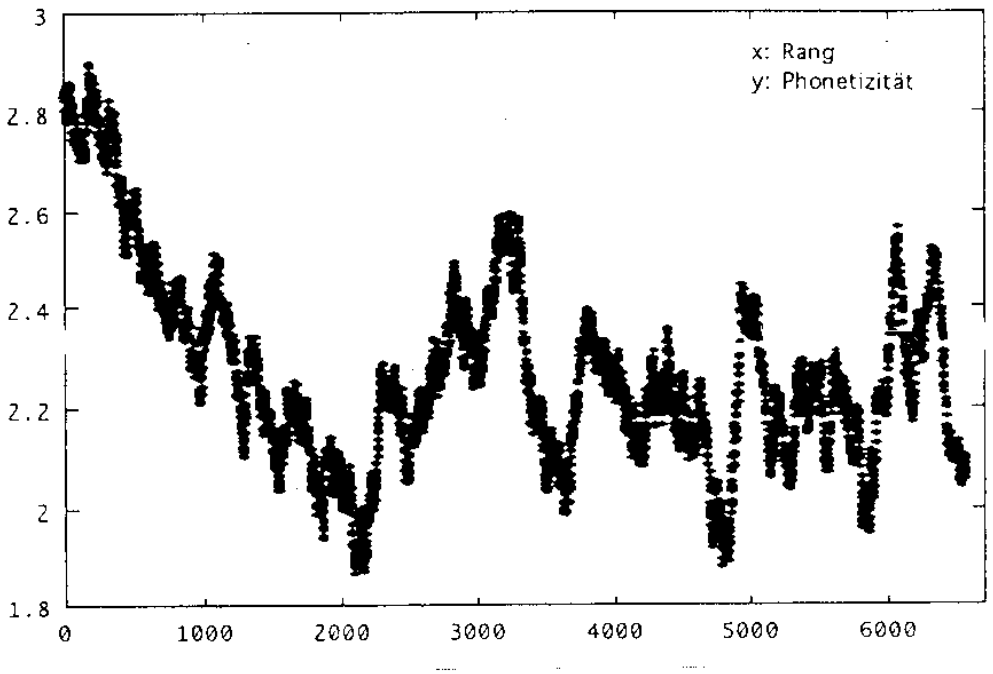


Abb. 5.4: Alle GB-Zeichen, gleitende Mittelwerte - Fensterbreite: 200

Im Bereich der 2.000 häufigsten Zeichen ist der erwartete Trend deutlich erkennbar - dieser Ausschnitt wird in Abbildung 5.5 noch einmal detaillierter dargestellt.

Im Bereich der selteneren Zeichen ist dagegen kein einheitlicher Verlauf der durch die gleitenden Mittelwerte berechneten Punktwolke zu erkennen.

Ein entscheidender Faktor für die uneinheitliche Phonetizitätsverteilung im Bereich niederfrequenter Zeichen könnte die Tatsache sein, dass mit abnehmender Frequenz immer mehr Zeichen dieselbe absolute Häufigkeit im Korpus besitzen, bei der Berechnung der gleitenden Mittelwerte aber genauso stark gewertet werden wie hochfrequente Zeichen. Es scheint demnach sinnvoll, die Phonetizität von Zeichen gleicher absoluter Häufigkeit zu einem Wert zusammenzufassen, das heißt das arithmetische Mittel der Phonetizitätswerte aller Zeichen gleicher Frequenz zu bilden und das Ergebnis als einen Rang zu werten.

Aus der graphischen Repräsentation der gemittelten Ränge in Abbildung 5.7 lässt sich keinerlei Rückschluss auf den Zusammenhang zwischen Frequenz und Phonetizität ziehen, so dass auch hier das Verfahren der gleitenden Mittelwerte

herangezogen werden muss, um in den Daten verborgene Tendenzen sichtbar zu machen.

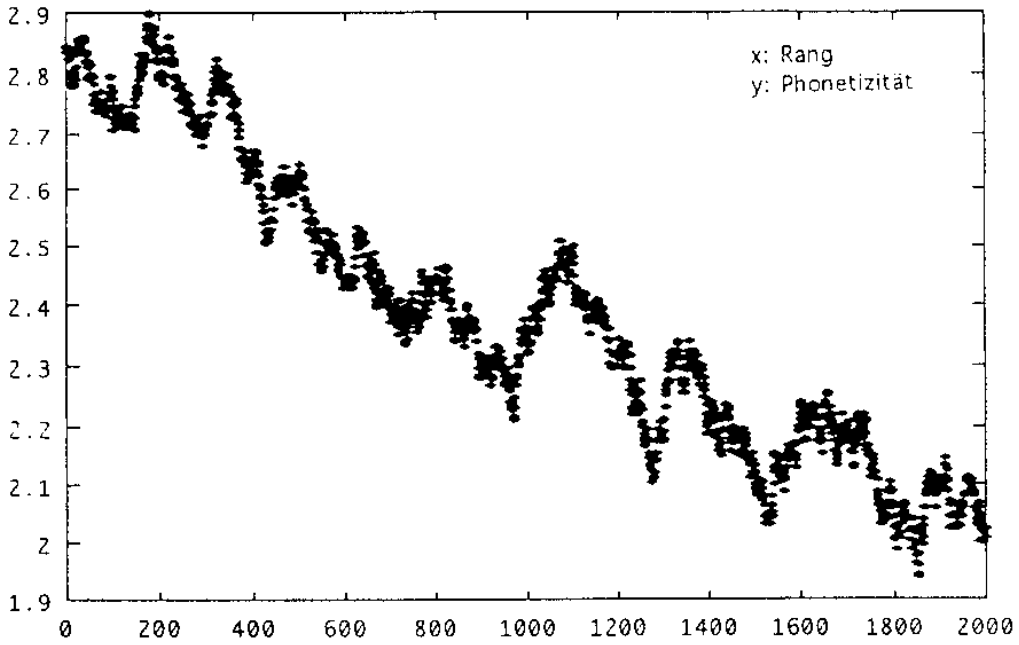


Abb. 5.5: Rang 0 bis 2.000 aller GB-Zeichen, gleitende Mittelwerte - Fensterbreite: 200

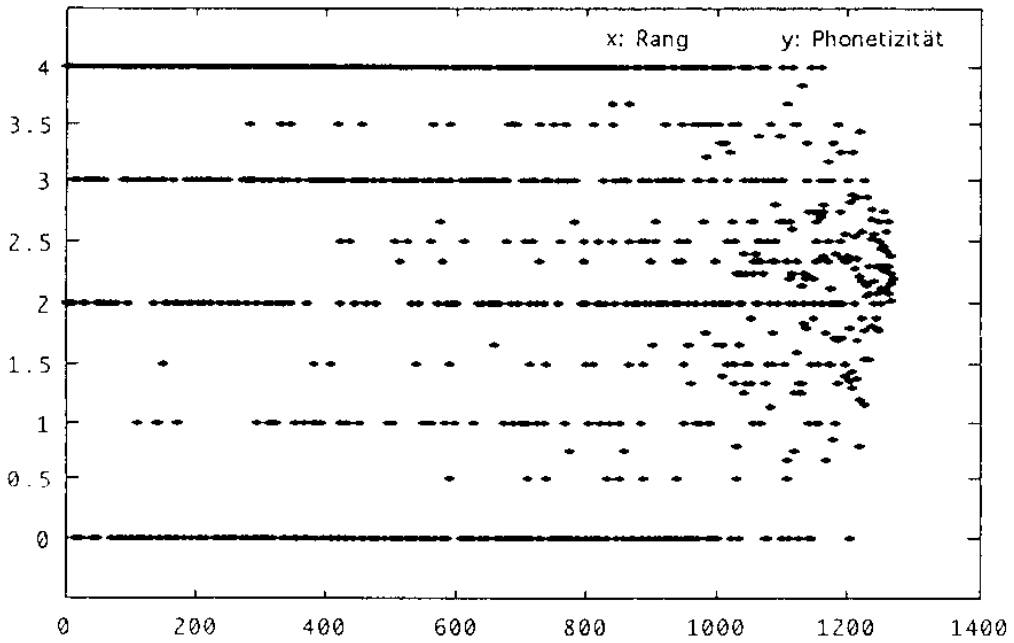


Abb. 5.7: Phonetizität aller GB-Zeichen, gemittelte Ränge

Das Ergebnis der Berechnung bei einer Fensterbreite von 100 bzw. 200 Datenpunkten wird in den Abbildungen 5.8 und 5.9 graphisch wiedergegeben und bestätigt deutlich den erwarteten Trend.

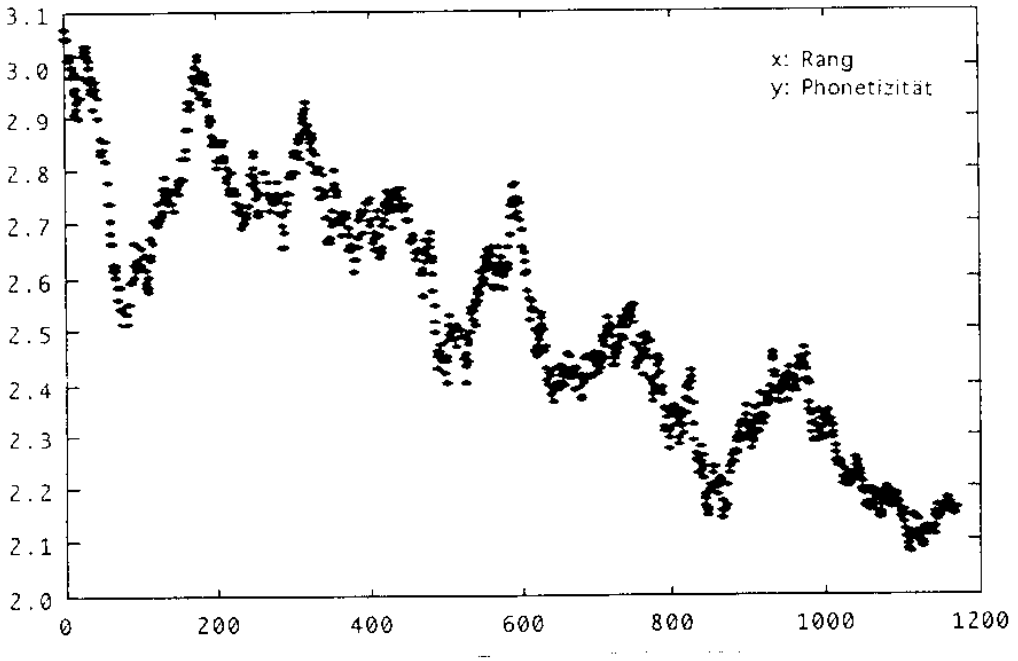


Abb. 5.8: Phonetizität aller GB-Zeichen - gleitende Mittelwerte (Fensterbreite 100)

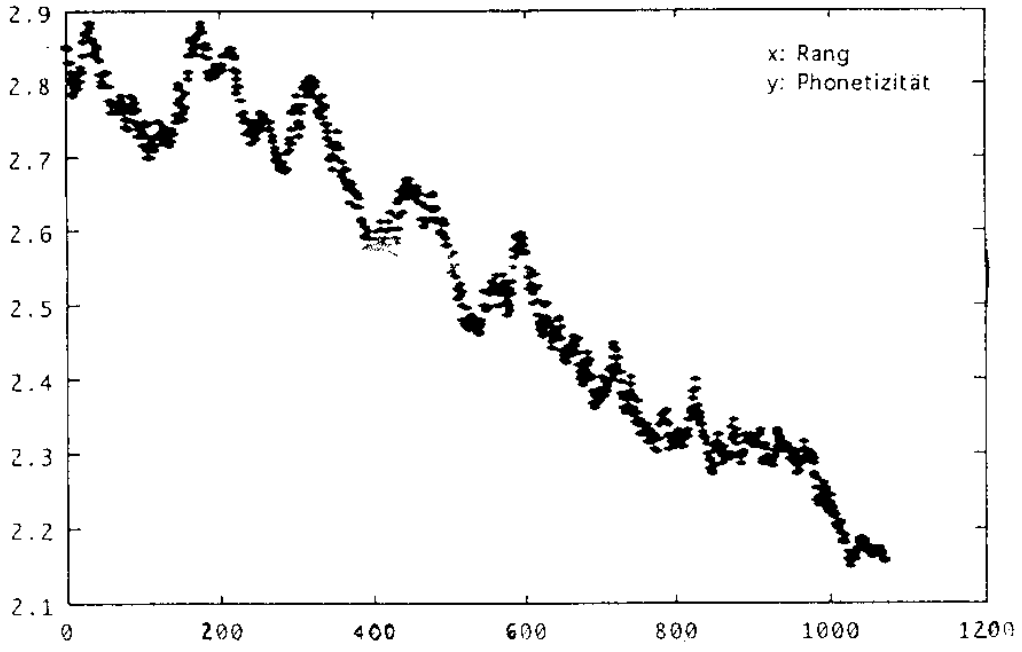


Abb. 5.9: Phonetizität aller GB-Zeichen - gleitende Mittelwerte (Fensterbreite 200)

5.5 Bewertung

Die Untersuchung hat gezeigt, dass tatsächlich ein Zusammenhang zwischen Schriftzeichenfrequenz und Phonetizität in der modernen chinesischen Schrift nachweisbar ist.

Wie erwartet handelt es sich bei diesem Zusammenhang jedoch nicht um einen deterministischen, so dass man etwa aufgrund einer gegebenen Frequenz eines Schriftzeichens zuverlässige Aussagen über dessen Phonetizität machen könnte. Die Beziehung zwischen Frequenz und Phonetizität hat vielmehr stochastischen Charakter, ist also nur als allgemeiner Trend über größere Datenmengen erfassbar.

Schließlich lässt sich das positive Untersuchungsergebnis als Bekräftigung des grundlegend phonetischen Charakters des chinesischen Schriftsystems deuten.

6. Literatur

Altmann, G. (1980). Prolegomena to Menzerath's Law. In R. Grotjahn (Hg.), *Glottometrika 2* (S. 1-10), Bochum: Brockmeyer.

Altmann, G. (1988). Verteilung der Satzlängen. In K.-P. Schulz (Hg.), *Glottometrika 9* (S. 147-169), Bochum: Brockmeyer.

Altmann, G. (1994). *Altmann-Fitter: Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen*. Lüdenscheid: RAM-Verlag.

Altmann, G, Beöthy, E., & Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 35/5, 537-543.

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

Arapov, M.V. (1982). A variational approach to frequency-rank distributions of text elements. In H. Guiter & M.V. Arapov (Hg.), *Studies on Zipf's Law* (S. 29-52), Bochum: Brockmeyer.

Best, K.-H., & Zhu, J. (1993). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit Ausblick auf das Chinesische). In U. Klenk (Hg.), *Computatio Linguae II.*, Stuttgart: Steiner.

DeFrancis, J. (1984). *The Chinese Language - Fact and Fantasy.* Honolulu: University of Hawaii Press.

Fickermann, I., Markner-Jäger, B., & Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In R. Köhler & J. Boy (Hg.), *Glottometrika 6* (S. 115-126), Bochum: Brockmeyer.

Gerlach, R. (1982). Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeld & U. Strauss (Hg.), *Glottometrika 4* (S. 95-102), Bochum: Brockmeyer.

Guo, J., & Lui, H. (1994). PH - A Chinese corpus for Pinyin-Hanzi transcription. *Chinesisch und Computer*, 9, 23-37.

Heups, G. (1983). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In R. Köhler & J. Boy (Hg.), *Glottometrika 5* (S. 113-133), Bochum: Brockmeyer.

HSZ (1988). Beijing Bibliothek (Hg.), *Hanzi shuxing zidian*. Beijing: Shumu wenxian chubanshe.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In W. Lehfeldt & U. Strauss (Hg.), *Glottometrika 4* (S. 103-113), Bochum: Brockmeyer.

Krott, A. (1994). Ein funktionalanalytisches Modell der Wortbildung. Unveröffentlichte Magisterarbeit. Trier.

Prün, C. (1994). Validity of Menzerath-Altmann's Law: Graphic Representation of Language, Information Processing Systems and Synergetic Linguistics. *Journal of Quantitative Linguistics*, 1.2, 148-155.

Rothe, U. (1983). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. In R. Köhler & J. Boy (Hg.), *Glottometrika 5* (S. 101-112), Bochum:Brockmeyer.

Sambor, J. (1984). Menzerath's Law and the Polysemy of Words. In R. Köhler & J. Boy (Hg.), *Glottometrika 6* (S. 94-114), Bochum: Brockmeyer.

Schwibbe, M.H. (1984). Text- und wortstatistische Untersuchungen zur Validität der Menzerathschen Regel. In R. Köhler & J. Boy (Hg.), Glottometrika 6 (S. 162-176), Bochum: Brockmeyer.

Stalph, J. (1989). *Grundlagen einer Grammatik der sinojapanischen Schrift*. Wiesbaden: Harrassowitz.

Teupenhayn, R., & Altmann, G. (1984). Clause Length and Menzerath's Law. In R. Köhler & J. Boy (Hg.), *Glottometrika 6* (S. 127-138), Bochum: Brockmeyer.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, 1.1, 98-106.

XHDC (1985). Fremdspracheninstitut Peking (Hg.). *Xin Han-Da cidian - Das neue Chinesisch-Deutsche Wörterbuch*. Beijing: Commercial Press.

Zhou, Y. (1978). Xiandai hanzi zhong shengpang de biaoyin wenti. *Zhongguo Yuwen*, 3, 172-177.

Zhou, Y. (1980). *Hanzi shengpang duyin biancha*. Changchun: Jilin renmin chubanshe.

Zhu, J., & Best, K.H. (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus*, 35, 45-60.

Anhang

Tabelle 1. Das Menzerathsche Gesetz auf Komponentenebene

x = Anzahl der Striche pro Komponente
n = Anzahl der Komponenten mit Strichzahl *x*
y empirisch = Durchschnittlicher Schreibaufwand der Striche in Komponenten mit Strichzahl *x*
y theoretisch = Erwarteter durchschnittlicher Schreibaufwand

x	*n*	*y empirisch*	*y theoretisch*
1	11	2.0909	2.0182
2	59	1.6016	1.6586
3	87	1.4291	1.4788
4	115	1.3173	1.3631
5	90	1.2577	1.2797
6	62	1.1935	1.2153
7	36	1.1984	1.1634
8	23	1.1413	1.1203
9	12	1.1759	1.0835
10	3	1.1333	-
11	2	1.1818	-

Tabelle 2. Das Menzerathsche Gesetz auf Schriftzeichenebene

x = Anzahl der Komponenten pro Schriftzeichen
n = Anzahl der Schriftzeichen mit Komponentenzahl *x*
y empirisch = Durchschnittliche Strichzahl der Komponenten in Schriftzeichen mit Komponentenzahl *x*
y theoretisch = Erwartete durchschnittliche Strichzahl

x	*n*	*y empirisch*	*y theoretisch*
1	298	4.7282	4.8513
2	1994	4.1166	3.9638
3	2565	3.5670	3.5220
4	1215	3.2835	3.2387
5	491	3.0159	3.0348
6	120	2.9889	2.8777
7	25	2.8286	2.7512
8	13	2.3077	2.6462
9	3	3.1111	-

Tabelle 3. Das Menzerathsche Gesetz auf Wortebene

x = Anzahl der Schriftzeichen pro Wort (Wortlänge)
n = Anzahl der Wörter mit Wortlänge *x*
y empirisch = Durchschnittliche Komponentenzahl der Zeichen in Wörtern mit Wortlänge *x*
y theoretisch = Erwartete durchschnittliche Komponentenzahl

x	*n*	*y empirisch*	*y theoretisch*
1	155	2.7032	2.6960
2	771	2.4371	2.4231
3	129	2.2171	2.2765
4	48	2.2187	2.1779
5	2	1.9000	-

Tabelle 4. Das Menzerathsche Gesetz auf Teilsatzebene (Korpus)

x = Anzahl der Wörter pro Teilsatz (Teilsatzlänge)
n = Anzahl der Teilsätze im Korpus mit Teilsatzlänge *x*
y empirisch = Durchschnittliche Schriftzeichenzahl der Wörter in Teilsätzen mit Teilsatzlänge *x*
y theoretisch = Erwartete durchschnittliche Schriftzeichenzahl

x	*n*	*y empirisch*	*y theoretisch*
1	6277	2.3361	1.9377
2	12239	1.7358	1.8251
3	14408	1.6996	1.7622
4	17423	1.6267	1.7190
5	18959	1.5975	1.6862
6	19686	1.5854	1.6598
7	19434	1.5653	1.6378
8	18377	1.5600	1.6190
9	16975	1.5396	1.6027
10	15050	1.5267	1.5881
11	13166	1.5236	1.5751
12	11353	1.5152	1.5633
13	9552	1.5093	1.5525
14	7953	1.5051	1.5426
15	6803	1.5034	1.5335
16	5441	1.5052	1.4249

x	*n*	*y empirisch*	*y theoretisch*
17	4663	1.4926	1.5170
18	3811	1.4933	1.5095
19	3085	1.4858	1.5024
20	2525	1.4851	1.4958
21	2099	1.4857	1.4895
22	1831	1.4860	1.4835
23	1459	1.4747	1.4778
24	1198	1.4708	1.4724
25	1017	1.4635	1.4672
26	841	1.4653	1.4623
27	706	1.4923	1.4575
28	597	1.4861	1.4529
29	446	1.4490	1.4485
30	376	1.4540	1.4443
31	323	1.4629	1.4402
32	253	1.4542	1.4363
33	242	1.4567	1.4325
34	176	1.4554	1.4288
35	134	1.4473	1.4252
36	103	1.4291	1.4217
37	101	1.4059	1.4184
38	73	1.4250	1.4151
39	74	1.4307	1.4119
40	56	1.4696	1.4088
41	42	1.4855	1.4058
42	50	1.4186	1.4029
43	43	1.4035	1.4001
44	29	1.3871	1.3973
45	31	1.3799	1.3946
46	26	1.4833	1.3919
47	19	1.4356	1.3893
48	22	1.4138	1.3868
49	20	1.4133	1.3844
50	14	1.3871	1.3819
51	17	1.4418	1.3796
52	8	1.4904	-
53	5	1.2830	-
54	2	1.2037	-
55	4	1.3864	-
56	7	1.4031	-

x	*n*	*y empirisch*	*y theoretisch*
57	3	1.6082	-
58	7	1.5000	-
59	4	1.2373	-
60	5	1.4533	-
61	4	1.5164	-
62	1	1.5161	-
63	4	1.2738	-
65	1	1.1846	-
66	2	1.5303	-
67	2	1.5224	-
68	3	1.2451	-
69	4	1.1775	-
70	1	1.0429	-
71	1	2.0563	-
72	2	1.2292	-
73	2	1.5342	-
75	3	1.2578	-
76	1	1.2368	-
78	2	1.1795	-
79	1	1.2785	-
80	2	1.5063	-
81	1	1.3704	-
86	2	1.0756	-
88	1	1.0341	-
90	1	1.5222	-
92	1	1.4565	-
94	1	1.4362	-
96	1	1.0104	-
104	1	1.0192	-
110	1	1.0273	-
121	2	1.0124	-
126	1	1.4127	-
130	1	1.4154	-
131	1	1.1756	-
141	1	1.5957	-
142	1	1.0352	-
146	1	1.0137	-
153	1	1.5948	-
481	1	1.1102	-

Tabelle 5. Das Menzerathsche Gesetz auf Teilsatzebene (Einzeltext)

x = Anzahl der Wörter pro Teilsatz (Teilsatzlänge)
n = Anzahl der Teilsätze im Text mit Teilsatzlänge *x*
y empirisch = Durchschnittliche Schriftzeichenzahl der Wörter in Teilsätzen mit Teilsatzlänge *x*
y theoretisch = Erwartete durchschnittliche Schriftzeichenzahl

x	*n*	*y empirisch*	*y theoretisch*
1	30	2.0000	1.9208
2	34	1.6324	1.7399
3	52	1.7051	1.6420
4	57	1.5219	1.5760
5	60	1.4833	1.5266
6	46	1.4384	1.4874
7	40	1.4821	1.4550
8	30	1.4333	1.4275
9	18	1.3580	1.4037
10	19	1.4368	1.3828
11	10	1.4364	1.3641
12	6	1.4583	-
13	6	1.4744	-
14	6	1.4881	-
15	3	1.4000	-
16	2	1.4375	-
17	1	1.5882	-
18	1	1.4444	-

Tabelle 6. Das Menzerathsche Gesetz auf Satzebene (Korpus)

x = Anzahl der Teilsätze pro Satz (Satzlänge)
n = Anzahl der Sätze im Korpus mit Satzlänge *x*
y empirisch = Durchschnittliche Wortzahl der Teilsätze in Sätzen mit Satzlänge *x*
y theoretisch = Erwartete durchschnittliche Wortzahl

x	*n*	*y empirisch*	*y theoretisch*
1	33875	10.9180	10.8069
2	29191	9.2937	9.2139
3	19704	8.2833	8.3933
4	10208	7.7265	7.8558
5	4590	7.3919	7.4627
6	2054	7.0763	7.1561
7	845	6.8629	6.9068
8	405	6.7438	6.6978
9	177	6.7332	6.5188
10	77	6.3182	6.3627
11	43	5.7780	6.2247
12	21	6.6032	6.1013
13	8	6.9615	-
14	3	4.7857	-
15	1	1.6000	-
16	2	5.2500	-
17	1	5.8824	-
18	2	5.1667	-
20	1	10.4000	-
22	1	5.2273	-
34	1	1.6765	-

Tabelle 7. Das Menzerathsche Gesetz auf Satzebene (Einzeltext)

x = Anzahl der Teilsätze pro Satz (Satzlänge)
n = Anzahl der Sätze im Text mit Satzlänge *x*
y empirisch = Durchschnittliche Wortzahl der Teilsätze in Sätzen mit Satzlänge *x*
y theoretisch = Erwartete durchschnittliche Wortzahl

x	*n*	*y empirisch*	*y theoretisch*
1	38	5.9737	6.1360
2	89	5.8258	5.6213
3	40	5.7583	5.3405
4	13	4.6923	5.1450
5	5	5.8000	-
8	1	5.2500	-

Tabelle 8. Komponentenkomplexität

x = Anzahl der Striche pro Komponente
f empirisch = Anzahl der Komponenten mit Strichzahl *x*
f theoretisch = Erwartete Anzahl der Komponenten mit Strichzahl *x* gemäß der Poisson-Verteilung

x	*f empirisch*	*f theoretisch*
1	11	14.47
2	59	51.26
3	87	90.79
4	115	107.22
5	90	94.95
6	62	67.28
7	36	39.72
8	23	20.10
9	12	8.90
10	3	3.50
11	2	1.85

Tabelle 9. Schriftzeichenkomplexität

x = Anzahl der Komponenten pro Schriftzeichen
f empirisch = Anzahl der Schriftzeichen mit Komponentenzahl *x*
f theoretisch = Erwartete Anzahl der Schriftzeichen mit Komponentenzahl *x* gemäß der Poisson-Verteilung

x	*f empirisch*	*f theoretisch*
1	298	1641.45
2	1994	2314.61
3	2565	1631.92
4	1215	767.06
5	491	270.41
6	120	76.26
7	25	17.92
8	13	3.61
9	3	0.78

Tabelle 10. Wortlängenverteilung

x = Wortlänge in Anzahl der Schriftzeichen
f empirisch = Anzahl der Wörter mit Wortlänge *x*
f theoretisch = Erwartete durchschnittliche Wortlänge (kein signifikantes Resultat)

x	*f empirisch*	*f theoretisch*
1	155	-
2	771	-
3	129	-
4	48	-
5	2	-

Tabelle 11. Teilsatzlängenverteilung (Korpus)

x = Teilsatzlänge in Anzahl der Wörter
f empirisch = Anzahl der Teilsätze mit Teilsatzlänge *x*
f theoretisch = Erwartete Teilsatzlänge gemäß negativer Binomialverteilung

x	*f empirisch*	*f theoretisch*
1	6277	7070.41
2	12239	12804.92
3	14408	16597.44
4	17423	18664.64
5	18959	19387.63
6	19686	19140.39
7	19434	18239.31
8	18377	16933.45
9	16975	15409.70
10	15050	13802.67
11	13166	12205.31
12	11353	10678.55
13	9552	9259.46
14	7953	7967.84
15	6803	6811.30
16	5441	5789.24
17	4663	4895.70
18	3811	4121.54
19	3085	3455.93
20	2525	2887.40
21	2099	2404.58
22	1831	1996.60
23	1459	1653.39
24	1198	1365.81
25	1017	1125.70
26	841	925.87
27	706	760.05
28	597	622.81
29	446	509.50
30	376	416.16
31	323	339.43
32	253	276.47
33	242	224.90
34	176	182.73

x	*f empirisch*	*f theoretisch*
35	134	148.30
36	103	120.23
37	101	97.37
38	73	78.79
39	74	63.69
40	56	51.44
41	42	41.51
42	50	33.48
43	43	26.97
44	29	21.72
45	31	17.48
46	26	14.05
47	19	11.29
48	22	9.07
49	20	7.28
50	14	5.84
51	17	4.68
52	8	3.75
53	5	3.00
54	2	2.41
55	4	1.92
56	7	1.54
57	3	1.23
58	7	0.98
59	4	0.78
60	5	0.62
61	4	0.50
62	1	0.40
63	4	0.32
65	1	0.25
66	2	0.20
67	2	0.16
68	3	0.12
69	4	0.10
70	1	0.08
71	1	0.06
72	2	0.05
73	2	0.04
75	3	0.03
76	1	0.02

x	*f empirisch*	*f theoretisch*
78	2	0.02
79	1	0.01
80	2	0.01
81	1	0.01
86	2	0.00
88	1	0.00
90	1	0.00
92	1	0.00
94	1	0.00
96	1	0.00
104	1	0.00
110	1	0.00
121	2	0.00
126	1	0.00
130	1	0.00
131	1	0.00
141	1	0.00
142	1	0.00
146	1	0.00
153	1	0.00
481	1	0.00

Tabelle 12. Teilsatzlängenverteilung (Einzeltext)

x = Teilsatzlänge in Anzahl der Wörter
f empirisch = Anzahl der Teilsätze mitTeilsatzlänge *x*
f theoretisch = Erwartete Teilsatzlänge gemäß negativer Binomialverteilung

x	*f empirisch*	*f theoretisch*
1	30	20.65
2	34	42.09
3	52	54.87
4	57	58.10
5	60	54.40
6	46	46.94
7	40	38.20
8	30	29.75
9	18	22.39
10	19	16.40
11	10	11.74

x	*f empirisch*	*f theoretisch*
12	6	8.25
13	6	5.70
14	6	3.89
15	3	2.62
16	2	1.75
17	1	1.16
18	1	2.13

Tabelle 13. Satzlängenverteilung (Korpus)

x = Satzlänge in Anzahl der Teilsätze
f empirisch = Anzahl der Sätze mit Satzlänge *x*
f theoretisch = Erwartete Satzlänge gemäß negativer Binomialverteilung

x	*f empirisch*	*f theoretisch*
1	33875	33724.82
2	29191	30064.56
3	19704	18695.36
4	10208	9945.27
5	4590	4843.62
6	2054	2228.38
7	845	985.14
8	405	422.87
9	177	177.45
10	77	73.13
11	43	29.70
12	21	11.92
13	8	4.73
14	3	1.86
15	1	0.73
16	2	0.28
17	1	0.11
18	2	0.04
20	1	0.02
22	1	0.01
34	1	0.00

Tabelle 14. Satzlängenverteilung (Einzeltext)

x = Satzlänge in Anzahl der Teilsätze

f empirisch = Anzahl der Sätze mit Satzlänge *x*
f theoretisch = Erwartete Satzlänge gemäß negativer Binomialverteilung

x	*f empirisch*	*f theoretisch*
1	38	55.60
2	89	65.48
3	40	40.16
4	13	17.08
5	5	5.66
8	1	2.04