Flut von Texten – Vielfalt der Kulturen

Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie

Herausgegeben von

Harald Burger, Annelies Häcki Buhofer, Gertrud Gréciano

Phraseologie und Parömiologie ; Bd. 14
ISBN 3-89676-677-5
Schneider Verlag Hohengehren, Baltmannsweiler 2003.

Schneider Verlag Hohengehren GmbH

Peter Grzybek, Graz

Zur lexikalischen Struktur von Sprichwörtern

1. Einleitung

Wie an anderer Stelle mehrfach moniert wurde, ist die sprachliche Organisation von Sprichwörtern kaum jemals wirklich ernsthaft im Hinblick auf Gesetzmäßigkeiten untersucht worden. Auch wenn dieses Pauschalurteil auf den ersten Blick vielleicht provokant erscheinen mag, so lässt es sich doch damit begründen, dass die Suche nach sprachlichen Regularitäten von Sprichwörtern im Grunde genommen kaum jemals über eine symptomatische Betrachtungsebene hinaus gekommen ist und eine systematische Ebene bislang eigentlich nicht erreicht hat (Grzybek 2000a,b, 2001, 2002).

Erste Versuche, die in diesem Bereich evidenten Forschungslücken zu füllen, haben sich naturgemäß auf eher spezielle Probleme konzentriert, so z.B. auf die Frage, ob sich die Satz- und Wortlänge von Sprichwörtern als gesetzmäßig beschreiben lässt. Besonderes Augenmerk wurde bei der Behandlung dieser Frage auf die Entwicklung geeigneter Methoden gerichtet, die über die traditionelle Herangehensweisen wie der Angabe von absoluten und/oder relativen Häufigkeiten, der Berechnung von Mittelwerten, der Erstellung einfacher Graphiken, o.ä. hinausgehen. Dabei ging es im Detail u.a. um die auf den ersten Blick trivial anmutende Frage, wie oft in einem Korpus von Sprichwörtern Wörter bzw. Sätze mit einer bestimmten Länge vorkommen, und ob sich diese Vorkommenshäufigkeiten durch Bezugnahme auf mathematische und/oder statistische Verfahren formalisieren lassen. Ungeachtet der scheinbaren Einfachheit der genannten Fragestellung liegen dieser Problematik eine Reihe von aufeinander aufbauenden Grundannahmen zugrunde, was auf einen zweiten Blick die dahinter stehende Komplexität transparent macht, nämlich die Annahmen,

- dass die Häufigkeit, mit der Wörter bzw. Sprichwörter einer bestimmten Länge in einem solchen Sprichwortkorpus enthalten sind, nicht zufällig (chaotisch) ist, sondern einer bestimmten Regel- oder Gesetzmäßigkeit folgt;
- dass sich diese Regelmäßigkeit nicht nur beschreiben, sondern auch formalisieren lässt;
- dass sich diese Regelmäßigkeit so formalisieren lässt, dass sich Querverbindungen zu allgemeinen (nicht nur auf Sprichwörter bezogenen) Untersuchungen und theoretischen Annahmen zur Satzlängenverteilung herstellen lassen;
- dass diese Querverbindungen Aussagen über die Spezifik von Sprichwörtern bzw. von sprichwörtlichen Sätzen erlauben.

Um den aufgeworfenen Fragen nachzugehen, wurde auf Vorschläge Bezug genommen, die in den vergangenen Jahren in Bereich der allgemeinen (quantitativen) Linguistik unterbreitet wurden, um solche Gesetzmäßigkeiten zu erfassen und im Rahmen eines synergetischen Ansatzes beschreiben zu können (Wimmer et al. 1994, Wimmer/Altmann 1996).

In den vorliegenden Überlegungen soll es darum gehen, weiteren Gesetzmäßigkeiten des Sprichworts nachzuspüren. Abermals wird sich dabei Bezug auf Ansätze der Quantitativen Linguistik nehmen lassen, auch wenn es um eine anders gelagerte Fragestellung geht. Im Vordergrund der vorliegenden Abhandlung soll es nämlich um die Frage gehen, ob sich das lexikalische Inventar eines gegebenen Sprichwortkorpus durch bestimmte Regularitäten der Häufigkeitsstruktur auszeichnet. Konkret lautet die Frage, ob sich die Häufigkeiten, mit der die einzelnen Wörter in einem Sprichwortkorpus vorkommen, theoretisch modellieren lassen. Während es also in den oben erwähnten Untersuchungen um die Länge der zur Disposition stehenden Einheiten (und deren Vorkommenshäufigkeiten) ging, steht im Vordergrund der folgenden Überlegungen nicht mehr und nicht weniger als die Frage nach der lexikalischen Häufigkeitsstruktur an und für sich.

2. Material

Als Untersuchungsmaterial soll – wie zum Teil auch schon in den o.a. Untersuchungen – die Sammlung slowenischer Sprichwörter *Pregovori, prilike in reki* von Kocbek (1887) dienen. Diese stellt die eigentlich erste umfassende, eigenständige Sprichwortsammlung des Slowenischen dar; natürlich beruht dieses Werk auf anderen vorbereitenden Arbeiten und kleineren, zuvor an unterschiedlichen Stellen veröffentlichten Sammlungen. Dennoch ist es nicht nur als ein geschichtlicher Meilenstein, sondern auch als Fundament der späteren slowenischen Parömiographie schlechthin anzusehen, das – zumindest zu einem großen Teil – auch die Grundlage für spätere Sammlungen darstellt, und zwar nicht nur für die Erweiterung von Kocbek/Šašelj (1934), sondern auch für "modernere" Sammlungen wie diejenigen von Bojc (1974 u.a.) oder Prek (1970 u.a.).

Die Sammlung von Kocbek besteht aus insgesamt 2.429 sprichwörtlichen Sätzen bei einem lexikalischen Umfang von 15.467 Wortformen. Bei diesen 15.467 handelt es sich um individuelle Vorkommnisse (Tokens), die auf 4.638 verschiedenen (Types) basieren. Die Beobachtung, dass 2.887 genau einmal vorkommen – hierbei handelt es sich um die sogenannten 'hapax legomena', die somit 18,66% der Tokens bzw. 62,25% der Types betragen – führt zur Frage einer systematischen Betrachtung der Vorkommenshäufigkeit aller Wortformen.

Gehen wir davon aus, dass jede der insgesamt n Wortformen (hier: $n = 15.467$) unterschiedlich häufig vorkommen kann, und zwar mindestens einmal und maximal m-mal. Bezeichnen wir weiterhin die konkrete Vorkommenshäufigkeit einer Wortform als i, so gilt $1 \leq i \leq m$. Im Falle unseres Sprichwortkorpus die maximale Vorkommenshäufigkeit einer Wortform $m = 487$, so dass $1 \leq i \leq 487$. Jede dieser (verschiedenen) Vorkommenshäufigkeiten i kann nun durch eine unterschiedliche Anzahl von Wortformen repräsentiert sein; bezeichnen wird diese Anzahl als f_i. In unserem Fall beträgt $f_i = 2.887$ für $i = 1$ (weil es 2.887 Wortformen gibt, die einmal vorkommen), und $f_i = 1$ für die maximale Vorkommenshäufigkeit von $i = $ m (in unserem Fall 487).

Entsprechend berechnet sich auch die **Anzahl der verschiedenen Wortformen** (d.h. der Types) als:

$$\sum_{i=1}^{m} f_i = 4.638,$$

und die **absolute Vorkommenshäufigkeit** aller Wortformen (d.h. der Tokens) als:

$$\sum_{i=1}^{m} i \cdot f_i = 15.467.$$

Tab. 1 fasst diese Voraussetzungen zum Zwecke der Demonstration auszugsweise zusammen:

Tab. 1: Struktur der lexikalischen Vorkommenshäufigkeiten

i	1	2	3	4	5	...	455	487	
f_i	2887	732	314				1	1	Σ 4.638
f_i	2887	1464	942				455	487	Σ 15.467

Vor diesem Hintergrund erscheinen nun auch übliche Fragen wie die nach den häufigsten Wortformen in systematischem Licht; diese zielen nämlich im Prinzip auf eine Ranghäufigkeitstabelle, in der die mit $f_i > 0$ besetzten Wortformen (i) in absteigende Reihenfolge gebracht werden. Die hieraus resultierende Rangreihenfolge der Ränge r (r = 1,2,...m) ist in Tab. 2 für die 10 häufigsten Wortformen unseres Sprichwortkorpus dargestellt; enthalten sind neben den konkreten Wortformen deren absolute (f_r) und relative (p_r) Vorkommenshäufigkeiten – es sind in unserem Fall all diejenigen Wortformen, die eine Frequenz von $f_r > 100$ aufweisen.

Tab. 2: Liste der 10 häufigsten Wortformen

		f_r	p_r
1	ne	487	0,0315
2	je	455	0,0294
3	se	381	0,0246
4	kdor	264	0,0171
5	v	241	0,0156
6	na	195	0,0126
7	pa	125	0,0081
8	za	109	0,0070
9	in	106	0,0069
10	ima	101	0,0065
...	...	...	...

Es soll hier nicht um einen inhaltlichen Vergleich des lexikalischen Bestands der Sprichwörtersammlung mit einer Wortfrequenzliste auf der Basis einer Textkorpusanalyse gehen; deshalb sei nur en passant erwähnt, dass acht der hier aufgeführten zehn frequentesten Wörter auch zu den "Top Ten" der Frequenzliste aus dem elektronischen Korpus slowenischer

Texte (CORTES)[1] gehörten – lediglich die Wortformen 'kdor' [wenn] und 'ima' [er/sie hat] weisen dort eine deutlich niedrigere Frequenz auf (Rang 330 respektive Rang 121). Das lässt auf der einen Seite den Schluss zu, dass sich sprichwörtliches Wortmaterial im Vergleich zu Wortfrequenzlisten durchaus durch bestimmte Spezifika auszeichnet, andererseits aber Regularitäten folgt, die durch allgemeine sprachliche Gesetzmäßigkeiten geprägt sind. In diesem Sinne soll im folgenden der Frage lexikalischer Regularitäten sprichwörtlichen Materials nachgegangen werden. Konkret soll es dabei um zwei de facto miteinander verwobene Fragestellungen gehen:

1. Steht die Häufigkeit, mit der die häufigste Wortform vorkommt, in einer spezifischen Beziehung zu der Häufigkeit, mit der die zweithäufigste Wortform vorkommt, steht diese ihrerseits in einer spezifischen Beziehung zur dritthäufigsten, usw. Mathematisch umformuliert: Lässt sich die Häufigkeit einer gegebenen Klasse zur Häufigkeit der jeweiligen Nachbarklasse in Form einer Relation $P_x \sim P_{x-1}$ (d.h. als dynamisches System) verstehen, nimmt diese Relation gegebenenfalls die Form einer spezifischen Funktion $P_x = g(x)\ P_{x-1}$ an und – wenn ja – um welche Art von Funktion handelt es sich?
2. Wenn man weiß, welchen (prozentualen) Anteil die am häufigsten, am zweit-, dritt- usw. häufigsten vorkommenden Wortformen am gesamten lexikalischen Inventar haben, und wenn man diese Häufigkeiten von der am häufigsten bis zur am seltensten vorkommenden Wortform Schritt für Schritt kumulativ aufaddiert, dann lässt sich sagen, welchen Anteil am lexikalischen Inventar die häufigste und zweithäufigste Wortform, die häufigste, zweit- und dritthäufigste usw. zusammen haben. Die sich daraus ergebende Frage lautet: Ist auch die stetige Zunahme dieser kumulierten relativen Häufigkeiten (deren Summe 1 betragen muss) gesetzmäßigen Charakters, und lässt sich die allfällig zu beobachtende Regularität ebenfalls formalisieren?

Beginnen wir unsere Untersuchungen mit der ersten Frage. Es handelt sich hierbei um eine Problemstellung, die in der Quantitativen Linguistik seit

[1] Materialbasis ist hierbei eine Worthäufigkeitsliste der 1000 frequentesten slowenischen Wörter aus dem CORTES-Korpus, das Primož Jakopin zusammengestellt hat. Dieses Korpus bestand zum Zeitpunkt der Analyse (Dezember 1999) aus 112 literarischen (überwiegend prosaischen) Texten aus dem 19. und 20. Jhd. Die Texte stammen von 41 verschiedenen Autor(inn)en, wobei es sich in 98 Fällen um original slowenische Texte, in 14 Fällen um Übersetzungen ins Slowenische handelt.

den 30er und 40er Jahren des 20. Jhd.s mit dem Namen von George Kingsley Zipf verbunden ist, auf dessen Arbeiten folglich kurz einzugehen ist.

3. Zipf und Mandelbrot

In seinem 1935 publizierten Buch *The Psycho-Biology of Language* mit dem bezeichnenden Untertitel *An Introduction to Dynamic Philology* begründete Zipf eine erste Konzeption der Wortvorkommenshäufigkeit; hier lieferte er u.a. Argumente dafür, dass die Vorkommenshäufigkeit von Wörtern in Texten nicht zufällig, sondern gesetzmäßigen Charakters ist. Genauer gesagt, hatte er einen Zusammenhang zwischen der Vorkommenshäufigkeit eines Wortes und der Anzahl von Wörtern, die diese Häufigkeit aufweisen, beobachtet: Demnach sind es in einem Text relativ wenige Wörter, die häufig vorkommen, und es sind relativ viele Wörter, die selten vorkommen. Diese Beobachtung über die Abnahme der Variabilität bei zunehmender Frequenz hatte Zipf mit der Annahme verbunden, dass dieses Wechselverhältnis gesetzmäßigen Charakters ist, und er hatte versucht, es mit einer relativ einfachen Gleichung mathematisch zu formulieren:

$$(1)\ a \cdot b^2 = k\,.$$

In dieser Formel entspricht a der Anzahl von Wörtern mit einer bestimmten Vorkommenshäufigkeit, b entspricht der jeweiligen Anzahl der Vorkommnisse, und k ist eine (für den gegebenen Text charakteristische) Konstante. Der genannten Formel zufolge würde sich also das Produkt des Quadrats der Vorkommenshäufigkeit eines bestimmten Wortes und der Summe seiner Vorkommenshäufigkeiten als eine Konstante darstellen.

Seine Berechnungen hat Zipf u.a. an den Daten einer Untersuchung von Eldridge (1911) zur Worthäufigkeit im amerikanischen Zeitungsenglisch veranschaulicht, die insgesamt ca. 44.000 Wörter (ca. 6.000 verschiedene Wörter) umfasste. In diesem Korpus kamen 2.976 Wörter genau einmal, 1.079 Wörter zweimal, 516 Wörter dreimal, usw. vor; ein Wort wie der Artikel 'the' hingegen kam 4.290 mal vor. Aus Gründen der Anschaulichkeit hat Zipf bei der Überführung der Daten in eine Graphik die (beobachteten und theoretischen) Werte logarithmiert. Abb. 1 stellt das recht überzeugende Ergebnis anschaulich dar.

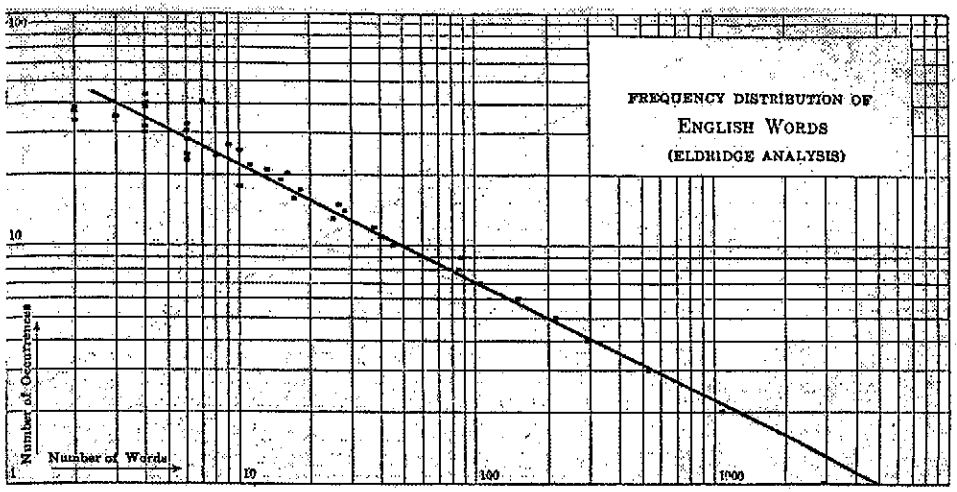

Abb. 1: Häufigkeitsverteilung von Wörtern nach Zipf (1935: 46)

Bei einem späteren Versuch der Formalisierung seiner Beobachtungen bezog Zipf (1949) im Gegensatz zu seiner früheren Herangehensweise zwar nach wie vor die absolute Vorkommenshäufigkeit (f) der Elemente in seine Formel ein, in Ergänzung dazu aber nun den Rang (r), den ein Wort innerhalb der untersuchten Textmenge mit einer bestimmten Häufigkeit einnimmt. Dies führte zu der (ebenfalls noch recht einfachen) Formel

$$(2)\ r \cdot f = k\,.$$

Diesem Ansatz zufolge erweist sich nunmehr das Produkt der absoluten Vorkommenshäufigkeit eines Wortes und seines Ranges als eine Konstante. Auch das Ergebnis dieser Überlegungen lässt sich an einem Beispiel veranschaulichen, und zwar an den oben bereits genannten und dargestellten Daten aus dem Eldridge-Korpus. Die Berechnung nach Formel (2) ergibt nach bilogarithmischer Transformation (s.o.) die Abb.2.

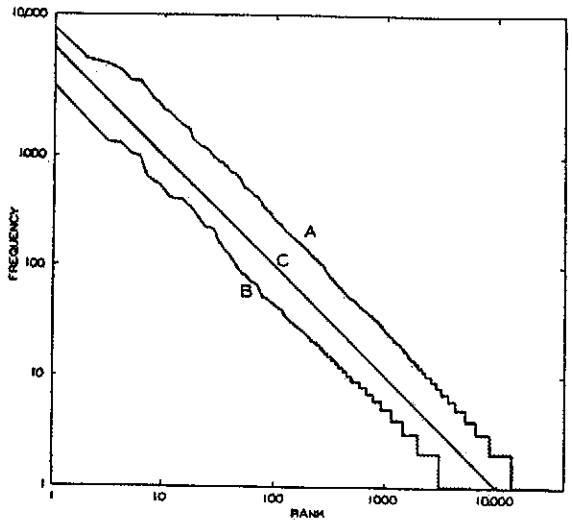

Abb. 2: (A) James Joyce; (B) Eldrige; (C) "Idealkurve" (Zipf 1949: 25)

Wie zu sehen ist, führen die Berechnungen offenbar zu überzeugenden Anpassungen; dies gilt allerdings, wie schon Zipf (1935: 43) selbst einräumte, für die besonders selten (und, wie im Anschluss an Zipf später von anderen Forschern bemängelt wurde) für die besonders häufig vorkommenden Wörter nur in eingeschränktem Umfang. Aus diesem Grunde ist der Zipf'sche Ansatz verschiedentlich modifiziert, erweitert und in allgemeinere Verteilungsmodelle überführt worden. Die hierbei erarbeiteten Modelle erwecken mitunter den Eindruck einer eigenen Wissenschaft, zu deren Verständnis ein mathematisches Grundstudium notwendig ist (vgl. z.B. Guiter / Arapov 1982, Baayen 2001, u.a.). Auf Sprache und sprachliche Texte bezogen, dürfte eine der wichtigsten und bekanntesten Ergänzungen wohl diejenige von Bénoit Mandelbrot (1953, 1954) gewesen sein, der an die (auch von Zipf konzedierte) Beobachtung anknüpfte, dass die von ihm beschriebene Gesetzmäßigkeit wohl für den "mittleren" Bereich des Vokabulars zutrifft, nicht aber für die extremen (d.h. die besonders häufigen und die besonders seltenen) Werte – und genau das zeigen ja auch die obigen Graphiken. Mandelbrot ging zunächst von der einfachen Zipf'schen Formel (2)

$$(2)\ r \cdot f = C$$

aus, die sich in geringfügiger Umformulierung als

$$(2a)\ f = \frac{C}{r}$$

lesen lässt. In weiterer Folge lässt sich die absolute Vorkommenshäufigkeit als eine Funktion des Rangs r und der Konstante C verstehen; definiert man nun den Rang als Variable x, erhält man die Funktion

$$(2b)\ f(x) = \frac{C}{x}.$$

Diese Funktion lässt sich leicht in eine Wahrscheinlichkeitsfunktion überführen, so dass sich für jedes x dessen theoretische Vorkommenshäufigkeit P_x berechnen lässt. Zu diesem Zweck muß (2b) rechts gestutzt werden – weil die harmonische Reihe nicht konvergiert – d.h. der Definitionsbereich wird $x = 1,2,3,\ldots,$ n. Hierbei wird C zu einer Normierungskonstante, definiert als $C^{-1} = \psi(n+1) - \psi(1)$, wo ψ die Digamma-Funktion (d.h. die logarithmierte Gamma-Funktion) ist, d.h.

$$(2b')\ P_x = \frac{1}{x[\Psi(n+1) - \Psi(1)]} \qquad x = 1,2,3, \ldots, n.$$

Diese Verteilung heißt Estoup-Verteilung (s. Wimmer / Altmann 1999). Eine Verallgemeinerung von (2b) ist

$$(2c)\ P_x = \frac{x^{-a}}{T(a)} \qquad x = 1,2,3, \ldots, n;\ a \in \Re; \qquad T(a) = \sum_{j=1}^{n} j^{-a}$$

Diese Wahrscheinlichkeitsfunktion wird gegenwärtig als die "klassische" Zipf-Verteilung (oder auch rechts gestutzte Zeta-Verteilung) bezeichnet, die sich von der ursprünglichen Zipf'schen Annahme in zweierlei Hinsicht unterscheidet: a) die Normierungskonstante C lässt sich nicht in geschlossener Form darstellen, b) der Exponent a ist $\neq 1$, während in der ursprünglichen Form gerade $a = 1$ war. Im Vergleich zur Formel (2c) beinhaltet die heute unter der Bezeichnung "Zipf-Mandelbrot-Verteilung" bekannte Verteilung eine zusätzliche Erweiterung um den Parameter b; somit ergibt sich:

$$(3)\ P(x) = \frac{C}{(b+x)^a} \qquad x = 1,2,3,\ldots,n.$$

Bei genauerem Hinsehen zeigt sich, dass die ursprüngliche Zipf'sche Formulierung (2b) sich in Form einer Wahrscheinlichkeitsfunktion (2b') als Spezialfall der *ZM*-Formel (3) darstellt, für den Fall nämlich, dass $a = 1$ und $b = 0$.

In der Anwendung gilt auch hier, dass die Parameter a und b bei einzelnen konkreten Texten variieren; dabei bleibt das Modell insgesamt unverändert, auch wenn sich der Wert für P_x je nach Variation der Parameter ändert. Die Konstante C erweist sich innerhalb der Wahrscheinlichkeitsfunktion (3) als eine Normierungskonstante, die sich aus den Parametern a und b bestimmen lässt:

$$(3')\ C^{-1} = \sum_{i=1}^{n} (b+i)^a$$

Damit ergibt sich insgesamt ein Verteilungsmodell, in dem sich für jedes P_x die theoretische Häufigkeit berechnen lässt. Dieses Modell – das sich in der Praxis keineswegs nur für sprachliche Erscheinungen als geeignet erwiesen hat – ist in der Sprach- und Textwissenschaft in erster Linie für Worthäufigkeiten von (vor allem längeren) Texten sowie von Wortfrequenzlisten bzw. Häufigkeitswörterbüchern auf der Basis von Textkorpora zur Geltung gekommen. Im folgenden soll untersucht werden, ob sich der theoretische Ansatz von Zipf bzw. Zipf/Mandelbrot auch zur Beschreibung der lexikalischen Struktur von Sprichwörtern fruchtbar machen lässt. Die Übertragung dieses Ansatzes auf Sprichwortmaterial ist ein Novum und alles andere als selbstverständlich; vermutlich ist ein solcher Versuch auch nicht unumstritten, denn bei einer Sprichwortsammlung handelt es sich weder um einen homogenen Text (insofern jedes Sprichwort im Grunde genommen ein eigener in sich geschlossener Text ist) noch um ein auf einer Textsammlung im allgemeinen Sinne basierendes Lexikon. Allerdings ließe sich dafür argumentieren, eine Sprichwortsammlung eben als Sprichwort-Lexikon zu verstehen, dessen Lemmata auf der Ebene des Satzes, nicht des Lexems zu definieren wären. Wenn jedoch die Anwendung des Zipf'schen Ansatzes auf unser Sprichwortmaterial gelänge, so wäre das nicht nur ein starkes Argument für die systematische Organisation der sprichwörtlichen Lexik, sondern auch ein weiterer Nachweis der gesetzmäßigen Organisation von Sprichwörtern überhaupt.

4. Analysen

Vor dem Hintergrund der oben dargelegten Überlegungen sollen im folgenden der Reihe nach drei Fragen an unser Sprichwortmaterial gerichtet werden:

- *Lexikalische Rangverteilung* (4.1.): Wie oft kommt die häufigste, zweithäufigste, dritthäufigste, usw. Wortform vor?
- *Lexikalisches Frequenzspektrum* (4.2.): Wie viele Wortformen kommen im Korpus jeweils genau 1, 2, 3, ...*n* mal vor?
- *Lexikalische Deckung* (4.3.): Welchen (prozentualen) Anteil am gesamten lexikalischen Bestand nehmen die häufigste, die beiden häufigsten, die drei häufigsten usw. Wortformen ein?

Natürlich soll es in allen drei Fragen nicht nur darum gehen, die empirischen Werte darzustellen, sondern vor allem auch darum zu untersuchen, ob bzw. wie sich die Werte mit theoretischen Modellen beschreiben lassen. Dabei ist davon auszugehen, dass die drei Fragen inhaltlich eng miteinander in Beziehung stehen, zumal sich die Rangverteilung mathematisch sowohl in das Frequenzspektrum als auch in die lexikalische Deckung überführen lässt. Im hier gegebenen Kontext soll jedoch der Schwerpunkt auf den einzelnen Fragen liegen, damit überhaupt erst einmal in einem ersten Schritt die Relevanz des Zipf-Mandelbrotschen Ansatzes auch für die Sprichwortforschung transparent gemacht werden kann. Beginnen wir mit der Rangverteilung, die im Hinblick auf die Zipf'schen Überlegungen so etwas wie einen Ausgangspunkt darstellen.

4.1. Rangverteilung

Die Frage nach der absoluten Vorkommenshäufigkeit (f_i) der häufigsten, zweithäufigsten, dritthäufigsten usw. Wortform ($i = 1,2,...,n$) beinhaltet im wesentlichen die Zipf'schen Überlegungen nach einer lexikalischen Ranghäufigkeitsverteilung (s.o.). Da es insgesamt nicht weniger als 15.467 Rangpositionen (i) gibt, wurde die gesamte Verteilung am Median geteilt, so dass wir es mit einer rechts-gestutzten Verteilung der oberen 50% der Wortformen zu tun haben, die sich auf die ersten 241 Ränge verteilen.

Tab. 3 enthält die Daten für die ersten 50 der 241 Ränge: in der ersten Spalte findet sich der jeweilige Rang (i), in der zweiten die jeweilige absolute Vorkommenshäufigkeit (f_i). In der dritten Spalte finden sich die theoretischen Werte (NP_i), die sich aufgrund der Anpassung der Zipf-Mandelbrot-Verteilung an die Daten ergeben. Die Güte der Anpassung wird in der Regel mit einem sogenannten χ^2-Anpassungstest geprüft. Da

dieser χ^2-Anpassungstest jedoch bei großen Stichproben (mit denen man bei sprachlichem Material oft zu tun hat) relativ schnell signifikant wird, verwendet man bei Stichproben mit großem N statt dessen auch den als χ^2 / N berechneten Diskrepanzkoeffizienten C; dieser Kontingenzkoeffizient wird bei $C < 0.02$ als Indiz einer guten, bei $C < 0.01$ als Indiz einer sehr guten Anpassung angesehen – in diesem Fall geht man somit mit anderen Worten, davon aus, dass die theoretische Berechnung geeignet ist, die empirisch ermittelten Werte in einem allgemeinen Modell zu erfassen. Insofern ist die Anpassung der Zipf-Mandelbrot-Verteilung an unser Sprichwortmaterial als sehr gut zu bezeichnen ($a = 0.91$, $b = 1.53$, $n = 241$; $\chi^2 = 103.66$, FG = 237, P > 0.99); auffällige Abweichungen finden sich lediglich im Bereich der Ränge 7-17 (vgl. Abb. 3).

Tab. 3: Ranghäufigkeit der Wortformen

i	f_i	NP_i	*i*	f_i	NP_i	*i*	f_i	NP_i	*i*	f_i	NP_i	*i*	f_i	NP_i
1	487	523,50	**11**	100	121,80	**21**	72	71,34	**31**	55	51,04	**41**	45	39,98
2	455	386,47	**12**	94	113,57	**22**	70	68,57	**32**	53	49,65	**42**	43	39,14
3	381	307,89	**13**	92	106,42	**23**	66	66,02	**33**	53	48,34	**43**	42	38,34
4	264	256,71	**14**	92	100,15	**24**	64	63,66	**34**	53	47,10	**44**	42	37,57
5	241	220,62	**15**	86	94,62	**25**	63	61,47	**35**	52	45,92	**45**	40	36,83
6	195	193,75	**16**	85	89,68	**26**	63	59,43	**36**	49	44,80	**46**	38	36,12
7	125	172,93	**17**	84	85,26	**27**	62	57,53	**37**	49	43,74	**47**	38	35,44
8	109	156,31	**18**	81	81,27	**28**	62	55,75	**38**	47	42,73	**48**	37	34,79
9	106	142,72	**19**	80	77,65	**29**	59	54,08	**39**	46	41,77	**49**	35	34,16
10	101	131,39	**20**	73	74,36	**30**	59	52,52	**40**	45	40,85	**50**	34	33,56

Abb. 3 veranschaulicht das Anpassungsergebnis der Zipf-Mandelbrot-Verteilung, ebenfalls aus Gründen der Anschaulichkeit beschränkt auf die ersten 50 Ränge:

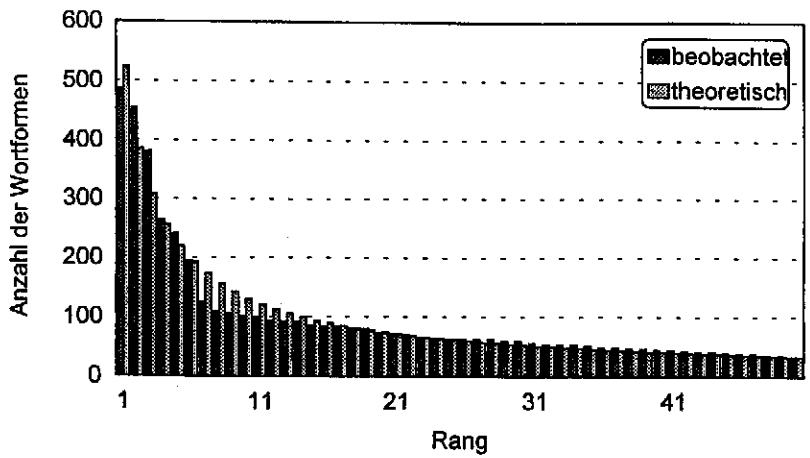

Abb. 3: Ranghäufigkeit der Wortformen (Rang 1-50)

4.2. Frequenzspektrum

Die Frage nach dem Frequenzspektrum entspricht von der Art ihrer Formulierung her den frühen Zipf'schen Überlegungen, in die der Rang der Vorkommenshäufigkeit noch nicht als entscheidende Variable einbezogen war. Mittlerweile konnte allerdings gezeigt werden, dass die Rangverteilung und das Frequenzspektrum ineinander überführbar sind (vgl. Zörnig / Boroda 1992, Chitashvili / Baayen 1993). Es geht konkret um die Frage, wie viele Wortformen (f_i) es gibt, die jeweils genau $i = 1,2,3,\ldots m$ mal vorkommen, so dass sich in der Folge prüfen lässt, ob die Zipf-Mandelbrot'sche Verteilung sich auch in diesem Fall als geeignet erweist. Tab. 4 enthält die entsprechenden Daten aus unserem Sprichwortmaterial.

Tab. 4: Absolute Vorkommenshäufigkeit der Wortformen

i	*fi*	*NP(i)*	*i*	*fi*	*NP(i)*	*i*	*fi*	*NP(i)*	*i*	*fi*	*NP(i)*	*i*	*fi*	*NP(i)*
1	2887	2772,57	**11**	19	18,61	**21**	4	3,90	**31**	1	1,50	**41**	0	0,75
2	732	807,63	**12**	13	15,12	**22**	0	3,48	**32**	1	1,39	**42**	2	0,71
3	314	354,59	**13**	14	12,48	**23**	3	3,13	**33**	2	1,29	**43**	1	0,67
4	171	191,00	**14**	11	10,44	**24**	2	2,82	**34**	3	1,20	**44**	0	0,63
5	119	116,32	**15**	10	8,84	**25**	1	2,55	**35**	1	1,11	**45**	2	0,60
6	85	76,90	**16**	8	7,56	**26**	3	2,31	**36**	0	1,04	**46**	1	0,57
7	54	53,91	**17**	10	6,53	**27**	2	2,11	**37**	1	0,97	**47**	1	0,54
8	41	39,50	**18**	9	5,68	**28**	2	1,93	**38**	2	0,91	**48**	0	0,51
9	29	29,96	**19**	7	4,98	**29**	4	1,77	**39**	0	0,85	**49**	2	0,48
10	19	23,35	**20**	8	4,40	**30**	1	1,63	**40**	1	0,80	**50**	0	0,46

Wie aus Tab. 4 ersichtlich ist, gibt es 2.887 Wortformen, die genau einmal vorkommen, 732 Wortformen, die zweimal vorkommen, usw. Es handelt sich hierbei also um die Wortformen, die insgesamt am seltensten vorkommen; es lässt sich leicht berechnen, dass allein die zehn seltensten Wortformen zusammen genommen nicht weniger als ca. 96%, die zwanzig seltensten Wortformen ca. 98% der lexikalischen Vorkommnisse, usw. abdecken. Abb. 3 enthält einerseits die Daten der Wortformen mit einer Frequenz von 1-50 Vorkommnissen (insgesamt weisen lediglich weitere 35 Wortformen eine höhere Frequenz auf). Abb. 4 veranschaulicht das Ergebnis der Anpassung der Zipf-Mandelbrot-Verteilung an die Daten, welches als äußerst gut anzusehen ist ($a = 2.51$, $b = 0.57$, $n = 50$; $\chi^2 = 43.99$, $FG = 38$, $P = 0.23$).

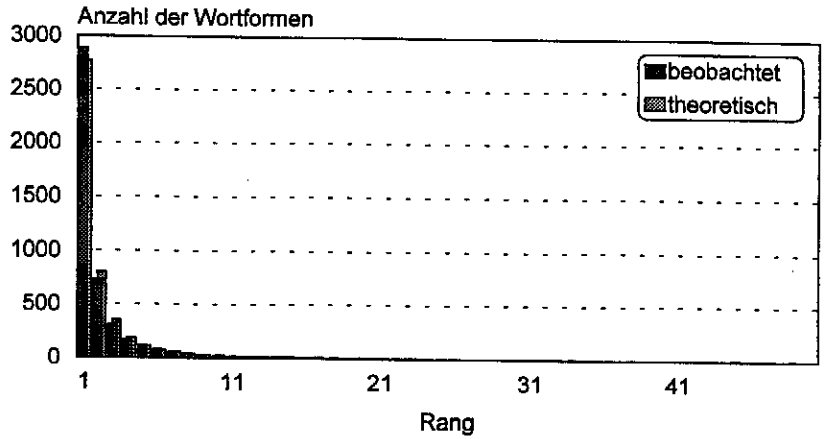

Abb. 4: Anzahl der Wortformen bei gegebener Vorkommenshäufigkeit

4.3. Lexikalische Deckung

Auch die Frage nach der lexikalischen Deckung steht in unmittelbarem Zusammenhang mit den beiden vorherigen, zumal sich auch die lexikalische Deckung mathematisch aus der Rangverteilung ableiten lässt (s.u.). Diese Frage zielt auf den (relativen) Anteil, der von der bzw. den häufigsten Wortformen eingenommen wird: Wie aus Tab. 4 ersichtlich ist, deckt die häufigste Wortform ('ne') 3,15% des gesamten lexikalischen Vorkommens der Sprichwortsammlung ab, die beiden häufigsten Wortformen zusammen 6,09%, die zehn häufigsten Wortformen insgesamt 15,93%. In der Quantitativen Linguistik und Textwissenschaft spricht man in diesem Zusammenhang von der "Textdeckung"; da wir es im gegebenen Fall de facto nicht mit einem homogenen Text zu tun haben, soll statt dessen der Begriff "lexikalische Deckung" verwendet werden.

Tab. 5 enthält die Daten für die ersten 50 Stützpunkte: In der ersten Spalte (*i*) finden sich die Ränge, in der zweiten und dritten die kumulierten absoluten (f_{cum}) und relativen (p_{cum}) Häufigkeiten.

Tab. 5: Kumulierte Vorkommenshäufigkeiten der Wortformen ($N = 15467$)

i	f_{cum}	p_{cum}	i	f_{cum}	p_{cum}	i	f_{cum}	p_{cum}	i	f_{cum}	p_{cum}	i	f_{cum}	p_{cum}
1	487	0,0315	11	2564	0,1658	21	3403	0,2200	31	4026	0,2603	41	4518	0,2921
2	942	0,0609	12	2658	0,1718	22	3473	0,2245	32	4079	0,2637	42	4561	0,2949
3	1323	0,0855	13	2750	0,1778	23	3539	0,2288	33	4132	0,2671	43	4603	0,2976
4	1587	0,1026	14	2842	0,1837	24	3603	0,2329	34	4185	0,2706	44	4645	0,3003
5	1828	0,1182	15	2928	0,1893	25	3666	0,2370	35	4237	0,2739	45	4685	0,3029
6	2023	0,1308	16	3013	0,1948	26	3729	0,2411	36	4286	0,2771	46	4723	0,3054
7	2148	0,1389	17	3097	0,2002	27	3791	0,2451	37	4335	0,2803	47	4761	0,3078
8	2257	0,1459	18	3178	0,2055	28	3853	0,2491	38	4382	0,2833	48	4798	0,3102
9	2363	0,1528	19	3258	0,2106	29	3912	0,2529	39	4428	0,2863	49	4833	0,3125
10	2464	0,1593	20	3331	0,2154	30	3971	0,2567	40	4473	0,2892	50	4867	0,3147

In der lexik-orientierten Forschung ist bei der Veranschaulichung der mit jedem Wort zunehmenden Textdeckung üblicherweise die kumulierte Verteilungsfunktion von einer diskreten in eine stetige Verteilung überführt worden, d.h. dass aus den Stützpunkten der stufigen Verteilung eine Kurve erstellt worden ist. Abb. 5 stellt diese Vorgehensweise in anschaulicher Form für die ersten 30 Datenpunkte des Sprichwortmaterials dar, durch die insgesamt ca. 25% des lexikalischen Vorkommens abgedeckt werden.

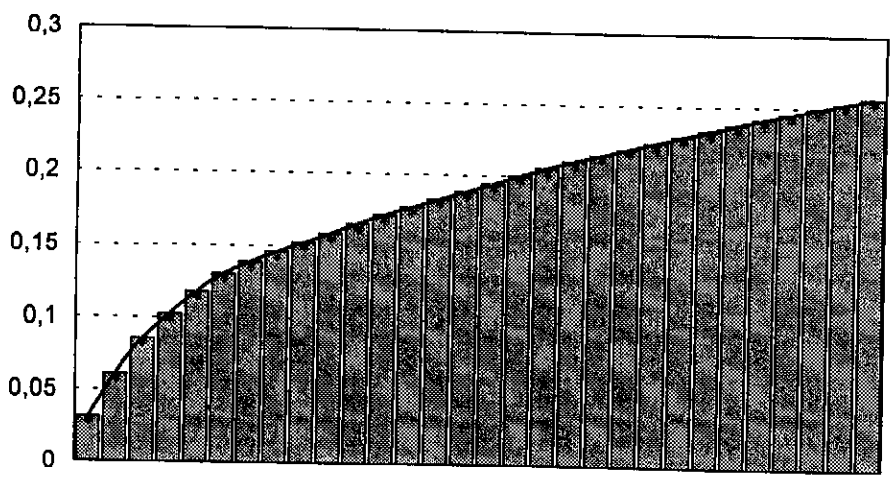

Abb. 5: Kumulierte Vorkommenshäufigkeit der Wortformen

Um den Trend theoretisch zu modellieren, und zwar nicht nur für die ersten 30 Datenpunkte, sondern für das gesamte lexikalische Inventar, soll im folgenden versucht werden, an die Daten ein nicht-lineares Regressionsmodell anzupassen. Abb. 6a und 6b stellen das Ergebnis der Anpassung zweier geeigneter Regressionsmodelle dar:

- Abb. 6a zeigt das ausgezeichnete Ergebnis ($R^2 = .985$) der Anpassung durch ein logarithmisches Modell: $y = a + b \cdot \ln(x)$.
- Abb. 6b zeigt das ebenfalls ausgezeichnete Ergebnis ($R^2 = .975$) der Anpassung durch ein Potenzmodell: $y = ax^v$.

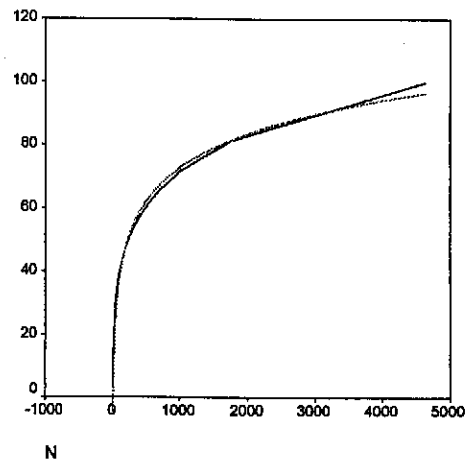

Abb. 6a: Logarithmische Anpassung

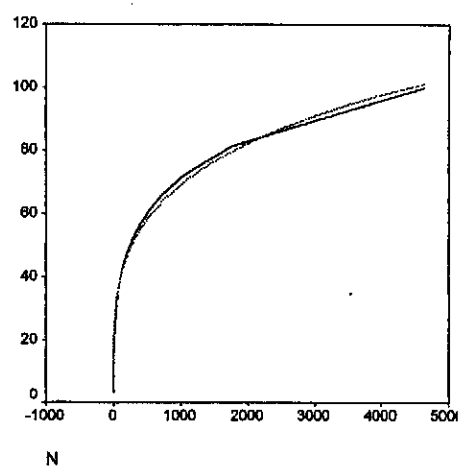

Abb. 6b: Potenzanpassung

Wie zu sehen ist, führen beide Modelle zu nahezu übereinstimmend guten Ergebnissen; dies ist – mathematisch gesehen – nicht weiter verwunderlich: Betrachtet man die gesuchte Funktion als kumulative Häufigkeit – d.h.: $F(x) = P(X < x)$ – dann zeigt die erste Ableitung von $F(x)$ die Ähnlichkeit beider Funktionen:

1. Für die Potenzfunktion

   ergibt sich: $y = a \cdot x^b \Rightarrow F'(x) = y' = ab \cdot x^{(b-1)}$;

   dies führt nach entsprechender Re-Parametrisierung zu:

   $y' = A \cdot x^{(-c)}$

2. Für die Logarithmusfunktion

ergibt sich: $y = a + b \cdot \ln(x) \Rightarrow F'(x) = y' = bx^{(-1)}$

Wie zu sehen ist, unterscheiden sich beide Funktionen nur mehr im Exponenten. Noch interessanter ist jedoch die Tatsache, dass beide Funktionen letztendlich einen Spezialfall der Zipf-Mandelbrot-Verteilung (3) darstellen:

$$(3'') \quad P_x = \frac{C}{(b+x)^a} \quad \Rightarrow \quad y = A \cdot (b+x)^m$$

Unschwer zu sehen ist, dass sich für $b = 0$ eben die Potenzfunktion $y = A \cdot x^m$ ergibt. Es steht außer Frage, dass die Herleitung der (kumulativen) Verteilungsfunktion aus der Zipf-Mandelbrot-Verteilung insofern von besonderer Bedeutung ist, als sich hier der Kreis der theoretischen Argumentation schließt. Diese eher mathematisch motivierte Frage und deren Lösung würde jedoch über den Rahmen der vorliegenden Darstellung hinausgehen und soll einer eigenen Erörterung vorbehalten bleiben (vgl. Antić / Grzybek / Stadlober 2002).

5. Resümee

Aus den vorangegangenen Überlegungen dürfte deutlich geworden sein, dass das lexikalische Inventar einer traditionellen Sprichwortsammlung keineswegs chaotisch organisiert ist, sondern bestimmten Regularitäten folgt. Hierbei handelt es sich ganz offensichtlich um exakt dieselben Gesetzmäßigkeiten, wie sie von der Quantitativen Sprach- und Textwissenschaft schon seit längerem auf der Basis von homogenen Texten bzw. von auf bestimmten Textkorpora basierenden Frequenzlisten erarbeitet wurden. Dass diese Regularitäten auch auf sprichwörtliches Material fruchtbar anzuwenden sind, ist ein bislang nicht beobachteter Befund – den dahinter steckenden mathematischen Zusammenhängen nachzugehen, wird nicht nur für die Parömiologie, sondern auch für die allgemeine Sprach- und Textwissenschaft von weiterreichender Bedeutung sein.[2]

[2] Ich danke G. Altmann (Lüdenscheid) und E. Stadlober (Graz) für hilfreiche Kommentare und Korrekturen.

Literatur

Antić, Gordana; Grzybek, Peter; Stadlober, Ernst (2002): "Lexikalische Vorkommenshäufigkeit und Textdeckung – Theoretische Überlegungen."

Baayen, R. Harald (2001): *Word frequency distributions*. Dordrecht (NL).

Bojc, E. (1974): *Pregovori in reki na Slovenskem*. Ljubljana [^{2}1980, ^{3}1987]

Chitashvili, Revas J.; Baayen, R. Harald (1993): "Word Frequency Distributions" In: Hřebíček, Luděk; Altmann, Gabriel (Hrsg.), *Quantitative Text Analysis*. Trier. (54-135.)

Grzybek, Peter (2000a): "Zum Status der Untersuchung von Satzlängen in der Sprichwortforschung – Methodologische Vor-Bemerkungen", in: *Слово во времени и пространстве. К 60-летию профессора В.М. Мокиенко*, Sankt Petersburg, 430-457.

Grzybek, Peter (2000b): "Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter", in: *Anzeiger für slavische Philologie, 27* (1999) [2000], 87-108.

Grzybek, Peter (2001): "Zur Satz- und Teilsatzlänge zweigliedriger Sprichwörter." In: Uhlířová, Ludmila; Wimmer, Gejza; Altmann, Gabriel; Köhler, Reinhard (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs*. Trier. (64-75)

Grzybek, Peter (2002): "Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter). " In: Palm, Christine (Hg.), *Europhras 2000*. Tübingen. [Im Druck]

Jackson, Willis (ed.) (1953): *Communication Theory*. London.

Kocbek, F. (1887): *Pregovori, prilike in reki.* Celje.

Kocbek, F.; Šašelj, I. (1934): *Slovenski pregovori, reki in prilike.* Ljubljana.

Mandelbrot, Bénoit (1953): "An informational theory of the statistical structure of language". In: Jackson (ed.) (1953); 486-502.

Mandelbrot, Bénoit (1954): "Structure formelle des textes et communication", in: *Word* 10; 1-27.

Prek, St. (1972): *Ljudska modrost. Pregovori, domislice in reki.* Maribor. [Ljubljana, ^{2}1974, ^{3}1982, ^{4}1986, ^{5}1996]

Wimmer, Gejza; Altmann, Gabriel (1996): "The Theory of Word Length Distribution: Some Results and Generalizations." In: Schmidt, Peter (Hg.), *Glottometrika 15*. Trier. (112-133).

Wimmer, Gejza; Altmann, Gabriel (1999): *Thesaurus of univariate discrete probability distributions*. Essen.

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel (1994): "Towards a Theory of Word Length Distribution." In: *Journal of Quantitative Linguistics, 1*; 98-106.

Zipf, George Kingsley (1935): *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, 1968.

Zipf, George Kingsley (1949): *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. New York/London, 1965.

Zörnig, Peter; Boroda, Moisei (1992): "The Zipf-Mandelbrot Law and the Interdependencies of the Frequency Structure and Frequency Distribution in Coherent Texts." In: Rieger, Burghard (Hrsg.), *Glottometrika 13*. Bochum (205-218).

Phraseologie und Parömiologie

Band 14

Schriftleitung / Anschrift der Redaktion

Christoph Chlosta
Universität GH Essen
FB 3 Literatur- und Sprachwissenschaften
D-45117 Essen