

Bibliography

- Collins Cobuild Dictionary of Idioms*. (1995); eds.: Sinclair, J./Fox, G./Moon, R. London.
- Čermák, F. (1998): Usage of Proverbs: What the Czech National Corpus Shows. In: Ďurčo, P.: *Europhras '97*. Bratislava; 37-59.
- Čermák F. (2001): Substance of Idioms: Perennial problems, Lack of data or Theory? In: *International Journal of Lexicography* 14; 1-20.
- Čermák F. (in print): *Paremiological Minimum of Czech: The Corpus Evidence*.
- Čermák F. (in print): *Propositional Idioms*.
- Ďurčo P. (2002): Parömiologische Konnektoren oder „Wie der Volksmund so schön sagt“. In: Piirainen, E./Piirainen, I.P. (Hrsg): *Phraseologie in Raum und Zeit*. Baltmannsweiler; 203-212.
- Ferguson, R. (1983): *The Penguin Dictionary of Proverbs*. Harmondsworth.
- Moon, R. (1998): *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford.
- Searle, J. R. (1979): *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge.
- Whiting, B. J. (1989): *Modern Proverbs and Proverbial Sayings*. Cambridge Mass.

Worthäufigkeit und Wortlänge in Sprichwörtern (Am Beispiel slowenischer Sprichwörter)

Peter Grzybek (Graz)

Einleitung: Regularitäten der Häufigkeit in Sprichwörtern

Bei der Analyse der sprachlichen Organisation von Sprichwörtern ist in den vergangenen Jahren wiederholt kritisch angemerkt worden, dass die entsprechenden Forschungen im Grunde genommen kaum jemals über eine symptomatische Betrachtungsebene hinausgekommen sind und eine systematische Ebene eigentlich noch nicht erreicht haben. Erste Versuche, die in diesem Bereich evidenten Forschungslücken zu füllen, haben sich zunächst auf die Frage der Satz- und Wortlänge in Sprichwörtern gerichtet.

So hat Grzybek (2000a) – ausgehend von einer kritischen Bestandsaufnahme der bisherigen Forschung in dieser Richtung – am Beispiel slowenischer Sprichwörter die Frage gestellt, 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. Mittlerweile lassen sich zumindest einige Antworten auf eine Reihe der aufgeworfenen Fragen geben, die gleichzeitig für weiterführende Untersuchungen an Sprichwörtern anderer Sprachen als plausible Hypothesen dienen können.

So konnte Grzybek (2000b) im Hinblick auf **Satzlängenhäufigkeit** an slowenischem Sprichwortmaterial nachweisen, dass die Häufigkeit, mit der Sprichwörter einer bestimmten Länge in einem solchen Sprichwortkorpus enthalten sind, nicht chaotisch ist, sondern einer bestimmten Regel- oder Gesetzmäßigkeit folgt. Im Detail stellte sich heraus, dass im Hinblick auf slowenische Sprichwörter „die Hyper-Poisson-Verteilung ein geeignetes Modell für die Verteilung von Satzlängen in Sprichwörtern darstellt, wenn

die Satzlänge in der Anzahl der den Satz konstituierenden Wörter gemessen wird (ibd., 105).“

In einer weiteren Studie stand das **Verhältnis von Satz- und Teilsatzlänge** in zweigliedrigen formelhaften Sprichwörtern im Vordergrund (Grzybek 2001). Dabei ließ sich an deutschsprachigen Sprichwörtern ein Trend nachweisen, dem zufolge das erste der beiden Kola sich im Vergleich zum zweiten Kolon als länger erweist. Dieser Trend, der sich als eine Tendenz zur Anti-Klimax interpretieren lässt, ist jedoch nicht von allgemeiner Gültigkeit: Er zeigt sich nämlich lediglich bei Sprichwörtern, deren formelhafte Einleitung nur im ersten Kolon (*Wer ...*) nicht durch eine formelhafte Einleitung im zweiten Kolon aufgegriffen wird. Bei Sprichwörtern mit formelhafter Einleitung im ersten und zweiten Kolon hingegen (*Wer ..., der ...*) kommt eine andere Tendenz zum Ausdruck: Hier tendieren erstes und zweites Kolon dazu, gleich lang zu sein, so dass Formelhaftigkeit in der Tat in einer syntaktisch-lexikalischen Parallelstruktur resultiert – allerdings ist diese Tendenz nur bei den Sprichwörtern mit besonders produktiven „prototypischen“ Strukturen statistisch signifikant.

Auch im Hinblick auf die **Wortlängenhäufigkeit** in Sprichwörtern konnten mittlerweile entsprechende Gesetzmäßigkeiten aufgezeigt werden. So demonstriert Grzybek (2003a) an slowenischen und estnischen Sprichwörtern, dass die Häufigkeit, mit der Wörter einer bestimmten Länge in einem Sprichwortkorpus enthalten sind, nicht chaotisch ist, sondern ebenfalls einer bestimmten Regel- bzw. Gesetzmäßigkeit folgt. Dabei stellte sich in einer Re-Analyse estnischer Sprichwörter (Krikmann 1967) heraus, dass die seinerzeit international so genannte „Fucks-Verteilung“ (eine einfache, 1-verschobene Poisson-Verteilung) nicht zur Modellierung der Wortlängenhäufigkeit in Sprichwörtern geeignet ist. Wohl aber ließen sich sowohl für estnische als auch slowenische Sprichwörter Modelle finden, die sich allerdings sowohl im Hinblick auf die Sprichwörter beider Sprachen als auch im Hinblick auf Ergebnisse zu Wortlängen in Fließtexten der jeweiligen Sprachen voneinander unterscheiden.

Erste Ergebnisse liegen mittlerweile auch in Hinsicht auf die **lexikalische Häufigkeit** in Sprichwörtern vor (Grzybek 2003b). Konkret geht es hierbei um das Problem, 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 Längenhäufigkeiten der zur Disposition stehenden Einheiten (d.h. Wörter, Satz- und Teilsatz und deren Häufigkeiten) ging, stand im Vordergrund der Überlegungen zur Wortvorkommenshäufigkeit

die Frage nach der lexikalischen Häufigkeitsstruktur an und für sich. Diese Frage stellte sich konkret in Form von drei Detailfragen dar:

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

Als Ergebnis auch dieser Untersuchungen stellte sich heraus, dass auch 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.

In der Zusammenschau der bislang vorliegenden Ergebnisse lässt sich also sagen, dass die sprachliche Struktur von Sprichwörtern gesetzmäßig organisiert ist. Diese Frage nach Gesetzmäßigkeiten der sprachlichen Organisation von Sprichwörtern soll im vorliegenden Beitrag weiter verfolgt und vertieft werden. Dabei soll – ausgehend von den oben genannten Arbeiten – erstmals die Frage von Zusammenhängen zwischen den beobachteten Regularitäten gestellt werden. Auch in dieser Hinsicht gibt es bislang mehr offene Fragen als Antworten; insofern soll es ausschließlich um einen spezifischen Aspekt solcher Zusammenhänge gehen, nämlich die Frage des Zusammenhangs von Worthäufigkeit und Wortlänge.

Worthäufigkeit und Wortlänge – Methodologische Grundlagen

Im Hinblick auf den Zusammenhang von Worthäufigkeit und Wortlänge gilt die ursprünglich von G.K. Zipf (insbesondere Zipf 1932, 1935) aufgestellte Hypothese „that the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences“ (1935: 25) als allgemein akzeptiert. Diese Idee von Zipf ist in den vergangenen Jahrzehnten wiederholt aufgegriffen und im Hinblick auf spezifische Fragen untersucht worden. In einer Zusammenschau dieser an Zipf anknüpfenden Arbeiten erweisen sich allerdings einige Probleme insofern als ungelöst, als die Grundlagen der unter-

schiedlichen Herangehensweisen nicht systematisch geklärt wurden. Da die Inkonsistenzen des Herangehens unlängst an anderer Stelle ausführlich diskutiert wurden (vgl. Strauss/Grzybek/Altmann 2003), kann eine Erörterung dieser Umstände an dieser Stelle ausgeblendet werden. Es sei deshalb hier lediglich resümierend auf eine Reihe von Unterschieden in den Grundannahmen bzw. in den konkreten Herangehensweisen hingewiesen.

So herrscht zum Beispiel keine Einigkeit darüber, ob die Frequenz von der Länge oder aber die Länge von der Frequenz abhängt. Auch die zugrunde liegenden Maßeinheiten wurden nicht einheitlich definiert: Während in einigen Untersuchungen die Wortlänge in Silben gemessen wurde, haben andere Untersuchungen Buchstaben als Längeneinheiten benutzt. Weiterhin berücksichtigten einige Studien bei der Betrachtung der Frequenz die absolute Vorkommenshäufigkeit, wohingegen andere den Häufigkeitsrang der Wortformen in Betracht zogen. Darüber hinaus ist festzustellen, dass Zipf und die ihm entlang dieser Linie Folgenden sich im Grunde genommen nicht auf einzelne Texte, sondern entweder auf Korpusmaterial oder aber auf Wörterbücher bzw. Frequenzlisten stützten. Gerade der letzte Umstand wirft die prinzipielle Frage nach der den Untersuchungen zugrunde liegenden Datenqualität, insbesondere das Problem der Datenhomogenität, auf (Altmann 1992).

In der erwähnten Untersuchung von Strauss/Grzybek/Altmann (2003) konnte nun gezeigt werden, dass – ausgehend von der Annahme, dass die Wortlänge eine Funktion der Wortfrequenz ist – ein homogener Text, wie lang auch immer er ist, die Nachweisbarkeit sprachlicher Gesetze nicht beeinträchtigt, während ein inhomogener Text Verzerrungen aufweisen bzw. bedingen kann: Insofern also insbesondere Textmischungen Phänomene hervorrufen können, die einzelnen Texten fremd sind, stellt sich die Frage nach der internen Homogenität von Textkorpora. Diese Beobachtung ist im hier gegebenen Zusammenhang insofern von unmittelbarer Relevanz, als sie direkt die Frage der Datenhomogenität von Sprichwortkorpora und Sprichwörteransammlungen aufwirft. Denn im Grunde genommen lässt sich eine Sprichwortsammlung ja als eine additive Sammlung einzelner (sprichwörtlicher) Texte auf der Satzebene verstehen, zwischen denen es de facto keine textuelle Kohärenz gibt.

Wenn im Folgenden also die Frage des Zusammenhangs von Worthäufigkeit und Wortlänge in Sprichwörtern behandelt wird, geht es auf der einen Seite um einen weiteren Aspekt der übergreifenden Frage der gesetzmäßigen Organisation von Sprichwörtern, auf der anderen Seite um Indizien im Hinblick auf die Frage der Datenhomogenität von Sprichwortsammlungen und -korpora.

Der Untersuchung der genannten Frage liegt dabei der folgende mathematische Ansatz zugrunde: Allgemein ist von einer Differentialgleichung auszugehen, in der die relative Veränderung der mittleren Länge (y) proportional zu der relativen Veränderung der Frequenz abnimmt (Köhler 1986). Da im allgemeinen 0-silbige Wörter nicht vorkommen bzw. sich als Klitika betrachten lassen, kann die mittlere Länge nicht kleiner als 1 werden. Aus diesem Grund muss die Funktion die Asymptote 1 haben. Die Gleichung bekommt dann die Form

$$(1) \quad \frac{dy}{y-1} = -b \frac{dx}{x}.$$

Hieraus ergibt sich die allgemein bekannte Formel

$$(2) \quad y = ax^{-b} + 1$$

mit $a = e^C$ (C ist die Integrationskonstante). Dabei ist y die mittlere Länge derjenigen Wörter, die jeweils x -mal im Text vorkommen. Zählte man auch Wörter der Länge 0, so würde die Konstante 1 entfallen und einige Werte würden (je nach Anteil der 0-silbigen Wörter in den jeweiligen Klassen) tiefer liegen.

Die Hypothese in Formel (2) hat im Vergleich zu anderen Ansätzen den Vorteil, dass die inverse Beziehung die gleiche Formel, jedoch mit anderen Parametern ergibt, d.h.

$$(3) \quad x = A(y-1)^{-B},$$

wobei $A = a^{1/b}$, $B = 1/b$. Daraus folgt, dass auch die Abhängigkeit der Frequenz von der Länge sich auf ähnliche Weise erfassen lassen sollte, jedoch mit transformierten Parametern.

Zur Frage des Zusammenhangs

Von diesem Grundansatz ausgehend, soll im Folgenden unter der Annahme, dass Wortlänge eine von der Wortfrequenz in gegebenen Texten abhängige Variable ist, das genannte Abhängigkeitsverhältnis an einem Sprichwortkorpus getestet werden. Als Untersuchungsmaterial soll dabei die auch schon in den oben genannten Analysen verwendete Sammlung slowenischer Sprichwörter von Kocbek (1887) verwendet werden, die insgesamt 2.429 sprichwörtliche Sätze aufweist.

Um bei der Untersuchung des Abhängigkeitsverhältnisses von Wortfrequenz und Wortlänge letztere in möglichst geringem Umfang weiteren Transformationen zu unterziehen, soll zum einen Wortlänge in der Anzahl

der Silben pro Wort berechnet werden; zum anderen soll jede Wortform separat (als Type) berücksichtigt werden, d.h. die lexikalischen Einheiten sollen nicht lemmatisiert werden. Die insgesamt 2.429 sprichwörtlichen Sätze beinhalten einen lexikalischen Umfang von 15.063 individuell vorkommenden Wortformen (Tokens), die auf 4.478 verschiedenen (Types) basieren.¹

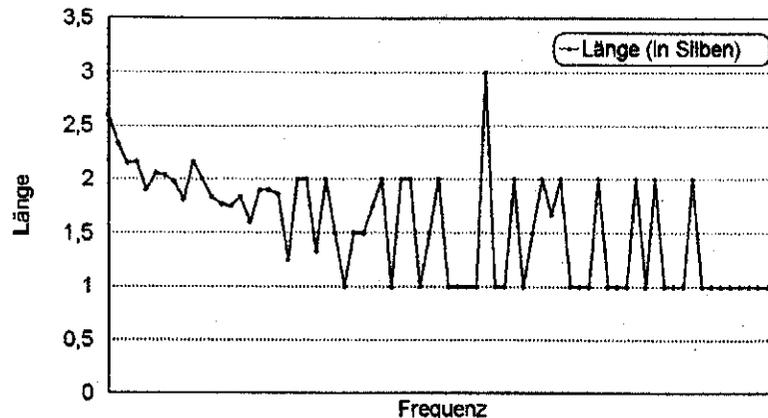
Tab. 1 enthält die Ergebnisse der entsprechenden Berechnungen: In der jeweils ersten Spalte (x) ist die Vorkommenshäufigkeit der jeweiligen Wortformen angegeben, in der jeweils zweiten Spalte (f_x) die Anzahl der Wörter, welche die entsprechende Vorkommenshäufigkeit aufweisen; so gibt es z.B. 2.755 Wortformen, die genau einmal vorkommen, 713 Wortformen, die genau zweimal vorkommen, usw. In der jeweils dritten und vierten Spalte ist die mittlere Länge der jeweiligen Wortformen in Silben und in Graphemen angegeben: So sind z.B. die Wörter, die genau einmal vorkommen, im Durchschnitt $\bar{x} = 2,5975$ Silben lang, die Wortformen, die genau zweimal vorkommen, $\bar{x} = 2,3310$ Silben, usw.

Tab. 1: Abhängigkeit von Wortfrequenz und Wortlänge

x	f_x	Wortlänge in Silben	x	f_x	Wortlänge in Silben	x	f_x	Wortlänge in Silben
1	2755	2,5975	25	2	1,5000	55	1	2,0000
2	713	2,3310	26	4	1,0000	59	2	1,0000
3	310	2,1516	27	2	1,5000	62	2	1,0000
4	168	2,1607	28	2	1,5000	63	2	1,0000
5	114	1,9035	29	4	1,7500	64	1	2,0000
6	84	2,0595	30	1	2,0000	66	1	1,0000
7	51	2,0392	31	1	1,0000	72	1	1,0000
8	42	1,9762	32	1	2,0000	81	1	1,0000
9	32	1,8125	33	1	2,0000	84	1	2,0000
10	18	2,1667	34	3	1,0000	85	1	1,0000
11	21	2,0000	35	2	1,5000	86	1	2,0000
12	12	1,8333	36	1	2,0000	92	1	1,0000
13	17	1,7647	37	1	1,0000	94	1	1,0000
14	12	1,7500	38	1	1,0000	99	1	1,0000
15	6	1,8333	40	2	1,0000	104	1	2,0000
16	10	1,6000	42	2	1,0000	106	1	1,0000
17	10	1,9000	43	1	3,0000	110	1	1,0000
18	10	1,9000	45	1	1,0000	125	1	1,0000
19	7	1,8571	46	1	1,0000	192	1	1,0000
20	8	1,2500	47	1	2,0000	196	1	1,0000
21	4	2,0000	48	1	1,0000	264	1	1,0000
22	1	2,0000	49	2	1,5000	381	1	1,0000
23	3	1,3333	52	1	2,0000	461	1	1,0000
24	2	2,0000	53	3	1,6667	489	1	1,0000

¹ Die hier angegebenen Werte sowie die folgenden Berechnungen unterscheiden sich aufgrund unterschiedlicher Wortdefinitionen zum Teil geringfügig von den in Grzybek (2003b) angeführten.

Abb. 1 stellt die Daten der Tab. 1 in anschaulicher Form dar: Auf der x-Achse ist die jeweilige Vorkommenshäufigkeit, auf der y-Achse die dazugehörigen (in den mittleren Anzahl an Silben pro Wort berechneten) durchschnittlichen Wortlängen abgetragen; zum Zwecke der besseren Anschauung sind die Datenpunkte mit einer Linie verbunden.



Unschwer zu sehen ist, dass durchaus eine allgemeine Tendenz zu bestehen scheint, der zufolge die besonders selten vorkommenden Wortformen länger sind als die häufiger vorkommenden Wortformen. Diese Tendenz kommt allerdings bei den häufiger vorkommenden Wortformen deutlich schlechter zum Ausdruck, da sich hier eine sehr viel höhere Streuung in den Daten bemerkbar macht. Die Ursache für diesen Umstand dürfte höchstwahrscheinlich darin begründet sein, dass die berechneten Mittelwerte insbesondere im Bereich der häufig vorkommenden Wortformen auf wenigen Vorkommnissen beruhen – da es ja nur sehr wenige Wortformen gibt, die häufig vorkommen (vgl. Tab. 1).

Um also in den einzelnen Frequenzklassen zu aussagekräftigeren („repräsentativeren“) Ergebnissen zu kommen, durch die eine allfällig bestehende Tendenz zuverlässiger zum Ausdruck gebracht werden kann, ist eine Glättung der Daten notwendig. Dazu bietet sich ein einfaches Verfahren an, das auch Strauss/Grzybek/Altmann (2003) angewendet haben: Längenklassen x werden zusammengefasst („gepoolt“), wenn $f(x)$ kleiner als 10 ist (d.h. im konkreten Fall, wenn eine Längenangabe auf weniger als 10 Vorkommnissen beruht). Bei dieser Art der Datenzusammenfassung gibt es mehrere Optionen:

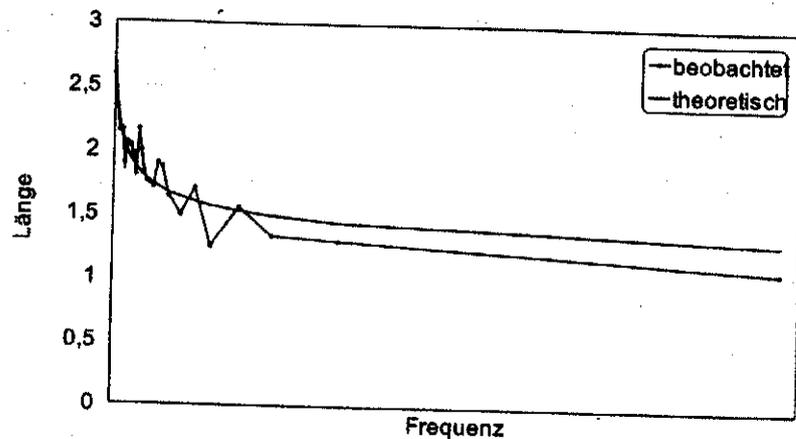
1. das Poolen lässt sich „von oben nach unten“ oder auch „von unten nach oben“ durchführen; in der Tab. 2 (s.u.) ist die Zusammenfassung von unten nach oben durchgeführt worden, d.h., es wird ausgehend von dem einen, am häufigsten vorkommenden Wort solange nach oben aufaddiert, bis mindestens 10 Vorkommnisse zusammengefasst sind (im gegebenen Fall also bis zu dem einen Wort, das 104 Mal vorkommt);
2. das Poolen lässt sich mit und ohne Gewichtung durchführen: ohne Gewichtung wird bei der Berechnung der mittleren Länge das arithmetische Mittel der in die neu gebildete Klasse aufgenommenen Vorkommnisse gebildet, mit Gewichtung wird die Anzahl der Vorkommnisse gewichtend berücksichtigt. In der Tab. 2 ist die durchschnittliche Länge der zusammengefassten Vorkommnisse ohne Gewichtung berechnet.

Tab. 2 stellt die auf die beschriebene Art und Weise erhaltenen Werte (x') neben den entsprechend berechneten mittleren Wortlängen dar. Ebenfalls enthalten sind die theoretischen Werte, die sich aus Formel (2) ergeben, wenn man $a = 1,7009$ und $b = -0,3038$ einsetzt; diese Werte erhält man als optimale Anpassungswerte in entsprechenden Iterationsprozeduren.

Tab. 2: Abhängigkeit von Wortfrequenz und Wortlänge (beobachtet und theoretisch)

x'	Wortlänge (in Silben)		x'	Wortlänge (in Silben)	
	beobachtet	theoretisch		beobachtet	theoretisch
1	2,5975	2,7009	14	1,7500	1,7630085
2	2,3310	2,3780	15,50	1,7167	1,73977892
3	2,1516	2,2183	17	1,9000	1,71930956
4	2,1607	2,1163	18,50	1,8786	1,70106908
5	1,9035	2,0432	21,50	1,6458	1,66978468
6	2,0595	1,9870	25,50	1,5000	1,63595424
7	2,0392	1,9418	30,50	1,7083	1,60228963
8	1,9762	1,9044	36,67	1,2500	1,56952526
9	1,8125	1,8726	46,50	1,5625	1,52987117
10	2,1667	1,8451	58,40	1,3333	1,49443586
11	2,0000	1,8210	82,30	1,3000	1,4455057
12	1,8333	1,7996	242,80	1,1000	1,32072344
13	1,7647	1,7804			

Abb. 2 veranschaulicht die Tendenz, die nunmehr hinreichend deutlich zum Ausdruck kommt.



Die Güte der Anpassung unserer theoretischen Kurve an die beobachteten Werte lässt sich statistisch testen, und zwar in Form des Determinationskoeffizienten (vgl. Grotjahn 1992), der in unserem Fall bei $R^2 = 0.8340$ liegt. Der Wert des Determinationskoeffizienten ist zwar nicht direkt interpretierbar, doch hat es sich im Bereich der quantitativen Linguistik und Textanalyse eingebürgert, Werte von ≥ 0.85 als Ergebnis einer guten bzw. sehr guten Anpassung anzusehen, womit unser R^2 -Wert ziemlich genau an dieser Grenze läge. Insofern liegt es nahe, ihn mit den R^2 -Werten aus den Textanalysen von Strauss/Grzybek/Altmann (2003) zu vergleichen, die ja genau auf die Frage der Datenhomogenität ausgerichtet waren. Dabei stellte sich heraus, dass in einzelnen Texten verschiedenster Sprachen die R^2 -Werte zwischen $R^2 = 0.84$ und $R^2 = 0.97$ lagen, solange die zugrunde liegenden Texte homogen waren; die Werte wurden deutlich schlechter, sobald man es mit irgendeiner Art von Textmischung zu tun hatte, und dies war nicht nur bei verschiedenen Texten, sondern auch schon bei der Kombination von mehr als einem Romankapitel der Fall: So lag z.B. der R^2 -Wert für das erste Kapitel von Tolstoj's *Anna Karenina* bei $R^2 = 0.88$, für das gesamte (aus 34 Kapiteln bestehende) erste Buch allerdings nur bei $R^2 = 0.86$; und während das erste Kapitel des Romans *The Portrait of a Lady* von Henry James auf einen Wert von $R^2 = 0.89$ kam, lag der Wert für den gesamten (aus 55 Kapiteln bestehenden) Roman bei $R^2 = 0.58$; lediglich radikalere Glättungen, etwa in Form von Klassen-

zusammenfassungen, die auf mindestens 20 Vorkommnissen beruhen, führten hier zu besseren Ergebnissen ($R^2 = 0.92$).

Insofern lässt sich das an unserem Sprichwortmaterial erhaltene Ergebnis von $R^2 = 0.834$ als durchaus akzeptables Ergebnis und als Beleg für die oben aufgestellten Annahmen ansehen: Auf der einen Seite bestätigt sich auch an Sprichwortmaterial die für homogene Texte nachgewiesene Regularität des Zusammenhangs von Wortfrequenz und Wortlänge, auf der anderen Seite liegt damit ein weiterer Beleg dafür vor, dass die sprachliche Organisation von Sprichwörtern nicht chaotisch, sondern gesetzmäßig ist. Keinesfalls gibt es Grund für die Annahme, dass man bei einer Sammlung bzw. einem Korpus Sprichwörtern a priori davon ausgehen darf, dass es sich um heterogenes Datenmaterial handelt.

Literatur

- Altmann, G. (1992): „Das Problem der Datenhomogenität“. In: *Glottometrika 13*. Bochum; 287-298.
- Grotjahn, R. (1992): „Evaluating the adequacy of regression models: Some potential pitfalls“. In: *Glottometrika 13*. Bochum; 121-172.
- Grzybek, P. (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, P. (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, P. (2001): „Zur Satz- und Teilsatzlänge zweigliedriger Sprichwörter.“ In: Uhlřřov, L./Wimmer, G./Altmann, G./Köhler, R. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs*. Trier; 64-75.
- Grzybek, P. (2003b): „Zur lexikalischen Struktur von Sprichwörtern“. In: Burger, H./Häcki-Buhofer, A./Gręciano, G. (eds.), *Flut von Texten – Vielfalt der Kulturen*. Baltmannsweiler; 97-116.

- Grzybek, P. (im Druck): „Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter).“ In: Palm, Ch. (Hg.), *Europhras 2000*. Tübingen.
- Kocbek, F. (1887): *Pregovori, prilike in reki*. Celje.
- Köhler, R. (1986): *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum.
- Krikmann, A. (1967): „Keelestatikat Eesti vanasõnadest“ [= Sprachstatistik estnischer Sprichwörter], in: *Emakeele Seltsi Aastaraamat*, 13; 127-154.
- Strauss, U./Grzybek, P./Altmann, G. (2003): „The more the better? Word Length and Word Frequency“. [Submitted]

La conceptualisation phraséologique: une construction conceptuelle langagière ou la prégnance du concept par la langue

Salah Mejri (Manouba)

Dans les études linguistiques, il est vraiment rare qu'on se hasarde sur le terrain de la conceptualisation soit parce qu'on considère tout simplement que cela ne fait pas partie des considérations linguistiques, soit parce qu'un tel phénomène ne cadre pas avec la répartition des différentes disciplines comme la syntaxe, la sémantique, la lexicologie, etc. Il faut reconnaître par ailleurs qu'elle représente l'une des questions les plus redoutables dans ce sens que son étude conduit à un réexamen critique ou à une remise en question d'un certain nombre de certains concepts de la linguistique, dont les définitions sont considérées comme relativement stabilisées. Or on assiste ces derniers temps, particulièrement avec les travaux sur la dénomination, un regain d'intérêt pour la question, même s'il n'est pas tout à fait direct. Pourquoi un tel regain d'intérêt? Il semble qu'il soit dû essentiellement au nombre de plus en plus important des études portant sur le figement en général et les phraséologismes en particulier, qui sont l'occasion de questionnements fondamentaux en rapport avec les mécanismes structurels des langues naturelles, dont nous retenons ici la relation que le signe linguistique peut avoir avec le domaine des objets, celui des concepts et celui de la signification. Pour ce faire, il faut faire une distinction très nette entre des entités comme le référent, le concept et le signifié, en en précisant les points communs et les points de différences. Ce n'est qu'après qu'il serait possible de montrer comment la conceptualisation peut se construire dans et par la langue, dans quelles conditions et selon quels mécanismes. Le résultat final serait de vérifier le tout dans des systèmes linguistiques différents pour illustrer concrètement l'impact de cette

Phraseologie und Parömiologie

Herausgegeben von
Wolfgang Eismann (Graz)
Peter Grzybek (Graz)
Wolfgang Mieder (Burlington VT, USA)

In Zusammenarbeit mit der
Europäischen Gesellschaft für Phraseologie
vertreten durch:

Harald Burger (Zürich), Wolfgang Eismann (Graz)
Peter Ďurčo (Bratislava), Gertrud Gréciano (Strasbourg)
Jarmo Korhonen (Helsinki), Christine Palm (Uppsala), Jan Wirrer (Bielefeld)

Band 15

Schriftleitung / Anschrift der Redaktion

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

Phraseologismen als Gegenstand sprach- und kulturwissenschaftlicher Forschung

Akten der Europäischen Gesellschaft für
Phraseologie (EUROPHRAS) und des
Westfälischen Arbeitskreises »Phraseologie/Parömiologie«
(Loccum 2002)

Herausgegeben von
Csaba Földes und Jan Wirrer



Schneider Verlag Hohengehren GmbH