

Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter

Peter Grzybek, Graz

0. Einleitende Vorbemerkung

Im vorliegenden Text geht es um die Häufigkeitsverteilung von Satz­längen in slowenischen Sprichwörtern. Die mit diesem Interesse verbundene Frage läßt sich – etwas naiver und etwas umständlicher zugleich – auch folgendermaßen formulieren: Wie oft kommen in einem Korpus von (im konkreten Fall slowenischen) Sprichwörtern Sätze mit einer bestimmten Länge vor?

Der Untersuchung dieser Fragestellung liegen eine Reihe von aufeinander aufbauenden Grundannahmen zugrunde, nämlich:

1. daß die Häufigkeit, mit der Sprichwörter einer bestimmten Länge in einem solchen Sprichwortkorpus enthalten sind, nicht zufällig (chaotisch) ist, sondern einer bestimmten Regel- oder Gesetzmäßigkeit folgt;
2. daß sich diese Regelmäßigkeit nicht nur beschreiben, sondern auch formalisieren läßt;
3. 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;
4. daß diese Querverbindungen Aussagen über die Spezifik von Sprichwörtern bzw. von sprichwörtlichen Sätzen erlauben.

Bei der konkreten Behandlung der aufgezeigten Problemstellungen werden wir, um den Gang des vorliegenden Textes kurz im Voraus zu skizzieren, folgendermaßen vorgehen: Im Anschluß an eine knappe Darstellung zur Geschichte der slowenischen Parömiographie (1.1.) und einer ebenso kurzen Beschreibung des unserer Untersuchung zugrundegelegten Sprichwortkorpus (1.2.) folgt eine Darstellung zum Status der Satz­längenforschung (2) – aus der sich dann auch die verschiedenen angewendeten methodologischen Ver-

fahren ergeben –, an die sich die eigentlichen Analysen anschließen (3), deren Ergebnisse in einer verallgemeinernden Zusammenfassung synthetisiert werden (4).

1. Zur slowenischen Parömiographie

1.1. Zur Geschichte der slowenischen Parömiographie

Zu Recht beklagt Toporišič (1984: 291) den insgesamt schlechten Zustand nicht nur der slowenischen Phraseographie im allgemeinen, sondern auch der Parömiographie im besonderen. So gilt für die slowenische Parömiographie im Grunde genommen nichts Anderes, als das, was Mieder (1988: 11, 1989: 187) für die Erstellung von Sprichwörteransammlungen grundsätzlich festgestellt hat, nämlich daß die verschiedenen Kompilatoren – zum Teil seit Jahrhunderten – schlichtweg voneinander abschreiben.

Auf welchen Zeitraum sich die frühesten Dokumentationen slowenischer Sprichwortgüter zurückführen lassen, ist keine eindeutig zu beantwortende Frage. Wollte man zum Beispiel die erste eigenständige slowenische Sprichwörteransammlung von Janez Mihelič (1780) als frühestes parömiographisches Werk ansehen, so wäre dem entgegenzuhalten, daß dieses Werk nicht erhalten ist. Oder wollte man die *Pareomiologia* von Hieronymus Megiser (1592 bzw. 1605) als die erste (erhaltene) Sammlung slowenischer Sprichwörter betrachten, so wäre zu bedenken, daß gerade über die hier dokumentierten 20 Sprichwörter immer wieder diskutiert worden ist, ob es sich bei dem Material überhaupt um slowenisches oder eher kroatisches Sprichwortgut handelt (vgl. Radics 1882; Levstik 1895, Rode 1989); jüngste Untersuchungen in dieser Frage legen dabei nicht nur deutlich offen, wie leichtfertig die bei der Behandlung dieser Frage notwendige akribische textphilologische Arbeit in der Vergangenheit genommen wurde, sondern argumentieren zudem recht überzeugend, daß die betreffenden Sprichwörter sich in historischer Sicht sowohl im Slowenischen als auch im čakavischen Kroatischen finden lassen, was Eismann (1997) veranlaßt, die Informationsquelle dieser Sprichwörter im istrischen Grenzgebiet zu vermuten. So hat die Unsicherheit über die slowenische Autochthonie der Megiser'schen Sprichwörter mitunter auch dazu geführt, die ca. 150 Sprichwörter aus der *Windischen Spachlehre* (1777) bzw. aus dem *Deutsch-Windischen Wörterbuch* (1789) von Oswald Gutschmann als älteste erhaltene Sammlung slowenischer Sprichwörter anzusehen (vgl. Grafenauer 1935).

Wie dem auch sei: Ungeachtet dieser für die historische Parömiographie sicherlich spannenden offenen Fragen, und abgesehen von zahlreichen klei-

neren Sammlungen, die im 18. und 19. Jahrhundert zusammengestellt worden sind¹ – stellt ohne Frage die Sammlung *Pregovori, prilike in reki* von Kocbek (1887) die eigentlich erste umfassende, eigenständige Sprichwortsammlung des Slowenischen dar und ist insofern nicht nur als ein Meilenstein, sondern als Fundament der slowenischen Parömiographie schlechthin anzusehen. Natürlich beruht auch dieses Werk auf vorbereitenden Arbeiten und kleineren zuvor an unterschiedlichen Stellen veröffentlichten Sammlungen, doch nimmt sie ungeachtet dessen einen besonderen Stellenwert ein, weil sie – 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 Sammlungen wie diejenigen von Bojc (1974 u. a.) oder Prek (1970 u. a.).²

Aus diesem Grunde soll die Sammlung von Kocbek (1887) für die vorliegende Untersuchung als Materialbasis dienen.

1.2. Kurze statistische Angaben zur Sprichwörtersammlung von Kocbek (1887)

Wenn in der Parömiologie überhaupt nach Satzlängen gefragt worden ist, dann nahezu ausschließlich nach der durchschnittlichen Satzlänge von Sprichwörtern eines bestimmten Sprichwortkorpus; Grundlage der Berechnungen war dabei entweder die Anzahl der Wörter oder die Anzahl der Silben pro sprichwörtlichem Satz. Diese durchschnittlichen Satzlängen sind in der Folge häufig zu Vergleichen mit anderen Sprichwortkorpora herangezogen worden, wobei es sowohl um intralinguale Vergleiche (z. B. Sammlungen verschiedener Epochen oder Sprichwörter mit unterschiedlichem Bekanntheitsgrad) als auch um interlinguale Vergleiche (also Sprichwörter in bzw. aus verschiedenen Sprachen) ging. Ein ausführlicher Abriß zum Status der Satzlängenforschung in der Parömiologie findet sich bei Grzybek (2000), so daß an dieser Stelle auf eine detailliertere Darstellung verzichtet werden kann. Es stellt sich dabei heraus, daß den meisten bisherigen Arbeiten zur

¹ Es ist hier nicht der Ort, auf diese Sammlungen – die übrigens auch in der kleinen Bibliographie von Klančar (1948) nicht annähernd erschöpfend erfaßt sind – detailliert einzugehen.

² Die mehrfachen Auflagen der Sammlung von Bojc und Prek sprechen für ihre Popularität; doch gilt es zu beachten, daß in den späteren Auflagen zum Teil in erhöhtem Maße nicht nur volkstümliche Sprichwörter, sondern auch Geflügelte Worte und Aphorismen (durchaus nicht nur slowenischer Herkunft) verstärkt Niederschlag finden.

Satzlänge in Sprichwörtern die notwendige statistische Solidität fehlt, insofern Vergleiche häufig auf rein intuitiver Basis (oder z.B. aufgrund optischer Eindrücke) angestellt worden sind.

Wie gesagt: Wenn in der Parömiologie bislang nach Satzlängen gefragt worden ist, dann nahezu ausschließlich im Hinblick auf solche Durchschnittswerte wie die hier angeführten. Gegen dieses Vorgehen ist methodologisch im Prinzip nichts einzuwenden.

Dennoch wollen wir uns an dieser Stelle nicht mit dem einfachen Berechnen von Mittelwert und Standardabweichung begnügen, und zwar u.a. aus dem Grunde, weil Mittelwerte mitunter insofern einen trügerischen Eindruck vermitteln, als sie nur eine bestimmte Perspektive auf das Datenmaterial werfen. So können identische Mittelwerte z.B. aufgrund von unterschiedlich variierenden Werten zustande kommen (z.B. $\bar{x} = 3.5$ für die Stichprobe n_1 mit den Werten 2,3,4,5 und $\bar{x} = 3.5$ für die Stichprobe n_2 mit den Werten 2,2,2,8) – dieser Unterschied wird in der Regel mit der sich unterscheidenden Varianz bzw. der Standardabweichung (d.h. der Wurzel aus der Varianz) beschrieben (im angeführten Beispiel $s_1 = 1.29$ vs. $s_2 = 3.00$). Andererseits können auch Stichproben mit identischer Standardabweichung aufgrund von vollkommen unterschiedlichen Voraussetzungen zustande kommen, was ebenfalls kurz an folgendem fingiertem Beispiel veranschaulicht sei: Gegeben sei die Stichprobe n_1 mit den Häufigkeiten 1, 2, 3, 4, 5, und die Stichprobe n_2 mit den Häufigkeiten 5, 4, 3, 2, 1. Der Mittelwert beider Stichproben ist nicht gleich, wohl aber die Standardabweichung. Aus diesem Grunde ist es sinnvoll, in Ergänzung zu Mittelwerten und Standardabweichungen nicht nur z.B. weitere „Maße der zentralen Tendenz“ und der „Streuung“, wie diese in der Statistik genannt werden, zu berechnen, sondern die gesamte Form der Häufigkeitsverteilungen zu untersuchen.

Im Hinblick auf Worthäufigkeiten wäre also entsprechend in einem ersten Schritt zu fragen, wie häufig Wörter einer bestimmten Länge vorkommen; darüber hinausgehend wäre dann in einem zweiten Schritt die Frage anzuschließen, ob sich die resultierende(n) Häufigkeitsverteilung(en) auch theoretisch formalisieren und interpretieren lassen. Diese Schritte werden wir unten im Abschnitt 4 der Reihe nach unternehmen.

In diesem Sinne seien zunächst die Ergebnisse der entsprechenden Berechnungen auch an unserem Korpus kurz referiert. Die Sammlung von Kocbek besteht aus insgesamt 2.429 Sätzen. In unserem Korpus beträgt die in Worten gemessene Satzlänge im Durchschnitt $\bar{x} = 6.37$ Wörter bei einer Standardabweichung von $s = 2.48$; die in Silben berechnete Durchschnittslänge der 2.429 Sätze beträgt $\bar{x} = 11.52$ bei einer Standardabweichung von $s = 4.29$. Diese beiden Ergebnisse zusammengenommen decken sich mit der

Tatsache, daß das Sprichwortkorpus einen Umfang von 15.467 Wörtern bzw. von 27.977 Silben aufweist (was seinerseits einer durchschnittlichen in Silben gemessenen Wortlänge von $\bar{x}=1.81$ Silben bei einer Standardabweichung von $s=0.85$ entspricht).

Mit diesen Angaben lassen sich also Aussagen darüber treffen, wie sich die einzelnen Werte im Durchschnitt zum Mittelwert der gesamten Stichprobe verhalten; mit einer Analyse der Häufigkeitsverteilung hingegen läßt sich überprüfen, ob die Besetzung der einzelnen Klassen zufällig ist oder aber einer bestimmten Regel folgt. Und in eben diesem Sinne wollen wir dafür argumentieren, die Untersuchung von sprichwörtlichen Satzlängen um weitere Verfahren als nur die Berechnung von Durchschnittslängen zu erweitern; Methodologisch heißt das – und dies gilt allgemein und betrifft die Sprichwortforschung nicht anders als andere Wissenschaftsdisziplinen auch – daß die Untersuchung von Durchschnittswerten und von Häufigkeitsverteilungen als komplementär zu verstehen sind und Hand in Hand gehen müssen. Allerdings hat diese Einsicht – und zwar keineswegs nur in der Sprichwortforschung – bislang noch nicht ausreichend um sich gegriffen. Insofern betreten wir mit der spezifischen Fragestellung nach der Häufigkeitsverteilung von Satzlängen mehr oder weniger Neuland nicht im Bereich der slowenischen Sprichwortforschung, sondern innerhalb der gesamten Parömiologie schlechthin.

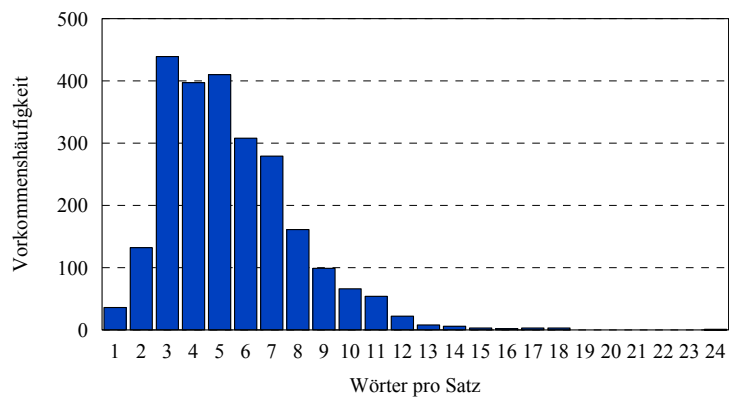
Bevor wir uns jedoch weiter in Fragen der Theorie verstricken, schauen wir uns die Rohdaten an. Tab. 1 gibt wieder, wie viele Sätze jeweils welcher Satzlänge in unserem Korpus vorkommen; angeführt sind sowohl die absoluten Werte (f_i) als auch die in Prozente umgerechneten Anteile ($f_i\%$). Die Ergebnisse sind in Abb. 1 anschaulich dargestellt.

Um nun in einem nächsten Schritt zu überprüfen, ob diese Häufigkeitsverteilung zufällig ist oder einer bestimmten Regelmäßigkeit folgt, müssen wir uns einige Grundlagen der Häufigkeitsanalyse aneignen; dabei wollen wir uns aber im gegebenen Zusammenhang darauf beschränken, den Status der Satzlängenforschung in der gebotenen Kürze zur Kenntnis zu nehmen, damit wir unsere spezifisch parömiologischen Interessen in diesen allgemeinen Forschungskontext einbetten können.

Tab. 1: Vorkommenshäufigkeit der in Worten berechneten Satzlängen in Kocbek (1887)

Wörter pro Satz	f(i)	f(i)%	Wörter pro Satz	f(i)	f(i)%
2	36	1,48	14	8	0,33
3	132	5,43	15	6	0,25
4	439	18,07	16	3	0,12
5	397	16,34	17	2	0,08
6	410	16,88	18	3	0,12
7	308	12,68	19	3	0,12
8	279	11,49	20	–	–
9	161	6,63	21	–	–
10	99	4,08	22	–	–
11	66	2,72	23	–	–
12	54	2,22	24	1	0,04
13	22	0,91			

Abb. 1: Vorkommenshäufigkeit der in Worten berechneten Satzlängen in Kocbek (1887)



2. Zur Geschichte der Satzlängenforschung

Die Frage der Satzlänge ist natürlich keine spezifisch parömiologische Fragestellung, sondern ein den engeren Bereich der Sprichwortforschung überschreitendes Problem. Dennoch könnte man das Sprichwort in gewissem Sinne als Prototyp zur Untersuchung von Satzlängen ansehen, insofern bei allen Versuchen seiner Definition immer wieder darauf hingewiesen wird, daß es (in der Regel) die Ebene des Satzes nicht übersteigt – diese Annahme der Prototypik wird es unten noch einzuschränken gelten, doch wenden wir uns zunächst der historischen Entwicklung zu.

In einem historischen Rahmen gilt es auf den Kanadier Sherman (1888) zu verweisen, der sich Ende des 19. Jahrhunderts als eigentlicher Begründer der Satzlängenforschung erwies: In seinen Untersuchungen zu Texten verschiedener Autoren gelangte er zu diskreten Durchschnittswerten und begründete so ein Verfahren zur Feststellung der Autorenschaft. Diese Diskussion wurde in unserem Jahrhundert unter verschiedenen Perspektiven weitergeführt, von denen hier nur einige exemplarisch genannt seien – zum Beispiel:

- Ist eine bestimmte durchschnittliche Satzlänge charakteristisch für den Stil eines Autors, für ein spezifisches Textgenre, oder für einen bestimmten Funktionalstil?
- Kann Satzlänge als Kriterium für fragliche Verfasserschaft herangezogen werden?
- Ist Satzlänge ein Indikator für den Schwierigkeitsgrad eines Textes?
- Gibt es einen Zusammenhang zwischen der Satzlänge und sprachlichen Einheiten anderer Ebenen (wie z.B. der Wortlänge)?
- Mit welchen psychischen oder kapazitären Prozessen der Sprachproduktion und Sprachrezeption korrelieren Satzlängen?

Die Liste möglicher Fragestellung ließe sich fortsetzen, was an dieser Stelle jedoch unterbleiben kann (vgl. auch Niehaus 1996). Die bekanntesten Versuche, Satzlänge als Stilcharakteristikum zu bestimmen, gehen sicherlich auf die Arbeiten von Yule (1939, 1944) zurück: Auf der Basis der Werke dreier englischer Autoren berechnete er in fortlaufenden Textpassagen die in Worten gemessene Satzlänge und wendete „einfache“ statistische Verfahren wie Mittelwertberechnung, Median und Standardabweichung an, auf deren Basis er den sogenannten Yuleschen Koeffizienten bestimmte. Abgesehen von begrenzten Schwankungen dieses Koeffizienten bei den Texten eines Autors fand er erheblichere Schwankungen bei den Werten verschiedener Autoren und zog so die Schlußfolgerung, daß Satzlänge als autorspezifisches Charakteristikum angesehen werden kann. Diese Annahme wurde aufgrund anderer

Untersuchungen teilweise bestätigt (vgl. Wake 1957, Clayman 1981), teilweise aber auch in Frage gestellt. Auf jeden Fall hat sich in diesen weiterführenden Untersuchungen gezeigt, daß die Wahl der jeweiligen Maßeinheit erheblichen Einfluß auf die Ergebnisse hat, insofern die Berechnung der Satzlänge in Worten zu ganz anderen Resultaten führt bzw. führen kann als eine Berechnung der Satzlänge z.B. in Silben, Phonemen oder aber, wie Altman (1988) vorgeschlagen hat, in Teilsätzen.

Ohne auf diesen gesamten Problembereich im folgenden detailliert einzugehen, kann man zusammenfassend Niehaus (1997: 215) darin zustimmen, „daß die Satzlänge kein eindeutiges Kriterium zur Bestimmung der Autorschaft eines Textes liefert“. Diese Feststellung – so wichtig sie auch ist – muß uns im folgenden jedoch nicht weiter beunruhigen. Denn zwar ist für uns von Belang, in welcher Maßeinheit die Satzlänge von Sprichwörtern sinnvollerweise zu berechnen ist – die Frage der Feststellung von Autorschaft auf der Grundlage von Satzlängen ist für unsere Fragestellung praktisch unerheblich, da es sich gerade bei Sprichwörtern ja um kollektive, stereotype Texte handelt, die mit individueller Autorschaft nichts zu tun haben. Insofern wollen wir uns eher auf die Frage konzentrieren, welche Modelle entwickelt worden sind, die Häufigkeit von Satzlängen in Texten zu formalisieren. Unter Formalisierung ist dabei eine solche mathematische Beschreibung zu verstehen, die es erlaubt, die tatsächliche Vorkommenshäufigkeit der einzelnen Satz­längen mit einem theoretisch zu berechnenden Verteilungsmodell so zu vergleichen, daß anschließend mit Hilfe eines statistischen Tests die Güte der Übereinstimmung der Ergebnisse geprüft werden kann. Solche Verteilungsmodelle (von denen es ca. 750-1000 verschiedene, wenn auch in bestimmten „Familien“ zusammengefaßte gibt) entsprechen bestimmten Formeln mit Parametern; es kommt also darauf an, für jedes Modell zunächst die Parameter so zu bestimmen, daß die sich aufgrund der Formel ergebenden theoretischen Werte möglichst gut mit den beobachteten übereinstimmen – dieser Vorgang wird als „Anpassung“ einer theoretischen Verteilung an beobachtete Daten bezeichnet –, um dann zu sehen, welches der Modelle insgesamt der beobachteten Häufigkeitsverteilung am besten entspricht.

3. Zur Modellierung der Satz­längen­häufigkeit

Vor dem Hintergrund der Tatsache, daß traditionellerweise die Satz­länge auf der Basis der Wortanzahl pro Satz berechnet wurde, sind in der Vergangenheit verschiedene Modelle der Satz­längen­häufigkeitsverteilung diskutiert worden. Eines von ihnen ist die sog. Lognormalverteilung, wie sie z.B. von

Williams (1969) vorgeschlagen wurde; es handelt sich hierbei im Grunde genommen um die allgemeine Normalverteilung, deren Werte einfach logarithmiert werden. Wie Altmann (1988: 149f.) vermutet, dürfte der Grund für die Annahme einer Lognormalverteilung darin zu suchen sein, daß man für einen Text bzw. für einen Autor einen festen Satzlängenmittelwert als sein Charakteristikum annimmt, und zufällige („normale“, jedoch logarithmisch transformierte) Abweichungen von diesem Mittelwert vermutet. Mit dieser Vermutung sieht Altmann (1988: 150) jedoch zwei Probleme verbunden. Erstens ist verschiedentlich gezeigt worden, daß der Satzlängenmittelwert auch innerhalb eines Textes nicht konstant ist, sondern sich im Textverlauf beträchtlich ändern kann. Und zweitens vernachlässigt die in der Vermutung implizierte Annahme der Zufälligkeit der Satzlängenerzeugung die Wirkung der anzunehmenden Zipf'schen Kräfte, die auf ein Gleichgewicht zwischen produktions- und rezeptionsökonomischen Faktoren ausgerichtet sind; sie erlaubt es somit nicht, das Konzept der Satzlängen in eine allgemeinere Theorie der Längen von Spracheinheiten zu integrieren, da außer dem Zufall keine (weiteren) Faktoren in Betracht gezogen werden. Wenn man jedoch annimmt, daß regulatorische Faktoren in den Prozeß der Satzlängenerzeugung eingreifen (oder diesen gar steuern), dann ist zu erwarten, daß die Abweichungen vom Mittelwert nicht zufällig sind, sondern ihrerseits einer bestimmten Regel- oder Gesetzmäßigkeit folgen, und daß erst die Abweichungen von diesem Gesetz zufällig sind (wenn nicht sogar auch sie ihrerseits wiederum einem bestimmten Gesetz folgen sollten).

Kritik an der Annahme der Lognormalverteilung hat – aus anderen, hier nicht zu diskutierenden Gründen – u.a. auch Sichel (1974) geübt. Er selbst hat seinerseits vorgeschlagen, Satzlänge als eine poissonverteilte (d.h. dem Gesetz der kleinen Zahlen folgende) Variable zu betrachten. Demnach würde die Satzlängenverteilung der Formel (1) folgen:

$$(1) f(x|r) = \frac{e^{-r} r^x}{x!}, \quad x = 0, 1, \dots$$

Dabei wird r nicht als ein konstanter Parameter, sondern als eine Zufallsvariable mit eigener Wahrscheinlichkeit angesehen. Genau hierin liegt aber das Problem, daß nämlich das Auffinden dieser sekundären Verteilung schwer zu begründen ist und häufig nicht gelingt. Zudem hat Sichel hierfür eine extrem komplizierte Funktion zu ihrer Beschreibung angegeben – das spricht zwar nicht gegen sein Modell (schließlich ist Einfachheit nicht entscheidendes Kriterium für die Qualität einer Annahme), läßt es aber in letzter Konsequenz auch nicht als sehr operational erscheinen.

Einen anderen Weg ist Altmann (1988) in seinen Überlegungen zur Satzlänge gegangen, ausgehend von zwei Annahmen, nämlich

1. daß jegliche Verteilung der Längen in einem Text (unter anderem, aber nicht ausschließlich also auch der Satzlängen) gesetzmäßig organisiert ist, und
2. daß es ausreichend ist, Annahmen über die Differenz zweier jeweils benachbarter Wahrscheinlichkeiten zu machen.

Aus der zweiten Annahme leitet sich ab:

$$(2) \quad P_x - P_{x-1} = \Delta P_{x-1}.$$

Die Differenz zwischen diesen beiden Wahrscheinlichkeiten ist dabei nicht konstant, sondern hängt vom jeweiligen P_{x-1} ab, so daß sich ergibt:

$$(3) \quad D = \frac{P_x - P_{x-1}}{P_{x-1}} = \frac{\Delta P_{x-1}}{P_{x-1}}.$$

Die Frage ist, welche Faktoren nun auf D Einfluß nehmen können. Unter Verweis auf die Zipf'schen Kräfte (s.o.) hat Altmann (1988: 152) die folgenden Faktoren genannt:

a – Wirkung des Produzenten (Stil u.a.)

c – Wirkung des Rezipienten (der Sprachgemeinschaft)

b – Faktoren des Textes

d – Faktoren der Sprachebene

Im Hinblick auf die Berechnung der Satzlänge in Anzahl der Teilsätze kommt der Faktor d folglich nicht zum Tragen; somit ergibt sich nach Altmann (1988) die folgende Gleichung:

$$(4) \quad D = \frac{b - ax}{cx}$$

Hierbei wirken die Faktoren a und b „gestaltend“, während c „bremsend“ wirkt (a hat ein negatives Vorzeichen, insofern der Produzent versucht, seinen eigenen Stil in die Bindung der Textsorte einzubringen); weiterhin wirkt b „global“, während a und c „lokal“ wirken. Falls allerdings die Satzlänge in Wortanzahl gemessen wird, kommt der intervenierende Faktor der Sprachebene (d) hinzu, was zu der Gleichung (5) führt:

$$(5) \quad D = \frac{b - ax}{cx + d}$$

Dies führt zu den beiden Ansätzen (6) und (7)

$$(6) \quad P_x = \frac{b - ax}{cx} P_{x-1}$$

$$(7) \quad P_x = \frac{b - ax}{cx + d} P_{x-1}$$

Wir können uns hier die weiteren detaillierten Ableitungen ersparen und wollen uns statt dessen auf die aus ihnen resultierenden Verteilungsmodelle beschränken. Hierbei handelt es sich im einen Fall (Satzlänge in Anzahl der

Teilsätze) um die negative Binomialverteilung (8), im anderen Fall (Satzlänge in Anzahl der Wörter) um die Hyperpascal-Verteilung (9):

$$(8) P_x = \binom{r+x-1}{x} p^r q^x \quad x = 0, 1, 2, \dots$$

$$(9) P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0 \quad x = 0, 1, 2, \dots$$

Interessanterweise hatte bereits Fucks (1968) in seinen Untersuchungen zur Satzlänge im Deutschen darauf hingewiesen, daß die negative Binomialverteilung sich gut eigne, die Verteilung von Satzlengthen zu erfassen, wobei er allerdings die Satzlengthe einer Reihe von deutschen Texten in der Anzahl der Silben pro Satz berechnet hatte.³ Die beiden in (8) und (9) dargestellten Modelle hat Altmann (1988: 155) zu Ehren des Begründers der Satzlengthenforschung als „Sherman-Gesetze“ bezeichnet, und sie sind von ihm selbst an umfangreichem Material bestätigt worden.

Es gilt im Hinblick auf unsere auf die Satzlengthe von Sprichwörtern ausgerichtete Fragestellung jedoch einen Punkt zu beachten, der oben bereits angesprochen wurde: Wir haben es bei den Sprichwörtern nicht mit einem fortlaufenden Fließ-Text zu tun, von dem angenommen werden kann, daß er insgesamt einem selbst-regulatorischen Mechanismus folgt (innerhalb dessen auch die Verteilung der Satzlengthen gesteuert wird). Vielmehr stellt im Grunde genommen jedes einzelne Sprichwort einen in sich abgeschlossenen Text dar, und die Ebene eines Gesamttextes wird de facto erst gar nicht erreicht. In diesem Sinne hätten wir es bei einem Sprichwortkorpus de facto mit etwas Anderem als mit einem homogenen Text zu tun:

1. entweder hätten wir etwas vor uns, was Orlov (1982) als „Quasi-Text“ bezeichnet hat, eine Text-Mischung also, die nach Altmann (1992) nicht die notwendige Datenhomogenität aufweist, welche die Wirksamkeit der genannten selbst-regulatorischen Prozesse gewährleistet;
2. oder aber wir hätten es mit etwas zu tun, was sich in Analogie zu einer Lexikonstruktur sehen ließe, d.h. ein paradigmatisches Inventar von

³ Dabei war Fucks allerdings von ganz anderen Voraussetzungen ausgegangen und hatte die Formel auf ganz andere Art und Weise abgeleitet: Ausgegangen war er von der Poisson-Verteilung, die dann die negative Binomialverteilung ergibt, wenn der Parameter der Poisson-Verteilung (der Mittelwert) einer Gammaverteilung folgt.

Einheiten, im gegebenen Fall allerdings von Sätzen, nicht von Lexemen als Einträgen.

Während die erste Annahme dazu führen dürfte, daß Versuche, ein theoretisches Verteilungsmodell für unsere Satzlängenhäufigkeiten zu finden, zum Scheitern verurteilt sind (weil in diesem Fall Chaos, keine Ordnung zu erwarten ist), würde die zweite Annahme insofern dem Status des Sprichworts gerecht, als sie zugleich als Einheiten der Sprache und als Texte anzusehen sind. Auch würde sich die zweite Annahme durchaus mit der umgangssprachlich motivierten Annahme decken, derzufolge wir eine Sprichwortsammlung ja durchaus auch als Sprichwörter*lexikon* bezeichnen.

Zusätzlich zu beachten wäre folgender Umstand: Beim Sprichwort haben wir es mit einem Text bzw. mit einer Textsorte zu tun, von der angenommen werden kann, daß sie im Laufe der Zeit bestimmten Anpassungs- oder Optimierungsprozessen unterlegen ist, die ihrerseits normierend auf die Textstruktur Einfluß genommen haben. Aufgrund dieser Annahme wäre zu erwarten, daß sich die Verteilung der Satzlängenhäufigkeiten von Sprichwörtern anders verhält als eine randomisierte Stichprobe beliebiger, aus einem gegebenen Fließtext extrahierter Sätze – eine Annahme, die im gegebenen Zusammenhang natürlich nicht weiter verfolgt werden kann.

Wenden wir uns im folgenden statt dessen den konkreten Analysen unseres slowenischen Sprichwortmaterials zu. Dabei werden wir nicht auf theoretischem, sondern auf empirischem Wege vorgehen; d.h. wir werden das Material nicht darauf hin überprüfen, ob eine im Vorhinein als begründbar postulierte bzw. hypostasierte Verteilung(sfunktion) sich am Sprichwortmaterial als passend erweist, sondern wir werden das Material dahingehend untersuchen, welche Verteilungsfunktionen sich unter welchen Bedingungen als geeignet erweisen. Wenn dies gelingt, werden wir im Nachhinein versuchen, auf deduktivem Wege eine Begründung für die gefundene Verteilung(sfunktion) zu erarbeiten.

4. Zu einem Modell der Verteilung von Satzlängen in slowenischen Sprichwörtern

Bei der Untersuchung der Satzlänge(n) von Sprichwörtern stellt sich die oben bereits aufgeworfene Frage nach der Maßeinheit, in welcher die Satzlänge zu messen ist. Dem von Altmann unterbreiteten Vorschlag der Messung in Teilsätzen zu folgen, erweist sich aus praktischer Sicht als sinnlos, weil Sprichwörter aufgrund ihrer Tendenz zur Kürze überwiegend aus nur einem Satz ohne weitere syntaktische Untergliederung bestehen, so daß sie in bezug auf

die Anzahl der Teilsätze zu wenig variieren, als daß hier brauchbare Ergebnisse abzuleiten wären. Andererseits ist nicht auszuschließen, daß eine Berechnung in Worten aufgrund der zu erwartenden Variation allgemeine Trends kaschiert, so daß sich eine zu große Unschärfe bemerkbar macht.

Wir werden aus diesem Grunde einerseits unsere Berechnungen auf der Basis der Wortanzahl pro Satz durchführen, andererseits aber alle Anpassungen mehrfach durchführen, indem wir verschiedene Klassenzusammenfassungen vornehmen; d.h. wir werden zwar die Satzlänge in Worten berechnen, aber nicht nur in der Anzahl der einzelnen Wörter, sondern auch in jeweils zusammengefaßten Zweier-, Dreier- und Vierergruppen. Die Zusammenfassung in solche Gruppen ist ein in der Satzlängenforschung durchaus übliches Verfahren, weil aufgrund der extremen Variabilität der Satzlängen oft die einzelnen Klassen in den Stichproben zu dünn besetzt sind, als daß hier verallgemeinernde Aussagen möglich sind. Ob und wie sich jedoch gegebenenfalls die Art der Zusammenfassung auswirkt, ist bislang noch nicht systematisch untersucht worden. Wenn wir also mehrere Klassenzusammenfassungen durchführen, sollte dabei nicht der Eindruck entstehen, als würde hier das Datenmaterial manipuliert, um die Anpassung an die theoretischen Modelle zu optimieren – ganz im Gegenteil: durch die Berechnung der unterschiedlichen Zusammenfassungen soll gerade dieses ausgeschaltet werden. Dabei wird noch ein weiterer Umstand zu bedenken sein, nämlich der jeweilige Anfangspunkt der Zusammenfassungen: Wir könnten die jeweils ersten Klassen ohne weiteres als diejenigen definieren, in der 1-2, 1-3, 1-4, usw. Wörter pro Satz vorkommen, die jeweiligen zweiten Klassen entsprechend als diejenigen, in der 3-4, 4-6, 5-8, usw. Wörter pro Satz vorkommen – dies ist auf den ersten Blick eine durchaus sinnvolle Zusammenfassung, da es ja keine Sprichwörter mit 0 Wörtern pro Satz geben kann. Allerdings fallen hierbei die unterschiedlichen Satzlängenhäufigkeiten in unterschiedliche zusammengefaßte Klassen, so daß erneut die Gefahr einer durch die Meßgröße bedingten Verschiebung besteht.

Aus diesem Grunde werden wir im folgenden für die wortbasierten Berechnungen alle möglichen Typen von Zusammenfassungen dahingehend überprüfen, ob es trotz unterschiedlicher Zusammenfassungen zu einem konsistenten Modell der Häufigkeitsverteilung von Satzlängen in unseren Sprichwörtern kommt.

Auf eine (methodologisch analoge) Berechnung der Satzlänge auf der Basis der Silbenanzahl pro Satz werden wir im vorliegenden Zusammenhang aus verschiedenen Gründen verzichten. Es könnte sich zwar – unter anderem aufgrund der in der Parömiologie immer wieder herausgestellten metrischen und rhythmischen Strukturierung von Sprichwörtern – als überaus sinnvoll

erweisen, die Satzlänge von Sprichwörtern nicht in Worten, sondern in Silben zu berechnen, doch scheint es angebracht, zunächst einmal die Ergebnisse am Sprichwortmaterial mit den bisher bekannten Modellen zu vergleichen, was bislang ja noch nie geschehen ist. Die Analyse der Satzlänge von Sprichwörtern auf der Basis von Silben sollte deshalb einer eigenen Abhandlung vorbehalten bleiben, in deren Rahmen das zu untersuchende Material auch auf weitere Sprachen auszuweiten wäre (vgl. Grzybek 2001).

5. Satzlänge in Worten

Wie nicht ganz unerwartet, gelingt keine gute Anpassung an eine theoretische Verteilung, wenn das Material ungruppiert untersucht wird. Hervorragende Anpassungen lassen sich hingegen vornehmen, wenn das Material, wie oben beschrieben, in verschiedenen Klassen zusammengefaßt wird. Unten werden die Ergebnisse im Detail dargestellt; doch zuvor seien die hierbei ins Spiel kommenden Modelle kurz erörtert.

Insgesamt gibt es drei verschiedene Verteilungsmodelle, die sich in allen Fällen, also ohne Rücksicht auf die Art der Zusammenfassung, als geeignet erweisen: die Dacey-Negativ-Binomialverteilung (12), die Dacey-Poisson-Verteilung (13), sowie die Hyper-Poisson-Verteilung (14).

$$(12) \quad P_x = (1-\alpha) \binom{k+x-1}{x} p^k q^x + \alpha \binom{k+x-2}{x-1} p^k q^{x-1}, \quad x = 0, 1, 2, \dots$$

$$k \geq 0, 0 < p \leq 1, q = 1-p, 0 \leq \alpha \leq 1$$

$$(13) \quad P_x = \frac{(1-\alpha)a^x e^{-a}}{x!} + \frac{\alpha x a^{x-1} e^{-a}}{x!}, \quad x = 0, 1, 2, \dots \quad \text{wenn } a \neq 0$$

$$a \geq 0, 0 \leq \alpha \leq 1$$

$$(14) \quad P_x = \frac{a^x}{{}_1F_1(1; b; a) b^{(x)}}, \quad x = 0, 1, 2, \dots$$

$$a \geq 0, b \geq 0$$

Sowohl bei der Dacey-Negativ-Binomialverteilung (12) als auch bei der Dacey-Poisson-Verteilung (13) handelt es sich um gemischte Verteilungen; die Tatsache, daß sich mit diesen gute Anpassungsergebnisse erzielen lassen, wäre gegebenenfalls damit zu erklären, daß das Datenmaterial in der Tat heterogen ist, so daß es nicht mit einer einzigen Verteilungsfunktion modelliert werden kann.

Keine gemischte Verteilung ist hingegen die Hyper-Poisson-Verteilung (14); schauen wir uns zunächst die Ergebnisse der Anpassungen dieser Verteilung an unsere Daten an, bevor wir einen Interpretationsversuch der Befunde unternehmen. Die Hyper-Poisson-Verteilung kommt ausnahmslos in allen sechs Arten der Zusammenfassung auf gute, in fünf der sechs Gruppierungsformen sogar auf sehr gute Anpassungen. Als gute bzw. sehr gute Anpassungen werden dabei solche Ergebnisse verstanden, bei denen der χ^2 -Anpassungstest (als Test für die Güte der Anpassung des theoretischen Modells an die beobachteten Daten) unter Berücksichtigung der entsprechenden Freiheitsgrade auf ein $P < 0.05$ bzw. $P < 0.01$ kommt oder der als χ^2 / N definierte Kontingenzkoeffizient auf ein $C < 0.02$ bzw. $C < 0.01$ (vgl. Grotjahn/Altmann 1993).

Die Ergebnisse der Anpassung der Hyper-Poisson-Verteilung – NP[i] – an unsere Daten – F[i] – sind für alle Typen der Zusammenfassung mit den entsprechenden Daten in den folgenden Tab. 2a-f bzw. Abb. 2a-f veranschaulicht.

Tab. 2a: 2er Intervalle (1-2 ff.)

Intervall	X[i]	F[i]	NP[i]
1-2	1	36	35,88
3-4	2	571	569,15
5-6	3	807	797,43
7-8	4	587	584,44
9-10	5	260	290,03
11-12	6	120	108,80
(≥13)	7	48	43,27
		<i>a</i>	1,54
		<i>b</i>	0,10
		χ^2_4	4,91
		<i>P</i>	0,2964**
		<i>C</i>	0,002**

Tab. 2b: 2er Intervalle (2-3 ff.)

Intervall	X[i]	F[i]	NP[i]
2-3	1	168	154,72
4-5	2	836	769,92
6-7	3	718	793,59
8-9	4	440	456,25
10-11	5	165	181,87
12-13	6	76	55,49
(≥14)	7	26	17,16
		<i>a</i>	1,30
		<i>b</i>	0,26
		χ^2_4	28,29
		<i>P</i>	0,00
		<i>C</i>	0,0116*

Abb. 2a: 2er Intervalle (1-2 ff.)

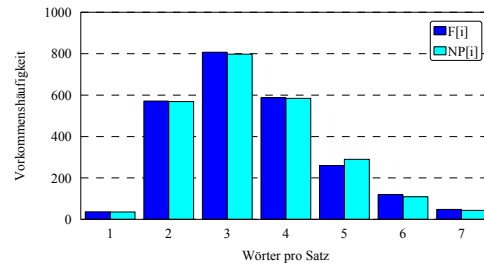
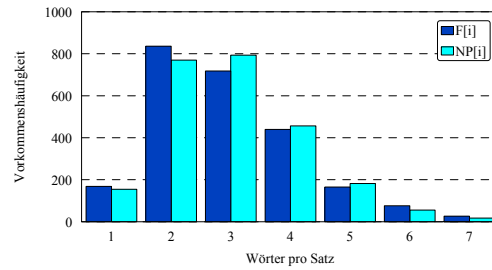


Abb. 2b: 2er Intervalle (2-3 ff.)



Tab. 2c: 3er Intervalle (1-3 ff.)

Intervall	X[i]	F[i]	NP[i]
1-3	1	168	168,71
4-6	2	1246	1251,26
7-9	3	748	732,06
10-12	4	219	222,94
13-15	5	36	45,89
(≥16)	6	12	8,13
		<i>a</i>	0,64
		<i>b</i>	0,09
		χ^2_3	4,42
		<i>P</i>	0,2196**
		<i>C</i>	0,0018**

Tab. 2d: 3er Intervalle (2-4 ff.)

Intervall	X[i]	F[i]	NP[i]
2-4	1	607	610,54
5-7	2	1115	1121,51
8-10	3	539	525,18
11-13	4	142	140,93
(≥14)	5	26	30,84
		<i>a</i>	0,63
		<i>b</i>	0,34
		χ^2_2	1,19
		<i>P</i>	0,5518**
		<i>C</i>	0,0005**

Abb. 2c: 3er Intervalle (1-3 ff.)

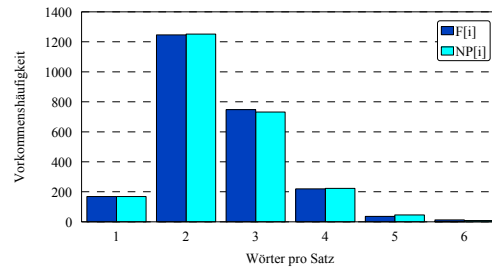
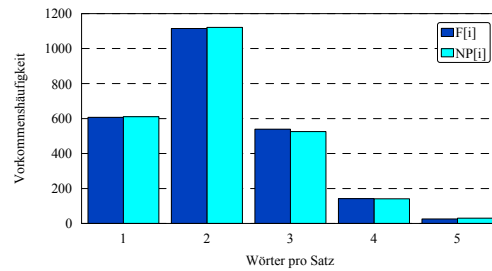


Abb. 2d: 3er Intervalle (2-4 ff.)

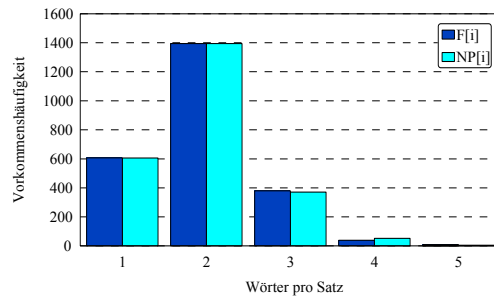
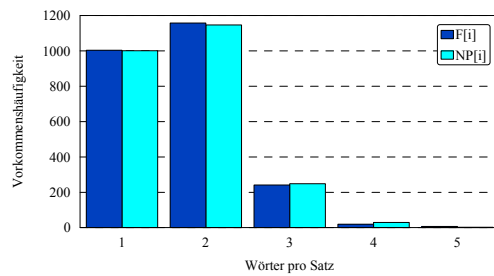


Tab. 2e: 4er Intervalle (1-4 ff.)

Intervall	X[i]	F[i]	NP[i]
1-4	1	607	605,79
5-8	2	1394	1394,39
9-12	3	380	371,01
13-16	4	39	52,39
(≥17)	5	9	5,42
	<i>a</i>		0,30
	<i>b</i>		0,13
	χ^2		6,00
	<i>P</i>		0,0498*
	<i>C</i>		0,0025**

Tab. 2f: 4er Intervalle (2-5 ff.)

Intervall	X[i]	F[i]	NP[i]
2-5	1	1004	1000,80
6-9	2	1158	1147,19
10-13	3	241	248,63
14-17	4	19	29,76
(≥18)	5	7	2,62
	<i>a</i>		0,27
	<i>b</i>		0,23
	χ^2		11,54
	<i>P</i>		0,0031*
	<i>C</i>		0,0047**

Abb. 2e: 4er Intervalle (1-4 ff.)**Abb. 2f: 4er Intervalle (2-5 ff.)**

Wie zu sehen ist, stellt die Hyper-Poisson-Verteilung in der Tat in allen Arten der Zusammenfassung ein bestens geeignetes Modell zur Modellierung der Satzlängenhäufigkeitsverteilung dar. Interessanterweise läßt sich nun ein Zusammenhang zwischen der Hyper-Poisson-Verteilung und der Hyper-Pascal-Verteilung (9) herstellen, die Altmann ja als generelles Modell für die (in Worten berechnete) Satzlängenverteilung in Texten postuliert hatte. Dies zeigt ein Blick auf die den Verteilungen zugrundeliegenden Differenzgleichungen. Für die von Altmann postulierte Hyper-Pascal-Verteilung sieht diese wie bereits oben in (7) dargestellt aus:

$$(7) P_x = \frac{b - ax}{cx + d} P_{x-1}.$$

Wenn man in dieser Gleichung nun $a = 0$ und $c = 1$ setzt, so ergibt sich

$$(15) P_x = \frac{b}{x + d} P_{x-1},$$

und genau aus dieser Differenzgleichung läßt sich die von uns empirisch als passend herausgearbeitete Hyper-Poisson-Verteilung ableiten.

Erinnern wir uns an die von Altmann vorgeschlagene Interpretation der Faktoren: a sollten diesem Modell zufolge Produktionsfaktoren (wie z.B. Stil u. a.) sein, b mögliche Textfaktoren, c ins Spiel kommende Rezeptionsfaktoren (z.B. Normierungen durch die Sprachgemeinschaft), und d schließlich war der gegebenenfalls intervenierende Faktor der Sprachebene(n).– Es wären also genau die autorenspezifischen (individualisierenden) Produktionsfaktoren (a) sowie die normierenden Rezeptionsfaktoren (c), die eliminiert – d. h. auf 0 bzw. 1 gesetzt würden (was mit der angesprochenen Kollektivität und dem daraus resultierenden „Einschleifen“ der Sprichwörter in Einklang zu sehen wäre).

Diese Interpretation entspricht im übrigen dem Umstand, daß die Hyper-Pascal-Verteilung (9) unter bestimmten Bedingungen – und zwar dann, wenn $k \rightarrow \infty$, $q \rightarrow 0$, $kq \rightarrow a$ – gegen die Hyper-Poisson-Verteilung (14) konvergiert.

6. Zusammenfassung und Ausblick

Die oben vorgenommenen Analysen können im Grunde genommen nicht mehr und nicht weniger als ein erster Schritt bei der Untersuchung sprachlicher Gesetzmäßigkeiten von Sprichwörtern sein. Die Analysen sollten auf jeden Fall an umfangreichem Material überprüft werden; für das Slowenische käme dazu z.B. die Sammlung von Šašelj/Kocbek in Frage. Doch müssen die Untersuchungen auch auf andere Sprachen erweitert werden, um sprachspezifische von sprachübergreifenden (sprichwortspezifischen?) Faktoren zu trennen. Aufgrund der oben erhaltenen Ergebnisse läßt sich jedenfalls für zukünftige Untersuchungen die Hypothese aufstellen, daß die Hyper-Poisson-Verteilung ein geeignetes Modell für die Verteilung von Satzlengthen in Sprichwörtern darstellt, wenn die Satzlengthe in der Anzahl der den Satz konstituierenden Wörter gemessen wird. Diese Hypothese steht natürlich auf der Basis des slowenischen Sprichwortmaterials allein auf tönernen Füßen; sie scheint sich aber, wie vorbereitende Analysen gezeigt haben (vgl. Grzybek/Krikmann 2001, Grzybek 2001), in der Tat an Sprichwörtern verschiedener Sprachen zu bestätigen.

Folgen müssen sodann ergänzende Analysen zur Satzlengthenverteilung auf der Basis der Silbenanzahl pro Wort, ebenso wie Häufigkeitsverteilung nicht nur von Satz-, sondern von Wortlengthen in Sprichwörtern. Schließlich sollten auch die Abhängigkeiten zwischen Satz- und Wortlengthe oder auch zwischen Wort- und Silbenlengthe in Sprichwörtern studiert werden; auch diese lassen sich nämlich auf Gesetzmäßigkeiten zurückführen, und es läßt sich zeigen,

daß diese Abhängigkeitsverhältnisse bei den Sprichwörtern denen von Wortlisten, nicht denen von Fließtexten gleichen – was der Annahme von der lexem-analogen Struktur der Sprichwörter ein weiteres Argument hinzufügt.

Auf diese Art und Weise werden wir jedenfalls Schritt für Schritt Einsicht in das Funktionieren der Sprichwörter als sprachliche Spezifika erhalten – denn eigentlich ist die Parömiologie bei der Untersuchung der sprachlichen Besonderheiten des Sprichworts bislang über eine symptomatische Betrachtungsebene nie hinaus gekommen und hat eine systematische Ebene (noch) nicht erreicht (vgl. für das Slowenische z. B. Förster 1992). Daß die Resultate nicht nur für die Sprichwortforschung von großer Bedeutung sind, steht dabei außer Frage.

Literatur

- Altmann, G. (1988): „Verteilung der Satztlängen.“ In: Schultz (Hrsg.) (1988), 147-169.
- Altmann, G. (1992): „Das Problem der Datenhomogenität.“ In: Rieger (Hrsg.) (1992), 287-298.
- Bojc, E. (1974): *Pregovori in reki na Slovenskem*. Ljubljana [²1980, ³1987]
- Borck, K. H.; Henss, R. (Hrsg.) (1968): *Der Berliner Germanistentag 1968. Vorträge und Berichte*. Heidelberg.
- Burger, H.; Zett, R. (Hrsg.) (1984): *Aktuelle Probleme der Phraseologie*. Bern u. a.
- Clayman, D. L. (1981): „Sentence length in Greek hexameter poetry.“ In: Grotjahn (Hrsg.) (1981), 107-136.
- Ľurčo, P. (Hrsg.) (1997): *Frazeologické štúdie II*. Bratislava.
- Eismann, W. (1997): „Заметки к сборнику пословиц Мегисера »Pargemio-logia polyglottos« 1592 и 1605.“ In: Ľurčo (Hrsg.) (1997), 63-77.
- Förster, W. (1992): „Zur Sprache slowenischer Sprichwörter.“ In: *Zeitschrift für Slawistik*, 37₂; 270-275.
- Fucks, W. (1968): „Analyse formaler Eigenschaften von Texten mit mathematischen Hilfsmitteln.“ In: Borck/Henss (Hrsg.) (1968), 42-52.
- Grafenauer, I. (1935): „Gutsmannov besednjak in njegova zbirka pregovorov, rekov in prilik.“ In: *Časopis za zgodovino in narodopisje*, 30; 1-29.
- Grotjahn, R. (Hrsg.) (1981): *Hexameter Studies*. Bochum.
- Grzybek, P. (2000): „Zum Status der Untersuchung von Satztlängen in der Sprichwortforschung – Methodologische Vor-Bemerkungen.“ In: *Слово во времени и пространстве. К 60-летию профессора В. М. Мокиенко*. Санкт Петербург, 430-457.

- Grzybek, P. (2001): „On the distribution of sentence length in proverbs.“
In: *Journal of Quantitative Linguistics*. (In Vorb.)
- Grzybek, P.; Krikmann, A. (2001): „Sentence length distribution in Estonian proverbs.“ (In Vorb.)
- Klančar, A.J. (1948): „A Tentative Bibliography on the Slovene Proverb.“
In: *Journal of American Folklore*, 61; 194-200.
- Kocbek, F. (1887): *Pregovori, prilike in reki*. Celje.
- Kocbek, F.; Šašelj, I. (1934): *Slovenski pregovori, reki in prilike*. Ljubljana.
- Levstik, F. (1882): „Zbirka slovenskih pregovorov iz leta 1592.“ In: *Levstikov zbrani spisi III*. Ljubljana 1895, 254-257.
- Mieder, W. (1988): „Einleitung.“ In: *Die deutschen Sprichwörter. Gesammelt von Karl Simrock*. Stuttgart, 7-15.
- Mieder, W. (1989): „Moderne Sprichwörterforschung zwischen Mündlichkeit und Schriftlichkeit.“ In: Röhrich/Lindig (Hrsg.) (1989), 187-208.
- Niehaus, B. (1997): „Untersuchung zur Satzlängenhäufigkeit im Deutschen.“
In: *Glottometrika 16. The Distribution of Word and Sentence Length*. Trier, 213-275.
- Orlov, Ju.K. (1982): „Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie ‘Sprache – Rede’ in der statistischen Linguistik).“ In: Orlov/Boroda/Nadarešvili (1982), 1-55.
- Orlov, Ju.K.; Boroda, M.G.; Nadarešvili, I.Š. (1982): *Sprache, Text, Kunst. Quantitative Analysen*. Bochum
- Prek, St. (1972): *Ljudska modrost. Pregovori, domislice in reki*. Maribor.
[Ljubljana, ²1974, ³1982, ⁴1986, ⁵1996]
- Radics, P. (1882): „Zbirka slovenskih pregovorov iz leta 1592.“ In: *Kres*, 2; 332-334.
- Rieger, B. (Hrsg.) (1992): *Glottometrika 13*. Bochum.
- Rode, M. (1989): „Slovenski pregovori in srednji vek.“ In: *Obdobje srednjega veka v slovenskem jeziku, književnosti in kulturi*. Ljubljana; 155-158.
- Röhrich, L.; Lindig, E. (Hrsg.) (1989): *Volksdichtung zwischen Mündlichkeit und Schriftlichkeit*. Tübingen.
- Schultz, K.-P. (Hrsg.) (1988): *Glottometrika 9*. Bochum.
- Sherman, L.A. (1888): „Some observations upon the sentence-length in English prose.“ *University of Nebraska Studies*, 1; 119-130.
- Sichel, H.S. (1974): „On a distribution representing sentence-length in prose.“ In: *Journal of the Royal Statistical Society*, A-137; 25-34.
- Toporišič, J. (1984): „Inhaltliche Aspekte der komparativen Phraseologeme in slowenischen Sprichwortsammlungen.“ In: Burger/Zett (Hrsg.) (1984), 291-321.

- Wake, W. C. (1957): „Sentence-length distribution of Greek authors.“ In: *Journal of the Royal Statistical Society, A-137*; 331-346.
- Williams, C. B. (1940): „A note on the statistical analysis of sentence-length as a criterion of literary style.“ In: *Biometrika*, 31; 356-361.
- Wimmer, G.; Altmann, G. (1999): *Thesaurus of univariate discrete probability distributions*. Essen.
- Yule, G. U. (1939): „On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship.“ In: *Biometrika*, 30; 363-390.
- Yule, G. U. (1944): *The statistical study of literary vocabulary*. Cambridge.