Komplexität sprachlicher Formen.
Die Singh-Poisson-Verteilung: ein Modell in der Wortlängenforschung?

Gordana Đuraš, Ernst Stadlober, Emmerich Kelih, Peter Grzybek

1. Einleitung

Ohne Zweifel sind Karl-Heinz Best große Verdienste auf dem Gebiet der Quantitativen Linguistik, insbesondere im Bereich der Modellierung von Wortlängenhäufigkeiten in unterschiedlichen Sprachen der Welt, zuzuschreiben. Der Umfang an Arbeiten, die im Zusammenhang mit dem Göttinger Projekt zur Quantitativen Linguistik[1] entstanden sind, geben ein eindrückliches Zeugnis dieser unermüdlichen Tätigkeit ab.

Parallel dazu, wenn auch zeitlich später einsetzend, setzte man sich auch im Grazer Projekt (unter maßgeblicher Beteiligung von Peter Grzybek, Ernst Stadlober, Gordana Đuraš und Emmerich Kelih)[2] ebenfalls intensiv mit Fragen der Quantitativen Sprach- und Textanalyse auseinander. Bei anfänglicher Konzentration auf das Problem der Modellierung von Wortlängenhäufigkeitsverteilungen aus theoretischer, methodologischer und empirischer Sicht wurden hier andere Fragen der Quantitativen Linguistik (QL) später und im Anschluss daran fokussiert. Und während sich Karl-Heinz Best im Rahmen des Göttinger Projekts mit der Wortlänge in einer Reihe von *unterschiedlichen* Sprachen auseinandersetzt (u.a. insbesondere Deutsch, aber auch viele andere Sprachen wie Althochdeutsch, Altisländisch, Chinesisch, Dänisch, Englisch, Erzamordwinisch, Estnisch, Färöisch, Französisch, Finnisch, Isländisch, Italienisch, Ketschua, Koreanisch, Latein, Mittelhochdeutsch, Niederdeutsch, Niederländisch, Norwegisch, Persisch, Russisch, Sami, Schwedisch, Spanisch, Türkisch, Tschechisch, Tscheremissisch, Ukrainisch, Ungarisch), war das Grazer Projekt von Anfang an hauptsächlich auf slawische Sprachen – insbesondere auf das Russische, Kroatische und das Slowenische – und dabei vor allem auf intralinguale Differenzierungen (Textsorten, Funktionalstile, Individualstile, usw.) fokussiert; nicht-slawische Sprachen sind erst in jüngster Zeit vereinzelt hinzugekommen und stellen nach wie vor nicht den Schwerpunkt des Projekts dar.

Im Göttinger Projekt, das ebenfalls zahlreiche andere Fragen aus dem Bereich der QL behandelte, hat sich bei der Frage der Modellierung der Vorkommenshäufigkeit verschiedener sprachlicher Einheiten im Laufe der Zeit immer wieder die Hyperpoisson-Verteilung als geeignetes Modell erwiesen; auf jeden Fall gilt

---

[1] Die bibliographische Erfassung der in diesem Projekt entstandenen Arbeiten findet sich im Internet unter dem URL: http://wwwuser.gwdg.de/~kbest/litlist.htm

[2] Für einen Überblick über die Ausrichtung dieser Gruppe vgl. www.project-quanta.org

dies im Hinblick auf Texte der deutschen Sprache (vgl. u.a. Best 2011). Im Vergleich dazu sind im Rahmen des Grazer Projekts im Hinblick auf Wortlängen wiederholt auch eine Reihe anderer Modelle diskutiert worden, auf die hier im Einzelnen nicht eingegangen werden muss. Eines dieser Modelle war im Hinblick auf Wortlängen u.a. die Singh-Poisson-Verteilung, auf die unten im Detail einzugehen sein wird. Vor diesem Hintergrund soll im Rahmen des hier gegebenen Zusammenhangs im Folgenden untersucht werden, inwiefern sich dieses Modell auch für das Deutsche eignet. Zu diesem Zweck soll im Folgenden eine unlängst von Karl-Heinz Best (2011) erstellte Studie zur Wortlänge in Texten des Deutschen zum Ausgangspunkt für eine Re-Analyse von Wortlängenhäufigkeiten genommen werden.

1.1. Wortlängenhäufigkeiten: Einleitung

Die Diskussion um adäquate diskrete Häufigkeitsmodelle für Wortlängenverteilungen wird innerhalb der Quantitativen Linguistik seit über 100 Jahren geführt und ist in der Vergangenheit wiederholt auf die Frage nach „dem einen" universal passenden Modell reduziert wurden. Die entsprechenden Etappen dieser Geschichte sind in Grzybek (2006) schrittweise und systematisch aufgearbeitet, so dass im gegebenen Zusammenhang nicht näher darauf eingegangen werden muss. Aus heutiger Sicht erweist sich vor allem der Ansatz von Wimmer/ Altmann (2005, 2006), auf den im Verlaufe dieses Textes noch einzugehen sein wird, von nachhaltiger Bedeutung, da es mit diesem möglich ist, aus einem gemeinsamen Proportionalitätsansatz eine Vielzahl von theoretischen Häufigkeitsmodellen abzuleiten (vgl. dazu auch schon Wimmer et al. 1994, Wimmer/ Altmann 2005).

Hinsichtlich der linguistischen Betrachtung ist die Frage nach den Faktoren, die einen Einfluss auf die Adäquatheit eines bestimmten Wortlängenhäufigkeitsmodells haben können, von nachdrücklichem Interesse. In erster Linie sind hierbei Faktoren wie Autorschaft, Textsorte, Funktionalstil und die Zugehörigkeit zu einem Diskurstyp relevant. Sie kommen insbesondere dann ins Spiel, wenn individuelle Texte als Basis für die Bestimmung der Wortlängenhäufigkeiten herangezogen werden, und nicht in etwa heterogene Korpora oder Stichproben aus Wörterbüchern.

Ein weiterer Faktor, der bislang allerdings kaum systematisch untersucht wurde, ist die Frage, ob und in welcher Form sprachenspezifische Unterschiede festzustellen sind. Oder, in anderen Worten: es herrscht – aufgrund der Vielzahl von Einflussfaktoren – Unklarheit darüber, ob Modelle nur für jeweils für eine Textsorte, einen Funktionalstil bzw. eine Sprache, oder aber eine ganze Sprachgruppe, eine Sprachfamilie, usw. gelten. Zu weiteren Faktoren und Problemen der Wortlängenmodellierung vgl. Grotjahn/Altmann (1993), Antić/Kelih/ Grzybek (2006), Altmann/Best/Wimmer (1997) oder auch die Beiträge von Grzybek und Popescu et al. in diesem Band.

1.2. Wortlängenhäufigkeiten im Slawischen

Wie oben bereits herausgestellt, war ein Teilziel des Grazer Projektes zur Quantitativen Text- und Sprachanalyse die Modellierung von Wortlängenhäufigkeiten im Kroatischen, Russischen und Slowenischen, wobei wiederholt auch Texte anderer Sprachen wie z.B. des Tschechischen, des Serbischen u.a. systematischen Untersuchungen unterzogen wurden. In den Untersuchungen wurde die Wortlänge überwiegenden in der Anzahl von Silben pro Wort gemessen. Im Verlaufe der Studien wurden verschiedene Verteilungsmodelle getestet, wobei sich in der Regel Erweiterungen und/oder Modifikationen von Poisson- und Binomial-Modellen als passend herausgestellt haben. In jüngerer Zeit hat sich dabei wiederholt die Singh-Poisson als ein geeignetes Modell erwiesen, eine zweiparametrische Verallgemeinerung der Poisson-Verteilung. So konnten Đuraš (2012) und Đuraš/Stadlober/Kelih (2013) an jeweils 120 Texten aus dem Slowenischen und Russischen, repräsentiert durch vier verschiedene Textsorten (Journalismus, Gedichte, Privatbriefe, Prosa), trotz der unterschiedlichen historischen Entwicklung dieser beiden slawischen Sprachen und ungeachtet offensichtlicher Unterschiede auf der phonologischen, morphologischen und lexikalischen Ebene zeigen, dass die Wortlängenhäufigkeiten in beiden Sprachen durch dieses Modell beschrieben werden können. Dasselbe Ergebnis stellte sich im Hinblick auf verschiedene Typen von 120 mündlichen slowenischen Texten (wie z.B. aufgezeichnete Telefonanrufe bei Hotelrezeptionen oder Tourismusbüros, verlesene Nachrichten und TV-Interviews) heraus, die Grzybek/Verdonik (2013) detailliert untersucht haben. Vor diesem Hintergrund stellt sich die Frage nach der Güte dieses Modells auch für deutsche Texte – diese Frage ist jedoch mehr als nur empirischer Natur, ergeben sich doch im positiven Fall weitere statistische und nicht zuletzt auch linguistische Unterschiede: So ist im Vergleich zu der oben genannten zweiparametrigen Hyperpoisson-Verteilung (λ,θ) der Parameter α der Singh-Poisson-Verteilung (α,θ) auf die Schätzung der ersten Häufigkeitsklasse beschränkt und dient als Gewichtungsfaktor für die übrigen, woraus sich im Ergebnis die Notwendigkeit einer im Vergleich zur Hyperpoisson-Verteilung anders angelegten linguistischen Erklärung ergäbe, der es in weiterer Folge für beide Sprachen nachzugehen gälte.

1.3. Wortlängenhäufigkeiten im Deutschen

In Anbetracht dieser Befunde liegt es nahe, einigen Eigenschaften der Singh-Poisson-Verteilung, insbesondere in ihrer Relation zur Hyperpoisson-Verteilung, ein wenig detaillierter nachzugehen und damit auch der Frage, ob sich diese Verteilung auch für Texte des Deutschen eignet. Dies soll im Verlauf der vorliegenden Studie anhand ausgewählter Datensätze überprüft werden, die der Studie von Best (2011) entstammen. In dieser Studie mit dem Titel „Silben-, Wort- und Morphlängen bei Lichtenberg“ hat Best in 20 kurzen Texten aus den *Sudelbüchern* von G. Chr. Lichtenberg (Heft H, 1784-1788, Lichtenberg 1971, 175-211)

[…] die Silbenlänge, die Wortlänge (gemessen in der Anzahl von Silben) und die Morphlänge bestimmt. Im gegebenen Zusammenhang ist die Aufmerksamkeit – für die Modellierung der Silben- und Morphlänge gibt es aus dem Grazer Projekt bislang keine entsprechenden Erfahrungen – auf die Wortlänge in der Anzahl von Silben zu richten. Best (2011: 5) schlägt als geeignetes Modell für die Modellierung der Wortlängenhäufigkeiten die 1-verschobene Hyperpoisson-Verteilung vor; diese Verteilung wird dabei aufgrund der reichhaltigen Erfahrungen im Göttinger Projekt als ein Grundmodell für die Wortlängenhäufigkeiten im Deutschen angesehen. Auch im Hinblick auf die untersuchten und hier zur Re-Analyse anstehenden Texte konnte damit von Best (2011) in den meisten Fällen ein überzeugendes Resultat gefunden werden; Details zu diesen Analysen finden sich in weiter unten in Abschnitt 3.

2. Modellierung der Wortlängen

Beide Verteilungsmodelle, die Hyperpoisson- ebenso wie die Singh-Poisson-Verteilung, lassen sich aus einem gemeinsamen Ansatz ableiten: Ein zentraler in der Geschichte der Modellierung der Wortlängenhäufigkeiten (vgl. Altmann/ Köhler 1995) verfolgter Ansatz besteht in der Annahme, dass Wortlängen auf einen rekursiven Generierungsmechanismus der Art

$$P_x = g(x) P_{x-1}$$

zurückgeführt werden kann, wobei P_x die Wahrscheinlichkeit einer gegebenen Wortlänge und $g(x)$ eine organisierende Proportionalitätsfunktion ist. In Abhängigkeit von der Beschaffenheit der Funktion $g(x)$ gelangt man so zu unterschiedlichen Modellen, so z.B. für

$$g(x) = \frac{\theta}{x}$$

zur üblichen Poisson-Verteilung, aus der sich durch die lokale Modifikation der ersten Wahrscheinlichkeit P_1 direkt die Singh-Poisson-Verteilung ableiten lässt (s.u.). Entsprechend führt die Erweiterung

$$g(x) = \frac{\theta}{\lambda + x}$$

zur Hyperpoisson-Verteilung, bei der es sich somit nicht um eine lokale Modifikation, sondern um eine komplexere Verallgemeinerung des Poisson-Modells handelt. Beide Modelle gehen natürlich auch aus dem allgemeinen Ansatz von Wimmer/Altmann (2005, 2006) hervor, worauf hier freilich nicht im Detail ein-

gegangen werden kann. Wohl aber ist eine detailliertere Betrachtung beider Modelle naheliegend, was im Folgenden geschehen soll.

2.1. Hyperpoisson-Verteilung

Die zweiparametrige Hyperpoisson-Verteilung (λ,θ) ist wahrscheinlich das am häufigsten benutzte Modell für Wortlängenhäufigkeiten, das in der Literatur mitunter auch als Modell für die Verteilung der Satzlänge erwähnt wird – vgl. dazu u.a. Antić et al. (2006), Best (2001), Kelih/Grzybek (2004), Altmann et al. (1997), Nemcová/Altmann (1994).

Diese Verteilung ist eine Verallgemeinerung der Poisson-Verteilung, welche die Poisson-Verteilung (θ) als einparametrigen Spezialfall enthält. Sie kann hergeleitet werden mit Hilfe des bedingten Poisson-Modells $X|Y \sim$ Poisson (θY) und der Annahme, dass der Parameter Y eine gestutzte Pearson-Typ III-Verteilung mit einer Wahrscheinlichkeitsdichte, die gegeben ist durch

$$g(y) = \frac{(\lambda-1)e^{\theta y}(1-y)^{\lambda-2}}{{}_1F_1[1;\lambda;\theta]}, \quad 0 \le y \le 1, \tag{1}$$

wobei $\lambda > 1$, $\theta > 0$ und ${}_1F_1[1;\lambda;\theta]$ die konfluente hypergeometrische Funktion (Kummers Funktion) mit dem ersten Argument gleich 1 darstellt, d.h.

$${}_1F_1[1;\lambda;\theta] = 1 + \frac{\theta}{\lambda} + \frac{\theta^2}{\lambda(\lambda+1)} + \ldots = \sum_{x=0}^{\infty} \frac{\theta^x}{\lambda^{(x)}} \tag{2}$$

mit Pochhammers Symbol $\lambda^{(x)} = \lambda(\lambda+1)\ldots(\lambda+x-1)$. Als Resultat ergibt sich die Wahrscheinlichkeitsfunktion der Mischverteilung als

$$P(X=x) = \int_0^1 \frac{e^{-\theta y}(\theta y)^x}{x!} \frac{(\lambda-1)e^{\theta y}(1-y)^{\lambda-2}}{{}_1F_1[1;\lambda;\theta]}\, dy = \frac{\theta^x \Gamma(\lambda)}{{}_1F_1[1;\lambda;\theta]\Gamma(\lambda+x)}$$

mit $\lambda > 1$, $\theta > 0$, daher ist ihre 1-verschobene Form gegeben durch

$$\pi_{x|\lambda,\theta} = P(X=x) = \frac{\theta^{x-1}}{{}_1F_1[1;\lambda;\theta]\lambda^{(x-1)}}, \; x = 1,2,\ldots \tag{3}$$

Die Berechnung der ersten beiden Momente ist sehr kompliziert; dies wird noch aufwendiger und zeitintensiver für Momente höherer Ordnung. Mittelwert und Varianz der Verteilung (3) sind gegeben durch

$$\mu = E(X) = 1 + \theta + (1-\lambda)\left(1 - {}_1F_1^{-1}[1;\lambda;\theta]\right) \tag{4}$$

$$\mathrm{var}(X) = \theta\mu + (\mu - 1)(2 - \mu - \lambda), \tag{5}$$

wobei sich der Dispersionsindex $\delta = \mathrm{Var}(X)/(E(X) - 1)$ ergibt als

$$\delta = \theta - \lambda - \mu + 2 + \frac{\theta}{\mu - 1}. \tag{6}$$

Parameter λ enthält Information über den Typ der Verteilung (3). Für $\lambda = 1$ haben wir offensichtlich $\lambda^{(x-1)} = (x-1)!$ und ${}_1F_1[1;\lambda;\theta] = e^{\theta}$, wodurch sich Verteilung (3) zum 1-verschobenen Poisson-Modell vereinfacht. Für $0 < \lambda < 1$ ist $\delta < 1$, wobei das Maximum bei eins erreicht wird, wenn θ groß genug ist – wir haben daher Unterdispersion ($\delta < 1$). Im Poisson-Fall haben wir $\delta = 1$. Für $\lambda > 1$hat man allerdings eine Überdispersion der Verteilung (vgl. Tabelle 1). Falls λ größer wird, verkleinert sich der Mittelwert, daher wird auch der Wert von δ größer, unabhängig vom Wert des Parameters θ.

Tabelle 1
Unter- und Überdispersion in der 1-verschobenen HP-Verteilung

λ		θ						
		0.1	0.5	0.8	1	2.4	5	8
0.3	μ	1.29	2.00	2.38	2.62	4.09	6.70	9.70
	δ	0.87	0.70	0.69	0.70	0.79	0.88	0.92
0.6	μ	1.16	1.71	2.08	2.31	3.78	6.40	9.40
	δ	0.96	0.89	0.87	0.86	0.88	0.93	0.95
0.9	μ	1.11	1.54	1.86	2.07	3.49	6.10	9.10
	δ	0.99	0.98	0.97	0.97	0.97	0.98	0.99
1	μ	1.10	1.50	1.80	2.00	3.40	6.00	9.00
	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.3	μ	1.08	1.40	1.66	1.83	3.14	5.70	8.70
	δ	1.01	1.04	1.06	1.06	1.08	1.06	1.04
1.4	μ	1.02	1.13	1.21	1.27	1.79	3.31	5.82
	δ	1.01	1.07	1.12	1.15	1.34	1.56	1.54

2.2 Singh-Poisson-Verteilung

Die Singh-Poisson-Verteilung[3] ist eine zwei-parametrige (α,θ) Alternative zur Poisson-Verteilung, anwendbar in Situationen, wenn die beobachteten Zähldaten eine spezifische Abweichung von der Poisson-Verteilung anzeigen. Diese Verteilung ist ein Spezialfall einer endlichen Mischung und resultiert aus der Kombination der Poisson-Verteilung mit der degenerierten (Ein-Punkt-)Verteilung, welche ihre Wahrscheinlichkeitsmasse an der Stelle 0 konzentriert – vgl. Đuraš/ Stadlober (2010), Đuraš/Stadlober/Kelih (2013). In ihrer 1-verschobenen Form ist die Wahrscheinlichkeitsfunktion der diskreten Zufallsvariable X mit der Singh-Poisson-Verteilung gegeben durch

$$p_{x|a,q} = P(X = x) = \begin{cases} 1 - a + ae^{-q}, & x = 1 \\ aq^{x-1}e^{-q} / (x-1)!, & x = 2,3,\ldots \end{cases} \qquad (7)$$

wobei $\theta > 0$ and $0 < \alpha \leq \alpha_{max} = \frac{1}{1-e^{-\theta}}$. Hier gibt α_{max} den maximal möglichen Wert von α für gegebenes θ an und resultiert aus der Bedingung $1 - \alpha + \alpha e^{-\theta} \geq 0$. Die ersten beiden Momente sind gegeben durch $E(X) = 1 + \alpha\theta$ und $\mathrm{Var}(X) = \alpha\theta(1 + \theta - \alpha\theta)$, daher ist der Dispersionsindex $\delta = 1 + \theta(1 - \alpha)$. Offensichtlich wird die Über- und Unterdispersion nur durch den Parameter α gesteuert, da θ positiv. Für $\alpha = 1$ hat man Equidispersion, für $0 < \alpha < 1$ Überdispersion, und im Fall von $1 < \alpha < \alpha_{max}$ Unterdispersion – für weitere Details siehe Đuraš (2012).

2.3 Parameterschätzung

Die beiden oben diskutierten Modelle unterscheiden sich nicht nur, wie oben erwähnt, im Hinblick auf die damit einhergehende linguistische Interpretation, sondern auch, und zwar gravierend, hinsichtlich der Komplexität der Schätzprozeduren. Das sei an Hand der drei am häufigsten verwendeten Schätzverfahren illustriert: die Momentenmethode (MM), die Maximum-Likelihoodmethode (ML) und die auf dem Stichprobenmittelwert und den Häufigkeiten der ersten Häufigkeitsklasse (hier also der einsilbigen Wörter) basierende Schätzung (FF).

[3] Die Singh-Poisson-Verteilung ist im Rahmen der Quantitativen Linguistik u.a. bereits von Wimmer/Witkovský/Altmann (1999) als passendes Modell für Wortlängenverteilungen ins Spiel gebracht wurden. Aus empirischer Sicht hat sich diese Verteilung auch für Texte verschiedener romanischer Sprachen (Altmann/Best/Wimmer (1997) und auch slawischer Sprachen (Slowenisch, Russisch) als passend erwiesen. Insofern ist es durchaus berechtigt nunmehr auch deutschsprachige Texte in Betracht zu ziehen.

2.3.1. Momentenmethode (MM)

Die Momentenschätzungen für die Parameter erhält man über die funktionale Beziehung zwischen den Parametern und den theoretischen Momenten, indem man die theoretischen Momente durch die Stichprobenmomente substituiert. Für das Hyperpoisson-Modell erhält man

$$\hat{\lambda}_{MM} = \frac{\hat{\theta}_{MM}\,\bar{x} + 3\bar{x} - m_2^{'} - 2}{\bar{x} - 1}, \tag{8}$$

wobei der Schätzer $\hat{\theta}_{MM}$ gegeben ist durch

$$\hat{\theta}_{MM} = \frac{m_3^{'}\left(\bar{x}-1\right) + m_2^{'}\bar{x} + m_2^{'} - \left(m_2^{'}\right)^2 - \bar{x}^2}{2\bar{x}^2 - \bar{x} - m_2^{'}} \tag{9}$$

Für das Singh-Poisson-Modell erhalten wir im Vergleich dazu wesentlich einfachere Lösungen:

$$\hat{\theta}_{MM} = \frac{m_{(2)}}{\bar{x}-1} - 2 \text{ und } \hat{\alpha}_{MM} = \frac{\bar{x}-1}{\hat{\theta}_{MM}}. \tag{10}$$

2.3.2. Maximum-Likelihood-Methode (ML)

Die Maximum-Likelihood-Schätzung ist jener Wert, welcher die Likelihood- bzw. die logarithmierte Likelihood-Funktion maximiert. Im Hyperpoisson-Modell erhält man diese Schätzung durch Lösung der so genannten Score-Gleichungen

$$\begin{aligned}
\frac{\partial l\left(\lambda,\theta \mid f\right)}{\partial \lambda} &= n\psi\left(\lambda\right) - \frac{n}{{}_1F_1\left[1;\lambda;\theta\right]}\frac{\partial\,{}_1F_1\left[1;\lambda;\theta\right]}{\partial\lambda} - \sum_{i=1}^{k} f_i\psi\left(\lambda+i-1\right) = 0 \\
\frac{\partial l\left(\lambda,\theta \mid f\right)}{\partial \theta} &= \frac{n\left(\bar{x}-1\right)}{\theta} - \frac{n}{{}_1F_1\left[1;\lambda;\theta\right]}\frac{\partial\,{}_1F_1\left[1;\lambda;\theta\right]}{\partial\theta} = 0
\end{aligned} \tag{11}$$

Für die Ermittlung der Lösungen lässt sich hier der sonst übliche Newton-Raphson-Algorithmus nicht anwenden, da ${}_1F_1\left[1;\lambda;\theta\right]$ nicht analytisch darstellbar ist. Gemäß Definition (2) ist offensichtlich, dass die Berechnung der partiellen Ableitungen von ${}_1F_1\left[1;\lambda;\theta\right]$ bzgl. λ eine schwierige Aufgabe darstellt. Butler/ Wood (2002) haben eine numerische Lösung für dieses Problem vorgeschlagen, wobei die Schätzung für das Hyperpoisson-Modell (3) auf einer Approximation

der log-Likelihood-Funktion basiert, bei der ${}_1F_1[1;\lambda;\theta]$ durch eine kalibrierte Laplace-Approximation ihrer Integraldarstellung ersetzt wird. Dieser Ansatz ist sehr komplex und zeitaufwendig; diesbezügliche Details findet man in Đuraš (2012).

Im Gegensatz dazu hat man für das Singh-Poisson-Modell die einfache Formel

$$\hat{\alpha}_{ML} = \frac{n - f_1}{n\left(1 - e^{-\hat{\theta}_{ML}}\right)}\,, \tag{12}$$

wobei $\hat{\theta}_{ML}$ die Lösung von

$$\frac{\theta(n - f_1)}{n(\bar{x} - 1)} + e^{-\theta} - 1 = 0 \tag{13}$$

ist.

2.3.3. Schätzung basierend auf Stichprobenmittelwert und erster Häufigkeitsklasse (FF)

Die Schätzungen basierend auf Mittelwert und erster Häufigkeitsklasse erhält man, indem der theoretische Mittelwert μ und die Wahrscheinlichkeit der ersten Klasse π_1 ersetzt werden durch den Stichprobenmittelwert $\bar{x}$ und die relative Häufigkeit der ersten Klasse f_1/n. Nach einigen algebraischen Vereinfachungen erhält man die expliziten Parameterschätzungen für das Hyperpoisson-Modell (3) als

$$\hat{\lambda}_{FF} = \frac{n(m_2 + 2) - \bar{x}(n + f_1)}{n - \bar{x}f_1} \tag{14}$$

und

$$\hat{\theta}_{FF} = \frac{m_2(n - f_1) - f_1(\bar{x} - 1)^2}{n - \bar{x}f_1}. \tag{15}$$

Es sei an dieser Stelle darauf hingewiesen, dass bereits Bardwell/Crow (1964) bemerkten, dass derartige Parameterschätzungen bei gleichem θ möglicherweise für Überdispersion besser geeignet sind als für Unterdispersion.

Für das Singh-Poisson-Modell sind diese Schätzer interessanterweise identisch mit dem ML-Schätzer: $\hat{\alpha}_{FF} = \hat{\alpha}_{ML}$ und $\hat{\theta}_{FF} = \hat{\theta}_{ML}$ – vgl. Đuraš / Stadlober (2010).

2.4. In-between-Resümee

Ein wichtiges Ergebnis der obigen Betrachtungen ist, dass beide Modelle sowohl Unter- als auch Überdispersion abdecken und daher prinzipiell dazu geeignet sind, die spezifische Silben- und Wortstruktur von Bests 20 kurzen Texten[4] abzubilden. Gestecktes Ziel dieser Arbeit ist es allerdings, heraus zu finden, ob das Singh-Poisson-Modell als die einfachere der beiden Alternativen geeignet ist, die Wortlängenverteilung deutscher Texte adäquat zu beschreiben.

Weitere Argumente, die aus statistischer Sicht für eine Bevorzugung des Singh-Poisson-Modells sprächen, wären dabei: (i) die Singh-Poisson-Parameter-Schätzungen sind stabil und einfach zu berechnen, (ii) eine Simulationsstudie in Đuraš (2012) zeigte gute Eigenschaften des Modells für alle Dispersionsszenarien, und (iii) für alle in Đuraš/Stadlober/Kelih (2013) analysierten Texte wurden brauchbare und stabile Schätzungen erzielt. Des Weiteren liefern die Parameterregionen des Singh-Poisson-Modells gut interpretierbare Charakterisierungen von Texttypen und Diskurstypen – besonders gute Ergebnisse ließen sich in dieser Hinsicht für slowenische Texte erzielen.

Im Gegensatz dazu zeigte die Simulationsstudie in Đuraš (2012) bzgl. des Hyperpoisson-Modells (3), dass die Resultate für alle drei möglichen Dispersionsfälle bzgl. aller drei Schätzprozeduren relativ ungenau waren; detaillierte Vergleiche dazu finden sich in Đuraš (2012).

3. Re-Analyse von Best (2011)

Es kann und soll an dieser Stelle nicht um einen „Konkurrenzkampf“ zweier Verteilungsmodelle gehen, d.h. um die Frage, welches von zwei gegebenen Modellen das „bessere“ ist – die Güte eines Modells wird nie allein nur durch Anpassungstests und deren Ergebnisse bestimmt, sondern ergibt sich durch eine ganze Reihe verschiedener Faktoren und Betrachtungsweisen (vgl. dazu u.a. Mačutek/Altmann 2008).

Ungeachtet dessen scheint es, bevor wir auf weitere Details eingehen, sinnvoll, zunächst die Ergebnisse der Anpassung an die Hyperpoisson-Verteilung zur Kenntnis zu nehmen. An 18 der 20 untersuchten Texte lässt sich, wenn man die Güte der Anpassung über die Wahrscheinlichkeit p der X^2-Verteilung bestimmt, die Hyperpoisson-Verteilung mit gutem bis sehr gutem Erfolg anpassen.[5]

---

[4] Die durchschnittliche Länge der 20 Texte im Umfang von $x_{min} = 87$ bis $x_{\max} = 369$ Wörtern pro Text beträgt $\overline{x} = 147.8$ Wörter, 18 der 20 Texte weisen weniger als 200 Wörter auf, was im Vorhinein erhebliche Schwankungen und damit verbundene Probleme der Modellierung erwarten lässt (s.u.).

[5] Üblicherweise wird die Anpassungsgüte mit dem X^2-Wert und der dazu gehörigen Wahrscheinlichkeit $p = P(X^2 > x)$ angegeben. Hier klassifizieren wir eine Anpassung als schlecht, wenn der p-Wert kleiner als 0.01 ist. Da der X^2-Wert jedoch linear mit der Stichprobengröße N ansteigt, hat es sich in der Quantitativen Linguistik eingebürgert,

Im Hinblick auf die beiden übrigen Texte[6] wäre zu sagen, dass einer der beiden (H155) den Wert von $p = 0.01$ geringfügig unterschreitet, allerdings gut mit der einfachen Poisson-Verteilung zu modellieren ist, so dass wir es offensichtlich mit dem Problem fehlender Freiheitsgrade zu tun haben, das sich aus dem aufgrund der dünnen Klassenbesetzung notwendigen Datenpooling ergibt. Der zweite Text (H146) weist eine atypische Häufigkeitsverteilung auf, insofern diese bimodalen Charakters ist, was entweder auf eine Textmischung hinweisen oder aber eine Folge der geringen Stichprobe sein könnte.

Die Tabellen 2, 3 und 4 zeigen die empirischen Häufigkeitsverteilungen aller 20 Texte von Best, deren Textlängen (*TL=N*) und die dazu gehörigen Dispersionsindizes $d = s^2/(\overline{x} - 1)$.

Wie eine Anpassung der Daten mit dem Altmann-Fitter und den darin implementierten Schätzverfahren zeigt, erweist sich die Singh-Poisson-Verteilung – auf vollkommen ähnliche Weise wie die Hyperpoisson-Verteilung – in denselben 18 von Fällen als gutes bzw. sehr gutes Modell; auch hier entzieht sich lediglich der bimodale Text H146 einer Modellierung, während Text H155, wie oben bereits gesagt, gut mit der einfachen Poisson-Verteilung zu modellieren ist.

Im Hinblick auf die obigen theoretischen Überlegungen werden in den Tabellen zusätzlich die mit *R*[7] berechneten ML-Parameterschätzungen für das Singh-Poisson-Modell, sowie die daraus hervorgehenden X^2-Werte, aus denen sich dann auch die Werte des Diskrepanzkoeffizienten $C = X^2/N$ ergeben die als Kriterien für die Güte der Anpassung aufgelistet. Wie aus den Anpassungsergebnissen ersichtlich ist, liefert das Singh-Poisson-Modell mit dieser Schätzmethode gute Anpassungen, zumindest für einen Großteil der Texte.[8]

---

alternativ bei größeren Stichproben (hier also: längeren Texten) den Diskrepanzkoeffizienten zu berechnen. Dabei werden wir Modellanpassungen als (a) ‚sehr gut' für $C \leq 0.01$, (b) als ‚gut' fürs $0.01 < C \leq 0.02$, und (c) als ‚akzeptabel' für $0.02 < C \leq 0.05$ bezeichnet. In der Studie von Best (2011) wurde aufgrund der relativ kurzen Texte (vgl. Fussnote 3) deshalb primär die Wahrscheinlichkeit des X^2-Werts als Bewertungsbasis genommen; da aber die einzelnen (vor allem die unteren) Klassen zusätzlich mitunter extrem dünn besetzt waren, und da aus diesem Grunde nach entsprechender Klassenzusammenfassung die Anzahl der Freiheitsgrade zu klein wurde, wurden beide Werte angegeben. Diesem Vorgehen soll auch im vorliegenden Text gefolgt werden, zumal es keine objektive Entscheidung darüber geben kann, wann eine Stichprobe als "klein" (folglich ein Text als „kurz") und wann als „klein" bzw. „lang" anzusehen ist.

[6] Best (2011) selbst spricht gar von vier Texten, bei denen das Hyperpoisson-Modell nicht passe, doch dürfte es sich hier um den Effekt unterschiedlicher Klassenzusammenfassungen handeln.

[7] *R* ist ein als Teil des GNU-Projekts frei verfügbares Statistikprogramm (http://www.r-project.org/).

[8] Im Vergleich zu den mit dem *Altmann-Fitter 3.1* erhaltenen Ergebnissen sind die Anpassungswerte insgesamt geringfügig niedriger, da keine zusätzlich optimierenden Iterationsprozeduren durchgeführt wurden. Da in einem Fall dadurch der Schwellwert von

Tabelle 2
Wortlängen in Lichtenbergs Sudelbuch mit geschätzten Parametern des Singh-Poisson-Modells (4)

	Absolute Häufigkeiten (f_x)							
x	H 10	H 13	H 14	H 15	H 19	H 52	H 53	H 66
1	62	42	70	43	93	72	56	93
2	30	25	28	23	60	34	45	58
3	12	14	12	14	22	13	19	10
4	8	10	6	9	9	4	3	6
5	0	2	0	1	2	0	1	1
6	0	0	1	0	0	0	0	1
TL	112	93	117	90	186	123	124	169
d	1.24	1.27	1.44	1.25	1.14	1.12	0.94	1.22
$\hat{\alpha}_{ML}$	0.72	0.76	0.63	0.74	0.86	0.80	1.05	0.91
$\hat{\theta}_{ML}$	0.97	1.30	1.02	1.24	0.87	0.73	0.73	0.69
X^2	1.058	1.932	0.376	2.334	0.872	0.03	0.492	5.997
FG	1	2	1	2	2	1	1	1
p	0.304	0.381	0.54	0.311	0.647	0.862	0.483	0.014
C	0.009	0.021	0.003	0.026	0.005	<0.001	0.004	0.035

Tabelle 3
Wortlängen in Lichtenbergs Sudelbuch mit geschätzten Parametern des Singh-Poisson-Modells (4)

Absolute Häufigkeiten (f_x)								
x	H125	H134	H135	H138	H146	H147	H148	H150
1	68	69	49	42	61	61	91	184
2	63	46	25	30	29	35	44	115
3	18	13	5	11	8	6	8	43
4	14	8	8	4	16	4	9	20
5	1	0	0	3	2	1	0	6
6	0	0	0	0	1	1	1	1
7	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0
9	0	0	0	0	1	0	0	0
TL	164	136	87	91	118	108	153	369
d	1.03	1.07	1.32	1.60	1.91	1.38	1.39	1.25
$\hat{\alpha}_{ML}$	0.99	0.92	0.71	0.78	0.61	0.80	0.71	0.80
$\hat{\theta}_{ML}$	0.89	0.77	0.96	1.17	1.60	0.79	0.85	0.98

$p < 0.01$ geringfügig unterschritten wird, sind die Anpassungen für 17 der 20 als gut bzw. sehr gut anzusehen.

X^2	8.303	2.323	6.183	3.83	11.755	4.752	7.879	3.309
FG	2	1	1	2	2	1	1	2
p	0.016	0.128	0.013	0.147	0.003	0.029	0.005	0.191
C	0.051	0.017	0.071	0.042	0.099	0.044	0.052	0.009

Tabelle 4
Wortlängen in Lichtenbergs Sudelbuch mit geschätzten Parametern des Singh-Poisson-Modells (4)

Absolute Häufigkeiten (f_x)				
x	H151	H155	H181	H191
1	166	88	92	53
2	65	55	64	31
3	24	7	13	16
4	11	7	8	6
5	3	3	3	0
6	0	0	1	0
TL	269	160	181	106
d	1.37	1.27	1.28	1.09
$\hat{\alpha}_{ML}$	0.63	0.86	0.87	0.83
$\hat{\theta}_{ML}$	0.93	0.74	0.83	0.92
X^2	1.509	11.451	9.463	0.152
FG	2	1	2	1
p	0.47	0.001	0.009	0.697
C	0.006	0.072	0.052	0.001

In Anbetracht dieser Befunde wäre daher einstweilen zusammenfassend der Schluss zu ziehen, dass sich aufgrund der Anpassungsergebnisse beide Modelle – die Hyperpoisson-Verteilung ebenso wie die Singh-Poisson-Verteilung – für die untersuchten deutschen Texte trotz deren geringer Länge als durchweg geeignet erweisen.

Da beide Verteilungsmodelle offenbar nicht nur zur Modellierung in diesen Texten[9] und nicht nur für deutschsprachige Texte geeignet sind, sondern auch bereits erfolgreich auf andere Sprachen angewendet wurden, liegt es nahe, sich abschließend mit der Frage des Parameterverhaltens beider Verteilungsmodelle genauer auseinanderzusetzen und zumindest aus empirischer Sicht den Zusam-

[9] In einer Simulationsstudie von Đuraš (2012) konnte auch gezeigt werden, dass die Parameterschätzungen des Singh-Poisson-Modells ein stabileres Verhalten aufweisen als die Parameterschätzungen des Hyperpoisson-Modells. Für jeden der untersuchten Texttypen kann man mit Hilfe der ML-Schätzungen eine zuverlässige, durch die entsprechenden Texte abgedeckte Parameterregion (Parameterlandschaft) angeben und so texttypenspezifische Eigenheiten und auch Gemeinsamkeiten von unterschiedlichen Texttypen über die Charakteristiken der Parameterlandschaften quantifizieren.

menhang zwischen diesen beiden Modellen genauer zu betrachten. Dies betrifft einerseits das Verhältnis der beiden Parameter der Hyperpoisson-Verteilung (λ,θ) zueinander, andererseits das Verhältnis dieser beider zu den Parametern α und θ der Singh-Poisson-Verteilung.

Abb. 1 zeigt zunächst den Zusammenhang zwischen den Parametern θ und λ der Hyperpoisson-Verteilung.[10] Es ist leicht zu sehen, dass es sich hierbei um einen klaren, und zwar linearen Zusammenhang handelt, der im gegebenen Fall bereits mit der einfachsten linearen Funktion $\lambda = 2.07\theta$ auf $R^2 = 0.96$ kommt – natürlich ließe sich mit komplexeren linearen Funktion ein noch besseres Ergebnis erzielen, worauf es im hier gegebenen Kontext freilich nicht ankommt.

Es sei an dieser Stelle explizit vermerkt, dass ein solcher Zusammenhang keineswegs zwangsläufig aus dem Modell der Hyperpoisson-Verteilung hervorgeht, sondern sich vielmehr empirisch ergibt. Zwischen den Parameter α und θ der Singh-Poisson-Verteilung hingegen besteht kein linearer Zusammenhang (vgl. Đuraš 2012), was dafür spricht, dass wir es in der Tat mit einer primär auf die erste Klasse beschränkten lokalen Modifikation zu tun haben, die nicht grundsätzlich das Verhalten aller übrigen Klassen regelnd beeinflusst.

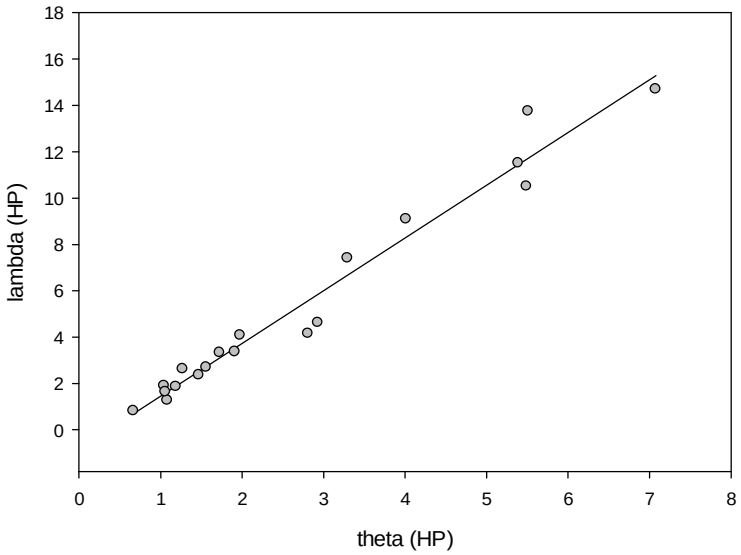

Abb. 1: Linearer Zusammenhang zwischen den Parametern θ und λ der Hyperpoisson-Verteilung

Über die Beobachtung des Zusammenhangs zwischen den beiden Parametern der Hyperpoisson-Verteilung hinausgehend stellt sich somit die Frage nach einer Beziehung zwischen den Parametern θ und λ der Hyperpoisson einerseits

[10] Grundlage sind 18 der 20 Datensätze, mit Ausnahme der beiden oben genannten (H146, H155).

und den Parametern α und θ der Singh-Poisson-Verteilung,[11] die für θ und α aus Abb. 2 ersichtlich sind.

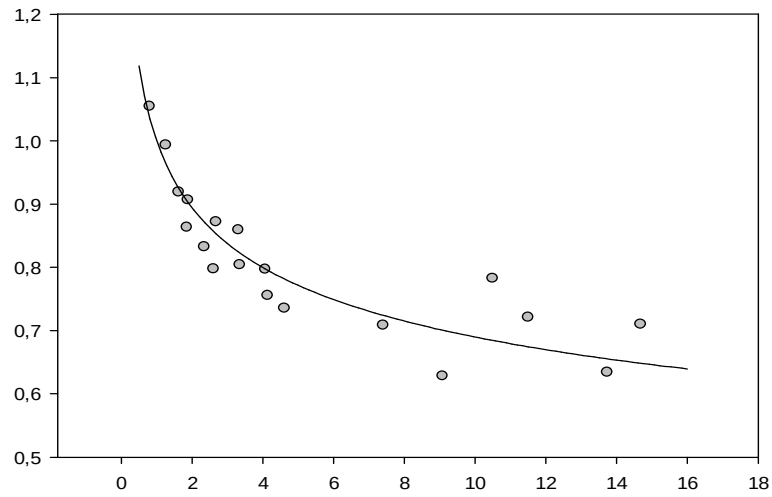

Abb. 2: θ (HP) – α (SP)

Wie aus Abb. 2 ersichtlich ist, scheint zumindest empirisch in der Tat ein nichtlinearer Zusammenhang zwischen dem Parameter θ der Hyperpoisson-Verteilung und dem Parameter α der Singh-Poisson-Verteilung vorzuliegen, der in einer ersten Annäherung mit der einfachen Potenzfunktion $\alpha = \theta^{-0.16}$ auf einen Determinationskoeffizienten von $R^2 = 0.84$ kommt. Es steht vollkommen außer Frage, dass weiterführende Schlussfolgerungen an dieser Stelle nicht zulässig sind, und dass der Möglichkeit eines solchen Zusammenhangs an anderer Stelle mit umfangreicherem Datenmaterial und größeren Stichproben nachgegangen werden muss.

Allerdings besteht kein erkennbarer Zusammenhang der Hyperpoisson-Parameter zum Parameter θ der Singh-Poisson-Verteilung. Dies spräche gegebenenfalls dafür, dass die Hyperpoisson-Verteilung als das insgesamt allgemeinere der beiden Modelle (s.o) auch und gerade deshalb so gut geeignet ist, weil es offenbar – unter anderem – lokale Spezifika wie die Modifikation der ersten Häufigkeitsklasse zu erfassen vermag.

4. Zusammenfassung

Neben der linguistischen Interpretierbarkeit der verwendeten Modelle – die allerdings bislang noch weitestgehend aussteht – sollte im Sinne des Occam'schen Prinzips der Parsimonie die generelle Einfachheit der Modelle ein Grundprinzip

[11] Aufgrund des linearen Zusammenhangs zwischen den Parametern θ und λ der Hyperpoisson-Verteilung ist klar, dass im Falle eines Zusammenhangs zur Singh-Poisson-Verteilung davon dann beide Parameter der Hyperpoisson-Verteilung davon betroffen sind.

bei Modellierungen sein. Dieses Prinzip ist auch für den Bereich der Parameterschätzungen erwägenswert. In diesem Sinne wurde in den obigen Darlegungen und Analysen einerseits gezeigt, dass die Schätzung der Parameter beim Singh-Poisson-Modell über die ML-Methode überraschend einfach ist und sogar mit der FF-Methode (Stichprobenmittelwert und erste Häufigkeitsklasse) zusammenfällt. Dies wäre zumindest als ein (kleiner) Vorteil gegenüber dem Hyperpoisson-Modell anzusehen, bei welchem zwar die MM-Schätzungen und die FF-Schätzungen einfach zu berechnen sind, aber die ML-Schätzungen überaus aufwendig zu ermitteln sind.

Insgesamt wäre allerdings, trotz der Vielzahl von im Detail noch zu klärenden Fragen (Textauswahl, Stichprobengröße, Verfahren Datenpooling, Parameterschätzung, u.a.m.) ein zentrales Resultat, welches u.a. auch aufgrund der Vielzahl der Arbeiten aus dem Göttinger und Grazer Projekt gewonnen werden kann, wie folgt zu formulieren: Die Wortlänge in Texten ist eine synergetisch organisierte Größe, die aus statistischer Sicht vor allem durch theoretische Verteilungen aus der Poisson-Familie zu erfassen ist – zumindest solange man das Wort auf einer orthographischen bzw. orthographisch-phonetischen Ebene definiert und seine Länge in der Anzahl der Silben pro Wort bestimmt und offenbar – so zumindest die Erfahrungen aus dem vorliegenden Text – germanische bzw. slawische Sprachen analysiert.

Literatur

Altmann, Gabriel; Best, Karl-Heinz; Wimmer, Gejza (1997). Wortlänge in romanischen Sprachen. In: Gather, A., Werner, H. (Hg.), *Semiotische Prozesse und Natürliche Sprache. Festschrift für Udo L. Figge zum 60. Geburtstag*. Stuttgart: Steiner, 1–13.

Antić, Gordana; Kelih, Emmerich; Grzybek, Peter (2006). Zero-syllable Words in Determining Word Length. In: Peter Grzybek (ed.): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer (Text, Speech and Language Technology, 31), 117–156.

Altmann, Gariel; Köhler, Reinhard (1995). 'Language Forces' and Synergetic Modelling of Language Phenomena. In: Schmidt, Peter (eds.), *Glottometrika 15. Issues in General Linguistic Theory and The Theory of Word Length*. Bochum: Brockmeyer, 62–76.

Bardwell, George E.; Crow, Edwin L. (1964). A two-parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association 59, 133–141.*,

Best, Karl-Heinz (2001). Wortlängen in Texten gesprochener Sprache. *Göttinger Beiträge zur Sprachwissenschaft 6, 31–42.*

Best, Karl-Heinz (2011). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 21, 1–13*.

Đuraš, Gordana; Stadlober, Ernst; Kelih Emmerich (2013). The Generalized Poisson Distributions as Models for Word Length Frequencies. In: Obradovič, Ivan; Köhler, Reinhard; Kelih, Emmerich (eds.), *Proceedings of Qualico 2013*. Beograd. [In print].

Đuraš, Gordana (2012). *Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages*. Dissertation, TU Graz.

Đuraš, Gordana; Stadlober, Ernst (2010). Modeling word length frequencies by the Singh-Poisson distribution. In: Grzybek, P., Kelih, E., Mačutek, J. (Eds.), *Text and Language. Structures· Functions · Interrelations. Quantitative Perspectives*. Wien: Praesens, 37–48.

Grotjahn, Rüdiger; Altmann, Gabriel (1993). Modelling the distribution of word length: Some methodological problems. In: Reinhard Köhler, Burghard B. Rieger (Eds.), *Contributions to Quantitative Linguistics. Proceedings of the First International Conference of Quantitative Linguistics, QUALICO, Trier, 199*1. Dordrecht; Boston; London: Kluwer Acad. Publ., 141–153.

Grzybek, Peter (2006). History and Methodology of Word Length Studies. The State of the Art. In: Peter Grzybek (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht, NL: Springer (Text, Speech and Language Technology, 31), 15–90.

Grzybek, Peter; Verdonik, Darinka (2013). Word length frequencies in oral Slovenian texts. [In prep.]

Kelih, Emmerich; Grzybek, Peter (2004): Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). *Glottometrics 8, 23–41*.

Mačutek, Ján; Altmann, Gabriel (2008). Testing Hypotheses in Quantitative Linguistics. In: Panchanan Mohanty, Reinhard Köhler (eds.), *Readings in Quantitative Linguistics*. Delhi: Radha Press, 33-44.

Nemcova, Emíliá; Altmann, Gabriel (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für Empirische Textforschung 1, 40–43.*

Wimmer, Gejza; Altmann, Gabriel (2005). Unified derivation of some linguistic laws. In: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), *791–801*.

Wimmer, Gejza; Altmann, Gabriel (2006). Towards a unified derivation of some linguistic laws. In: Peter Grzybek (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer, 329–337.

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1(1), 98–106.*

Wimmer, Gejza; Witkovský, Viktor; Altmann, Gabriel (1999). Modification of probability distributions applied to word length research. *Journal of Quantitative Linguistics 6(2), 257–268.*