ASIAN AND AFRICAN STUDIES, *7*, 1998, 1, 45–51

WORTLÄNGENHÄUFIGKEITEN IN CHINESISCHEN KURZGESCHICHTEN

Jinyang ZHU, Universität Hamburg, 20146 Hamburg
Karl-Heinz BEST, Georg-August-Universität Göttingen, 37073 Göttingen, Germany

Frequency of Word Lengths in Chinese Short Stories

Chinese short stories are used in this paper with the aim of testing once again the theory of word length distribution in texts as proposed by Wimmer et al. (1994) and Wimmer/Altmann (1996). The theory assumes that the word lengths in texts are not distributed randomly but have to correspond to quite specific, theoretically justifiable distributions. The paper is another step in the investigation of Chinese as well as another 36 languages covered by the quantitative linguistic project carried out in Göttingen and introduced to the readers of this journal in Best/ Song (1996).

0. Eine naheliegende Aufgabe bei der Erforschung jeder Sprache besteht darin, entweder für ein größeres Textkorpus insgesamt oder für viele Texte einzeln zu untersuchen, mit welcher Häufigkeit Wörter verschiedener Länge zu beobachten sind. So hat Fucks die Wortlänge als Kriterium verwendet, um den literarischen Stil einzelner Autoren zu charakterisieren, aber auch dazu, ganze Sprachen miteinander zu vergleichen (Fucks 1968: 6, 80). Als einzige asiatische Sprache wurde dabei das Japanische berücksichtigt, das sich als eine Sprache mit vergleichsweise hoher durchschnittlicher Wortlänge erwies (Fucks 1968: 91).

Einer etwas anderen Fragestellung soll in dieser Arbeit nachgegangen werden: Wie sind die Wortlängen in einzelnen Texten verteilt? Kann man das Auftreten der verschiedenen Wortlängenklassen als gesetzmäßig verstehen? Zeigt das Chinesische aufgrund seiner Sprachstruktur Unterschiede zu andern Sprachen oder findet man wieder die gleichen Verteilungen?

Als weitere Perspektiven sind Fragen nach den Zusammenhängen zwischen den Gesetzmäßigkeiten der Wortlängenverteilungen und anderen Aspekten der Sprachstruktur, z.B. stilistischen und typologischen Eigenschaften, ins Auge zu fassen. Daß die Wortlänge eine zentrale Größe der Sprachstruktur ist, konnte am Beispiel des Deutschen (Köhler 1986: 74) und des Polnischen (Hammerl 1991: 219) überzeugend nachgewiesen werden.

1. Hier geht es nun darum, Beobachtungen zur Wortlänge chinesischer Kurzgeschichten vorzustellen und zu prüfen, ob sie sich mathematisch modellieren lassen. Es stellt sich die Frage, ob die Wortlängenverteilungen dieser literarischenTexte sich von denen der andern, bisher untersuchten Textsorten (Zhu/Best 1992a, b; Best/Zhu 1994: 27f.; Zhu/Best 1997; Bohn 1998) unterscheiden. Eine Besonderheit chinesischer Texte besteht darin, daß – vor allem bei Texten mit einem höheren Anteil an fachsprachlicher Lexik – z.B. vier- und sechssilbige Wörter häufiger als drei- und fünfsilbige vorkommen (vgl. Best/Zhu 1994: 28), sodaß man einen „oszillierenden" Datenverlauf erhält. Ein solches Phänomen konnte noch in keiner weiteren der 37 bisher untersuchten Sprachen beobachtet werden (Best/Altmann 1996).

Diese Besonderheit ist bei Kurzgeschichten aber nicht zu erwarten, da sie keine längeren Wörter in nennenswertem Umfang enthalten. Stattdessen ist zu prüfen, ob die Ergebnisse, die mit Mao-Briefen erzielt wurden (Zhu/Best 1997), für die literarischen Texte bestätigt werden können, auch wenn beide Textsorten natürlich zu unterschiedlichen Funktionalstilen gehören (Fleischer/Michel/Starke 1993: 28).

2. Die vorliegende Untersuchung schließt sich so eng wie möglich an die Prinzipien an, die in andern, vergleichbaren Arbeiten (vgl. Best/Song 1996) befolgt wurden. D.h.: Es wurde für jede Kurzgeschichte gesondert die Zahl der einsilbigen, zweisilbigen usw. Wörter festgestellt. Anders als bei andern Sprachen kann das „Wort" im Chinesischen aber nicht als orthographische Einheit aufgefaßt werden; es wird stattdessen als eine distributionell-semantische Einheit bestimmt (Zhu/Best 1992b). Die Zahl der Silben pro Wort bemißt sich nach der Zahl der in ihm enthaltenen Vokale. Es wird immer nur der laufende Text – also ohne Zusätze wie Überschriften – ausgewertet.

3. Nach Erarbeitung der Daten für die einzelnen Texte wurde geprüft, ob die untersuchten Kurzgeschichten einem bestimmten mathematischen Verteilungsmodell folgen oder nicht. Zu diesem Zweck wurde der Altmann–Fitter (1994) verwendet, eine Software, mit deren Hilfe man die diskreten Wahrscheinlichkeitsverteilungen, die für Wortlängenverteilungen als Modelle begründet wurden (Wimmer u.a. 1994; Wimmer/Altmann 1996), an gegebene Dateien anpassen kann. Der Altmann-Fitter (1994) testet die Güte der Anpassung der Verteilungen an die einzelnen Dateien mit Hilfe des Chiquadrattests (X^2) und gibt so Auskunft darüber, ob eine gute Übereinstimmung zwischen Theorie und Beobachtung erzielt werden kann.

Die Auswahl eines geeigneten Modells unter den theoretisch möglichen ist nicht ganz unproblematisch, da sie sich nicht von vornherein aufgrund bekannter Bedingungen wie Sprache, Textsorte, Autor etc. treffen läßt. Für die chinesischen Kurzgeschichten hat sich aber ebenso wie für die Briefe Maos (Zhu/Best 1997) die positive Cohen-Poisson-Verteilung als geeignet erwiesen, deren Formel wie folgt lautet:

$$P_x = \begin{cases} \dfrac{(1-\alpha)\,\mathrm{a}}{e^{a}-1-\alpha\,\mathrm{a}}, & x=1 \\ \dfrac{a^x}{x!\,(e^{a}-1-\alpha\,\mathrm{a})}, & x=2,3,4,... \end{cases}$$

$$\mathrm{a} > 0;\ 0 < \alpha < 1.$$

(Die andern erwähnten Untersuchungen werden hier nicht weiter berücksichtigt, da sie nur einzelne Texte berücksichtigen.)

4. In den folgenden Tabellen finden sich die Ergebnisse der Anpassung der positiven Cohen-Poisson-Verteilung an die Daten der chinesischen Kurzgeschichten. Dabei bedeuten: x die Wortlängen, n_x die im jeweiligen Text beobachtete Häufigkeit, mit der Wörter der Länge x darin vorkommen; NP_x die berechnete Häufigkeit dieser Wortlängenklasse. a und α sind die Parameter der Verteilung. X^2 ist das Chiquadrat, FG sind die Freiheitsgrade; P ist die Überschreitungswahrscheinlichkeit des ermittelten Chiquadrats, C der Diskrepanzkoeffizient. P wird als zufriedenstellend angesehen, wenn $P \geq 0.05$ ist; eine Anpassung mit $0.01 \leq P < 0.05$ wird als noch akzeptabel betrachtet. In den meisten Fällen kann aber P nicht bestimmt werden, dann nämlich, wenn bei der Anpassung des Modells keine Freiheitsgrade erhalten bleiben. In diesen Fällen kann als Prüfkriterium nur der Diskrepanzkoeffizient $C = X^2/N$ verwendet werden, dessen Wert dann als zufriedenstellend gewertet wird, wenn $C \leq 0.02$ ist.

Die Ergebnisse der Anpassung der positiven Cohen-Poisson-Verteilung an die Kurzgeschichten:

	Text 1		Text 2		Text 3	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	37	37.05	51	47.67	49	49.92
2	24	23.93	36	39.44	54	55.83
3	10	9.94	19	21.37	8	8.24|
4	3	3.10|	11	8.68	4	1.01|
5	1	0.98|	2	2.82|		
6			1	0.76|		
7			1	0.26|		
a =	1.2467		1.6258		0.4429	
α =	0.0351		0.0176		0.8020	
X^2 =	0.00		1.42		0.90	
FG =	1		2		0	
P =	0.97		0.49			
C =					0.01	

Text 1: Wang Meng, *Bu ru suanlatang.* Quelle: *Zhongguo dangdai xiaoxiaoshuo jingxuan.* Hong Kong: xin yazhou 1991.
Text 2: Wang Meng, *Xiao xiao xiao xiao xiao.* Quelle: wie Text 1.
Text 3: Jiang Shengqun, *Na jiu shi wo.* Quelle: wie Text 1.

Anmerkung: Die senkrechten Linien in den Tabellen zeigen an, daß die entsprechenden Wortlängenklassen zusammengefaßt wurden.

	Text 4		Text 5		Text 6	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	37	37.27	66	66.03	48	47.95
2	57	58.29	52	51.97	38	38.04
3	5	8.43|	2	2.86|	9	9.10
4	6	1.01|	1	0.14|	2	1.91
a =	0.4339		0.1653		0.7177	
α =	0.8613		0.8950		0.5476	
X^2 =	0.29		0.0001		0.01	
FG =	0		0		1	
P =					0.94	
C =	0.003		0.0000008			

Text 4: Ling Guang, *Zuijia jiangyan.* Quelle: wie Text 1.
Text 5: Cao Qiang, *Shu.* Quelle: wie Text 1.
Text 6: Chen Shuqin, *Baba de huixin.* Quelle: wie Text 1.

	Text 7		Text 8		Text 9	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	31	31.60	40	39.96	41	41.91
2	42	45.02	54	54.00	61	60.30
3	5	7.36|	5	8.04|	7	7.56
4	7	1.02|	4	1.00|		
a =	0.4907		0.4469		0.3437	
α =	0.8278		0.8346		0.8829	
X^2 =	1.79		0.0001		0.05	
FG =	0		0		0	
C =	0.02		0.0000009		0.0005	

Text 7: Jin Jiang, *Jijide canying.* Quelle: *Lao lü tui mo.* Shanghai: Shaonianertong 1981.
Text 8: Jin Jiang, *Tuzide huayuan.* Quelle: wie Text 7.
Text 9: Qu Guanghui, *Houzi zhao jing.* Quelle: wie Text 7.

	Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	58	58.00	57	57.04	53	47.74
2	69	69.00	66	65.95	37	42.89
3	2	2.90|	3	6.49|	10	12.89
4	1	0.10|	4	0.52|	7	3.52
a =	0.1261		0.2952		0.8990	
α =	0.9470		0.8724		0.4997	
X^2 =	0.0001		0.0001		5.53	
FG =	0		0		1	
P =					0.02	
C =	0.0000007		0.0000007			

Text 10: Jin Jiang, *Shimo.* Quelle: wie Text 7.
Text 11: Jin Jiang, *Yanzi he jiage.* Quelle: wie Text 7.
Text 12: Jin Jiang, *Feizaopao.* Quelle: wie Text 7.

5. Diese Ergebnisse bedeuten, daß die positive Cohen-Poisson-Verteilung an alle Texte angepaßt werden kann. Die Anpassung an Text 12 ist allerdings nicht gut, aber doch noch akzeptabel. Gründe für solche relativ schwachen Ergebnisse sind vermutlich Inhomogentäten von Texten, die z.B. dadurch zustandekommen können, daß ein Autor sie nachträglich noch einmal bearbeitet hat, wie dies ja bei literarischen Texten häufig geschieht. Es gibt eine Vielzahl weiterer Faktoren, die zur Inhomogenität von Texten beitragen können, die hier aber vorläufig außer Betracht bleiben sollen. In Übereinstimmung mit den Ergebnissen in Zhu/Best (1997) kann jedenfalls festgestellt werden, daß auch die vorliegende Untersuchung keinen Anlaß ergeben hat, die Anfangshypothese, daß Wortlängen sich gemäß theoretisch begründbaren Verteilungen verhalten, abzulehnen. Die positive Cohen-Poisson-Verteilung gehört zu den von Wimmer u.a. (1994) und Wimmer/Altmann (1996) vorgeschlagenen Modellen.

6. Für weitere Forschungen sind folgende Überlegungen anzustellen: Die Tatsache, daß einige, mehr oder weniger zufällig ausgewählte Kurzgeschichten – so wie auch die Mao-Briefe – der positiven Cohen-Poisson-Verteilung folgen, darf nicht voreilig verallgemeinert werden. Texte anderer Autoren, anderer Zeitabschnitte der Entwicklung des Chinesischen oder anderer Textsorten können

ganz andere Ergebnisse erbringen. Insofern handelt es sich bei den bisher durchgeführten Untersuchungen lediglich um erste, noch vorläufige Ergebnisse. Es könnte auch sein, daß sich die Kürze der bearbeiteten Texte negativ bemerkbar macht: Die Chance, daß auch etwas längere Wörter in solchen Texten vorkommen, dürfte mit der Länge der Texte zunehmen. Dies würde wiederum die Überprüfung des Modells verbessern.

Auf einen weiteren Aspekt derartiger Untersuchungen sei ausdrücklich aufmerksam gemacht: Auch wenn man nicht das Ziel verfolgt, der sprachtheoretischen Frage nachzugehen, ob Strukturphänomene immer benennbaren und begründbaren Gesetzmäßigkeiten folgen, sind Daten, wie sie hier vorgelegt wurden, nützliche Grundlagen für stilistische (Mistrík 1973) oder sprachtypologische (Altmann/Lehfeldt 1973) Zwecke. So hat erst vor wenigen Jahren Silnitsky (1993) das Chinesische mit Hilfe typologischer Indizes charakterisiert und in eine typologische Klassifikation einbezogen. Es ist klar, daß der von Silnitsky benutzte Synthese-Index ebenfalls die Wortkomplexität betrifft. Der Vorteil von Untersuchungen wie der hier vorliegenden besteht darin, daß nicht nur ein Durchschnittswert – das sind nämlich die Indizes, wie sie auch Silnitsky verwendet – für eine Sprache angegeben wird. Dies berührt Fragen der Repräsentativität, die hier aber nicht weiter verfolgt werden sollen.

7. Abschließend sei noch ein Blick auf andere ostasiatische Sprachen gestattet. Erste Ergebnisse liegen zu japanischen und koreanischen Texten vor.

Kim/Altmann (1996) haben 24 koreanische Texte verschiedener Textsorten (Schulbuchtexte, Pressetexte und Erzählungen) bearbeitet, wobei sich herausstellte, daß diese Texte überwiegend der Conway-Maxwell-Poisson-Verteilung, z.T. mit Modifikationen, folgen. Drei der Texte ließen sich aber weder mit dieser noch mit einer anderen Verteilung modellieren. Best/Song (1996) haben Briefe und Pressetexte untersucht und konnten mit der Dacey-Poisson-Verteilung und mit der Hyperpoisson-Verteilung gute Ergebnisse erzielen.

Zu 11 japanischen Pressetexten aus „Asahi Shinbun“ hat Riedemann (1997) entsprechende Erhebungen durchgeführt. Hier zeigte sich, daß alle Texte der Hirata-Poisson-Verteilung folgen; dies ist eine Verteilung, die sich bisher vor allem bei französischenTexten bewährt hat (Dieckmann/Judt 1996; Feldt/Janssen/Kuleisa 1997).

Diese Hinweise zeigen, daß die Wortlängen in den bisher untersuchten ostasiatischen Sprachen unterschiedlichen Verteilungsmodellen entsprechen, die nur an nur einzelne Texte nicht angepaßt werden konnten. Es zeichnet sich bisher aber keine weitergehende Übereinstimmung zwischen diesen Sprachen ab, die aufgrund langen Kontakts immerhin denkbar wäre. Dennoch hat sich aber die von Wimmer u.a. (1994) und Wimmer/Altmann (1996) entwickelte Theorie auch bei diesen Sprachen bewährt.

LITERATUR

ALTMANN, Gabriel/ LEHFELDT, Werner. 1973. *Allgemeine Sprachtypologie.* München: Fink

BEST, Karl-Heinz/ALTMANN, Gabriel. 1996. Project Report. *Journal of Quantitative Linguistics* 3: 85-88

BEST, Karl-Heinz/SONG, Hea-Yean. 1996. Wortlängen im Koreanischen. *Asian and African Studies* 5: 39-49

BEST, Karl-Heinz & ZHU, Jinyang. 1994. Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: KLENK, U. (Hg.), *Computatio Linguae II*. Stuttgart: Steiner. 19-30

BOHN, Hartmut. 1998. *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift.* Hamburg: Kovač

DIEKMANN, Sabine/JUDT, Birga. 1996. Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: SCHMIDT, Peter (Hg.), *Glottometrika 15.* Trier: Wissenschaftlicher Verlag Trier. 158-165

FELDT, Sabine/JANSSEN, Marianne/KULEISA, Silke. 1997. Untersuchungen zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: BEST, Karl-Heinz (Hg.), *Glottometrika 16.* Trier: Wissenschaftlicher Verlag Trier (erscheint)

FLEISCHER, Wolfgang/MICHEL, Georg/STARKE, Günter. 1993. *Stilistik der deutschen Gegenwartssprache.* Frankfurt/M. u.a.: Peter Lang

FUCKS, Wilhelm. 1968. *Nach allen Regeln der Kunst.* Stuttgart: Deutsche Verlags-Anstalt

HAMMERL, Rolf. 1991. *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.*Trier: Wissenschaftlicher Verlag Trier

KIM, Icheon/ALTMANN, Gabriel. 1996. Zur Wortlänge in koreanischen Texten. In: SCHMIDT, Peter (Hg.), *Glottometrika 15.* Trier: Wissenschaftlicher Verlag Trier. 205-213

KÖHLER, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer

MISTRÍK, Jozef. 1973. *Exakte Typologie der Texte.* München: in Komm.: Sagner

RIEDEMANN, Gesa. 1997. Wortlängenhäufigkeiten in japanischen Pressetexten. In: BEST, Karl-Heinz (Hg.), *Glottometrika 16.* Trier: Wissenschaftlicher Verlag Trier (erscheint)

SILNITSKY, George. 1993. Typological Indices and Language Classes: A Quantitative Study. In: ALTMANN, Gabriel (ed.), *Glottometrika 14*. 139-160

WIMMER, Gejza/ALTMANN, Gabriel. 1996. The Theory of Word Length: Some Results and Generalizations. In: SCHMIDT, Peter (Hg.), *Glottometrika 15*. Trier: Wissenschaftlicher Verlag Trier. 112-133

WIMMER, Gejza/KÖHLER, Reinhard/GROTJAHN, Rüdiger/ALTMANN, Gabriel. 1994. Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics* 1: 98-106

ZHU, Jinyang/BEST, Karl-Heinz. 1992a. Zum Monosyllabismus im Chinesischen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 45: 341-355

ZHU, Jinyang/BEST, Karl-Heinz. 1992b. Zum Wort im modernen Chinesisch. *Oriens Extremus* 35: 45-60

ZHU, Jinyang/BEST, Karl-Heinz. 1997. Zur Modellierung der Wortlängen im Chinesischen. In: BEST, Karl-Heinz (Hg.), *Glottometrika 16.* Trier: Wissenschaftlicher Verlag Trier (erscheint)

SOFTWARE

Altmann-Fitter. 1994. Lüdenscheid: RAM-Verlag

AKTUELLE INFORMATIONEN IM INTERNET

http://www.gwdg.de/~kbest/projekt.htm